# Estudio del comportamiento de la comunidad científica: palabras clave y revisión por pares


UNIVERSIDAD
DE GRANADA

**TESIS DOCTORAL**

**Jorge Chamorro Padial**

Programa de Doctorado en Tecnologías de la Información y la Comunicación

**Departamento de Ciencias de la Computación e Inteligencia Artificial**

**Escuela Internacional de Posgrado**

**Universidad de Granada**

**Mayo 2023**

Editor: Universidad de Granada. Tesis Doctorales
Autor: Jorge Chamarro Padial
ISBN: 978-84-1117-996-6
URI: https://hdl.handle.net/10481/84445

Documento maquetado con TEXIS v.1.0.

Este documento está preparado para ser imprimido a doble cara.

# Estudio del comportamiento de la comunidad científica: palabras clave y revisión por pares

*Memoria que presenta para optar al título de Doctor por la Universidad de Granada*

**Jorge Chamorro Padial**

*Dirigida por la Doctora*

**Rosa Rodríguez-Sánchez**

Programa de Doctorado en Tecnologías de la Información y la Comunicación

**Departamento de Ciencias de la Computación e Inteligencia Artificial**

**Escuela Internacional de Posgrado**

**Universidad de Granada**

**Mayo 2023**

UNIVERSIDAD
DE GRANADA

D[a] ROSA MARÍA RODRÍGUEZ SÁNCHEZ. PROFESORA TITULAR DE UNIVERSIDAD ADSCRITA AL DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN E INTELIGENCIA ARTIFICIAL DE LA UNIVERSIDAD DE GRANADA.

**INFORMA:**

Que los trabajos de investigación que se exponen en la presente Memoria de Tesis Doctoral, publicada bajo la modalidad de Agrupación de Publicaciones y titulada: **Estudio del comportamiento de la comunidad científica: palabras clave y revisión por pares**, que presenta Jorge Chamorro Padial, han sido realizados bajo mi dirección y tutela durante los años de desarrollo de su correspondiente Tesis Doctoral, siendo expresión de la capacidad técnica e interpretativa de su autor en condiciones propicias que lo hacen merecedor del Título de Doctor, siempre y cuando así sea considerado por el Tribunal que designe la Universidad de Granada.

FDO. PROF. DRA. D[a] ROSA MARÍA RODRÍGUEZ SÁNCHEZ

En Granada, a 12 de mayo del 2023.

# Agradecimientos

En esta memoria, se sintetiza el resultado de algo más de cinco años de trabajo. Un proceso que ha atravesado varias etapas y que ha quedado plasmado en forma de compendio de publicaciones. Como no será difícil de imaginar; el desarrollo de esta tesis no ha sido ajeno a todas las circunstancias sociales, sanitarias y, por supuesto, personales que han acontecido desde su inicio.

Esta tesis no habría podido ser realizada sin el apoyo de personas que han resultado claves y que merecen una mención en estas líneas.

En primer lugar, debería hablar de los antecedentes que han dado lugar a la tesis, para poder darle contexto a la misma. Y es que este trabajo no sería una realidad si en el año 2010 no hubiera decidido matricularme en el Grado de Ingeniería Informática de la Universidad de Granada (que, por cierto, fue el primer año del grado, enmarcado dentro del denominado *Plan Bolonia*). Realmente ese pudo haber sido mi primer y último año en la Universidad o, al menos, en el Grado de Informática. Para alguien que no traía consigo conocimientos matemáticos ni físicos, el primer año de carrera no fue especialmente sencillo y sí realmente frustrante. Cuando ya tenía prácticamente decidido abandonar la carrera, tuve la suerte de tener de profesor a **Joaquín Fernández Valdivia**, y desde ese, momento entendí que las Ciencias de la Computación eran para mi y nunca más volví a plantearme dejar mis estudios. Tengo muchas cosas más que agradecer a Joaquín, pero creo que esta es la más importante de cara a la presente tesis.

Sin embargo, hasta tercero de carrera (y más bien entrando ya en cuarto...) la idea de investigar no estaba en mis planes. Y es en ese momento cuando apareció **Julio Ortega Lopera**; mi profesor de la asignatura de *Arquitectura de Computadores*, quien me animó a escribir en la *Revista de Enseñanza y Aprendizaje de Ingeniería de Computadores* [1], editada por el *Departamento de Arquitectura y Tecnología de Computadores* de la Universidad de Granada. El empujón final me lo dio **Héctor Pomares Cintas**, quien me terminó de convencer en una reunión improvisada, y en ese momento descubrí que la experiencia de redactar un artículo científico me re-

[1] https://icar.ugr.es/informacion/actividades/informacion/actividades-diversas/revista Accedida el 22 de Abril del 2023

sultaba bonita, motivante y muy constructiva. Esa fue la primera vez que publiqué un paper (Chamorro Padial, 2014) y, aunque no fuera un trabajo que objetivamente podamos catalogar de relevante, si fue la semilla que terminó germinando y desarrollando mi faceta investigadora. Me gustó tanto la experiencia que en menos de un año la quise mejorar, descubriendo que era aún mejor colaborar con alguien. Y junto al hoy Doctor **Fernando Palacios López**, conseguí publicar mi segundo trabajo (Palacios López y Chamorro Padial, 2015) y confirmar mi interés en la carrera investigadora, que me llevó a estudiar un máster enfocado plenamente en la investigación y con una idea clara de tener una tesis doctoral tarde o temprano.

Comenzó esta tesis un cinco de junio del 2017, cuando también estaba dando mis primeros pasos en el mundo laboral, como programador novato (o *junior*, anglicismo con el que se les suele denominar en el ámbito laboral a los que están empezando). Dado que la totalidad de la tesis la he desarrollado compaginándola con mi vida laboral, no es de extrañar que esta tesis también haya evolucionado a medida que yo lo iba haciendo como profesional. Hasta el punto en el que creo que es imposible saber qué habilidades profesionales me han repercutido positivamente en el desarrollo de la tesis y qué competencias del mundo de la investigación me han permitido crecer en el ámbito laboral. Lo que sí tengo claro es que esta sinergia formada por la unión de estos dos mundos: laboral e investigación, ha sido posible gracias a muchos compañeros de trabajo con los que he tenido la suerte de coincidir, y que me han dado la oportunidad de asumir cada vez más responsabilidades y salir continuamente de mi zona de confort. En este sentido, tengo que mencionar especialmente a **Almudena González-Recio**.

Justamente nada más empezar mi carrera profesional en T-Systems, pude conocer a **Francisco Javier Rodrígo-Ginés**, además de ser coautor de uno de los artículos de esta tesis, ha sido una persona clave en todo su desarrollo. Sus brillantes ideas, su capacidad de análisis y su gran inteligencia han supuesto una tremenda ayuda a la hora de darle forma a los artículos que he ido redactando.

**José Antonio García Soria** también debe, como no podría ser de otra manera, figurar en estos agradecimientos, por todas las aportaciones que ha hecho en buena parte de mis artículos así como por mantenerme informado de referencias bibliográficas claves para mi investigación así como por la revisión final y los comentarios realizados a este documento. Quizás, en este punto, es buen momento para mencionar en su conjunto al **Grupo de Visión por Computador** de la Universidad de Granada, que me ha apoyado estos años y me ha acompañado en el desarrollo de esta tesis.

Volviendo al análisis del contexto de esta tesis. No solamente la he podido desarrollar gracias a tener un entorno laboral propicio y estar rodeado de grandes investigadores y profesores, sino que también ha sido posible gracias a poder contar con el apoyo de mis padres y de grandes amigos, como

**Guillermo Landa Sánchez**, que siempre ha estado ahí cuando hacía falta.

Quería, finalmente, reservar el último agradecimiento, y el más especial de todos, para la persona que me ha enseñado, orientado y *aguantado* con paciencia durante tantísimo tiempo: **Rosa Rodríguez Sánchez**. Es completamente imposible transcribir en estas líneas todo lo que me ha aportado y posiblemente ni yo mismo soy plenamente consciente de la suerte que ha sido, para mi, tener a Rosa como Directora de Tesis.

Rosa ha conseguido que este proceso de varios años, de trabajar fines de semana y noches se haya convertido en una experiencia de la que no me arrepiento para nada en absoluto. Mención aparte se merece la libertad que he tenido este tiempo para desarrollar mis inquietudes en la tesis, siempre recibiendo un apoyo expreso por su parte.

En estos agradecimiento he tratado de reconocer la importancia que, a título individual, han tenido diferentes personas en la consecución de esta tesis. Pero es igualmente importante mencionar que esta tesis la he desarrollado en la Universidad de Granada, institución pública y, por tanto, ha sido financiada con los impuestos pagados por el conjunto de la Sociedad, quien también ha financiado mi estancia en todas las etapas educativas hasta llegar a donde estoy. Considero de justicia mencionar este hecho a fin de poner en valor la importancia de la Educación Pública.

# Resumen

En esta tesis doctoral realizamos, por medio del desarrollo y propuesta de diferentes modelos matemáticos y computacionales, un estudio sobre la comunidad científica y su comportamiento enmarcado, principalmente, en dos ámbitos de investigación: El proceso de revisión por pares desde el punto de vista de autores y editores, y el proceso de selección de palabras clave de autor para categorizar trabajos científicos.

En cuanto a la revisión por pares, aplicamos un modelo concebido originalmente para el ámbito de la biología, el modelo de cuasi especie, mediante el cual hemos modelado el comportamiento del proceso de revisión por pares en diferentes escenarios.

Las palabras clave de autor afectan al impacto y la visibilidad de los trabajos a los que catalogan. Por este motivo, hemos realizado diferentes trabajos que tratan de crear herramientas para asesorar a los autores a la hora de optimizar el proceso de selección de palabras clave utilizando información sobre la popularidad de un término y el número de trabajos ya existentes que utilizan estos mismos términos. Por otro lado, mediante el análisis de redes complejas, identificamos las palabras clave centrales y aquellas que ocupan una posición más periférica.

Fruto de esta tesis, se han escrito un total de diez artículos, de los cuales nueve ya han sido publicados y un décimo artículo está en elaboración. Además, se han creado y publicado bases de datos y aplicaciones, a raíz de los trabajos anteriormente citados.

**Palabras clave**: Revisión por pares; Palabras clave; Modelo de Cuasi-especie; Modelos computacionales; Comunidad Científica; Teoría de Juegos

# Summary

In this doctoral thesis we conducted, through the development and promotion of various mathematical and computational models, a study on the scientific community and its behavior, mainly framed in two research areas: the peer review process from the point of view of authors and editors, and the author keyword selection process to categorize scientific works.

Regarding peer review, we applied an originally conceived model for the biology field, the quasi species model, through which we have modeled the behavior of the peer review process in different scenarios.

Author keywords affect the impact and visibility of the works they catalog. For this reason, we have carried out different studies that try to create tools to advise authors when optimizing the keyword selection process using information about a term's popularity and the number of existing works that use those same terms. On the other hand, through complex network analysis, we identify central keywords and those that occupy a more peripheral position.

As a result of this thesis, ten articles have been written; nine of them have already been published, and a tenth is being drafted. Additionally, databases and applications created from previous work have been published.

**Keywords**: Peer review; Keywords; Quasiespecies model; Computational models; Scientific Community; Game theory

# Índice

# Capítulo 1

# Sobre este trabajo

## 1.1. Introducción y motivación

En este trabajo, se recoge el fruto de cuatros años de investigación realizada durante el desarrollo de una tesis doctoral titulada *Estudio del comportamiento de la comunidad científica: palabras clave y revisión por pares*. El objetivo de esta tesis, que podemos enmarcar dentro del área de las Tecnologías de la Información y la Comunicación, en el subárea de la *Sociología Computacional*, es el de colaborar a arrojar luz sobre la Academia y, más concretamente, sobre dos fenómenos que marcan el comportamiento de buena parte de la comunidad científica: Por un lado, la revisión por pares; paso necesario de cara a la publicación de trabajos científicos. Por otro lado, la búsqueda del impacto académico; mediante la cual esos trabajos publicados se ofrecen al resto de personas que conforman la Academia, así como al resto de la sociedad, y que son registrados y categorizados mediante *palabras clave*, que juegan un papel relevante a la hora de permitir que un documento sea encontrado (o no) por otros investigadores.

Esta tesis se desarrolla mediante la modalidad de *compendio de publicaciones*. Por este motivo, en este trabajo se describirán todas las publicaciones y otras actividades relevantes que conforman la tesis. Hay dos líneas de investigación diferenciadas. La primera, relacionada con la revisión por pares, es una línea recientemente desarrollada en el ámbito del grupo de investigación de *Visión por Computador*, del Departamento de Ciencias de la Computación e Inteligencia Artificial (DECSAI) de la Universidad de Granada y una segunda línea, de análisis de palabras clave, que se corresponde a una investigación original desarrollada en el contexto de esta tesis.

A la hora de escoger estas líneas de investigación se primó, por un lado, el enfoque pedagógico que puede tener para un estudiante de doctorado el hecho de integrarse en un grupo de investigación y continuar con su trabajo y, por otro lado, la capacidad de plantear nuevas preguntas e hipótesis que abren camino en el conocimiento científico. En este segundo enfoque debemos

enmarcar el estudio de las palabras clave en la Ciencia. Es una línea que se encuentra escasamente investigada y donde todavía hay mucho que aportar. Es también objetivo de esta tesis el de proponer a la comunidad científica diferentes posibilidades que tiene por delante la Ciencia en este ámbito.

## 1.2. Finalidad y objetivos

El fin último de la tesis doctoral es el de estudiar, a través de modelos matemáticos y computacionales, el comportamiento de la comunidad científica en lo que a producción de documentación científica se refiere. En este trabajo se plantean dos objetivos principales. Por un lado, nos proponemos analizar el comportamiento de autores, editores y revisores en el proceso de revisión por pares, identificando estrategias a seguir y proponiendo modelos para maximizar la recompensa obtenida por los actores que participan en el proceso de producción y validación de trabajos científicos. Por otra parte, queremos proponer modelos y estrategias que permitan maximizar el impacto de un trabajo científico mediante la selección de las palabras clave más apropiadas [1]. Dentro de estos objetivos mencionados, también nos proponemos una serie de subobjetivos que serán especificados a continuación:

1. Analizar el comportamiento de autores, editores y revisores en el proceso de revisión por pares, identificando estrategias a seguir y proponiendo modelos para maximizar la recompensa obtenida por los actores que participan en el proceso de producción y validación de trabajos científicos.

    - Modelar, mediante la Teoría de Juegos, el proceso de sumisión de un artículo científico a una revista, desde la perspectiva de cada uno de los roles que intervienen en el proceso: los autores, los editores y los revisores.
    - Identificar las diferentes estrategias seguidas dentro de cada rol, así como la recompensa generada por cada una.
    - Proporcionar a la comunidad científica herramientas que permitan mejorar su nivel de auto-conocimiento sobre sus comportamientos y estrategias seguidas.

2. Proponer modelos y estrategias que permitan maximizar el impacto de un trabajo científico mediante la selección de las palabras clave más apropiadas.

[1] Si bien en esta tesis hemos orientado nuestro modelo a la categorización de palabras clave en artículos científicos, el modelo propuesto permitiría ser aplicado en otro tipo de documentos o incluso en formatos audiovisuales.

- Poner de relieve la relevancia que tienen las palabras clave en la actualidad.
- Proponer modelos que permitan a los autores mejorar la visibilidad y el impacto de los artículos, mediante una correcta selección de palabras clave de autor.
- Aprovechando las palabras clave como estrategia de búsqueda, crear un modelo de búsqueda y recuperación de información que permita a los usuarios obtener documentos que den respuesta a preguntas complejas.

Por otro lado, en este trabajo se hace una apuesta por la publicación de datos abiertos como forma de contribuir al desarrollo de la ciencia. Por este motivo, los bancos de datos que hemos construido a lo largo de estos años han sido publicados en repositorios de datos abiertos.

## 1.3. Estructura

Como suele ser habitual en los compendios de publicaciones, en este trabajo se describirán todos los artículos publicados desarrollados durante el periodo predoctoral. Como paso previo a esta descripción, he considerado oportuno realizar una introducción teórica a las dos líneas de investigación que estamos tratando, ya que es necesario introducir el contexto teórico en el que se mueve el proceso de revisión por pares, que emana de la idea de comunidad científica que fue fruto de debate en el seno de la Filosofía de la Ciencia durante el siglo pasado. No podemos entender completamente la revisión por pares si no nos paramos a conocer el por qué de su existencia, su rol y su justificación. Para ello, se ha descrito brevemente la postura de diferentes autores relevantes del pasado siglo XX como Karl Popper, Thomas Kuhn, Karl Polanyi, John Ziman o Paul Feyerabend.

De igual manera, considero relevante entender el rol de las palabras clave no solo desde el punto de vista de la cienciometría, o como un mero elemento necesario para recuperar información: las palabras clave también tienen una explicación cultural y son una expresión de nuestra forma de comprender la realidad.

Esta introducción teórica tiene lugar en el segundo capítulo de esta tesis. Mientras que el tercer capítulo, y los siguientes, tienen como fin presentar y explicar de forma práctica la investigación que se ha realizado en esta tesis.

Tabla 1.1: Relación de trabajos publicados y línea de investigación asociada.

| Título | Referencia | Temática | Sección |
|---|---|---|---|
| An evolutionary explanation of assassins and zealots in peer review | Chamorro-Padial et al. (2019) | TICs y Sociología de la Ciencia | Sección 3.1 |
| The author's ignorance on the publication fees is a source of power for publishers | García et al. (2019) | TICs y Sociología de la Ciencia | Sección 3.2 |
| What is the sensitivity and specificity of the peer review process? | García et al. (2022) | TICs y Sociología de la Ciencia | Sección 3.3 |
| The Relevance of Title, Abstract, and Keywords for scientific paper quality and potential impact | Chamorro-Padial y Rodríguez-Sánchez (2023b) | TICs, Sociología de la Ciencia y Palabras Clave | Sección 3.4 |
| Text categorization through dimensionality reduction using Wavelet Transform. | Chamorro-Padial y Rodríguez-Sánchez (2020d) | TICs y Palabras Clave | Sección 4.1 |
| Finding answers to COVID-19-specific questions: An information retrieval system based on latent keywords and adapted TF-IDF | Chamorro-Padial et al. (2022) | TICs y Palabras Clave | Sección 4.2 |
| Attention –Survival Score: A Metric to Choose Better Keywords and Improve Visibility of Information | Chamorro-Padial y Rodríguez-Sánchez (2023a) | Palabras Clave | Sección 4.3 |
| Corner Centrality of Nodes in Multilayer Networks: A Case Study in the Network Analysis of Keywords | Rodríguez-Sánchez y Chamorro-Padial (2022) | TICs y Palabras Clave | Sección 4.4 |
| Clasificación de texto. Utilizando métricas de ganancia de información para categorizar disposiciones legales. | Chamorro-Padial y Rodríguez-Sánchez (2019) | Otros | Sección 5.1 |

# Capítulo 2

# Introducción

## 2.1. La comunidad científica: Abordaje desde la Filosofía de la Ciencia

Con ánimo de realizar una primera toma de contacto al marco teórico en el que se encuadra la presente tesis, nos centraremos, en primer lugar, en la importancia y el rol que tiene la comunidad científica. Para comenzar, parece conveniente introducir el concepto de *comunidad científica*. Para ello, podemos remitirnos a los trabajos desarrollados por *Karl Popper*, *Michael Polanyi*, *John Ziman* así como por los autores influenciados por las corrientes del denominado Racionalismo Crítico.

Para comprender el contexto en el que los diferentes autores han desarrollado sus teorías, debemos entender la *disputa* entre la escuela lógica, defensora de análisis empírico y racionalista, y la escuela socio-histórica, que defiende el estudio de los procesos por encima de los resultados. En el campo de la Filosofía de la ciencia, la escuela lógica argumenta que la metodología de la ciencia debe centrarse en aspectos lógicos y sistemáticos, evitando entrar en el análisis psicológicos e históricos. La corriente socio-histórica, sin embargo, considera clave estudiar estas facetas y dejar en una posición secundaria los principios metodológicos (Munévar, 2005).

Karl Popper es uno de los autores más destacados, si no el que más, dentro de la Escuela Lógica y el Racionalismo Crítico.

Para Karl Raimund Popper (Viena, 28 de julio de 1902 - Londres, 17 de septiembre de 1994), filósofo austriaco impulsor del *falsacionismo*, el papel de los científicos debe ser el de comprobar la falsabilidad [1] y tratar de refutar sus propias teorías mientras continúan el trabajo comenzado por otros científicos, asumiendo su fundamento y preservando, de esta forma, una serie

[1] De acuerdo con el Diccionario de la Real Academia Española, el verbo *falsar* es definido de la siguiente manera: *En la ciencia, desmentir una hipótesis o una teoría mediante pruebas o experimentos.*

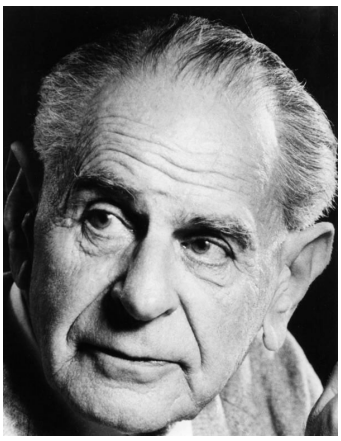


Figura 2.1: Karl Popper. Fuente: *Library of the London School of Economics and Political Science.*

de reglas y comportamientos que forman parte de la *tradición de la ciencia*. La comunidad científica tiene el rol de mostrar los aspectos más importantes de esta tradición científica. Sin embargo, la Ciencia no debe crecer por la mera acumulación de conocimiento, ya que si de esto se tratase, entonces la tradición científica carecería de una gran relevancia. Sin embargo, la ciencia avanza gracias a la *crítica*, que tiene el poder de destruir, modificar y alterar los mitos generados por la ciencia (Popper, 1962).

La Escuela Socio-histórica nos plantea otro punto de vista diferente. Aquí podemos mencionar, por ejemplo, a Thomas Kuhn (Cincinnati, 18 de julio de 1922 - Cambridge, 17 de junio de 1996) quien defiende que la importancia de la comunidad científica radica en los pequeños grupos de investigadores de una especialidad concreta que poseen conocimientos similares y un transfondo de conocimientos parecido. Son estos grupos, y no la comunidad científica en su conjunto, los que se encargan de generar conocimiento y de validarlo.

Por su parte, Paul Feyerabend (Viena, 13 de enero de 1924 - Zúrich, 11 de febrero de 1994) es el representante de una visión más alternativa y crítica con respecto a la ciencia. De acuerdo con el autor, la ciencia debe regirse por principios anarquistas. Basa esta afirmación en la necesidad de mantener la conversación científica sin restricciones y abierta a todas las opciones posibles, evitando cualquier obstrucción al avance de la misma. La ciencia, por lo tanto, debe fundamentarse en el principio de *todo vale* (*anything goes*, en inglés) como vía para no inhibir el progreso. Para Feyarabend, el anarquismo es una *excelente medicina* para la epistemología y la filosofía de la ciencia (Reale y Antiseri, 1988).

La ciencia puede avanzar mediante procesos contraintuitivos. Por ejemplo, se pueden formular hipótesis que contradigan teorías confirmadas y resultados experimentales ya establecidos. Lo que supone crear *contrarreglas*

opuestas a las reglas familiares y aceptadas en la comunidad científica (Feyerabend, 1975). El anarquismo epistemológico de Feyarabend argumenta que toda norma en la ciencia es suceptible de ser quebrantada o simplemente ignorada. Estas violaciones de las normas son necesarias para que se produzca un avance científico.

Feyarabend realiza una crítica abierta a Popper y otros autores que pretenden definir reglas y metodologías que rijan el comportamiento de la comunidad científica (Reale y Antiseri, 1988).

Imre Lakatos (Debrecen, 9 de noviembre de 1922 - Londres, 2 de febrero de 1974) fue un economista y filósofo. En el ámbito de la Filosofía de la Ciencia, destaca por su análisis del progreso y degeneración de la investigación científica. Si bien inicialmente Lakatos fue partidario de los postulados del Racionalismo Crítico, más adelante también adoptó planteamientos de Kuhn.

Para Lakatos, la ciencia puede progresar en un sentido favorable o en uno desfavorable y, por lo tanto, degenerarse. El problema con el que nos enfrentamos es que no tenemos herramientas para evaluar, a priori, si estamos en una tendencia de degeneración o en una de progreso, solamente podemos saberlo a posteriori. En cada campo de la ciencia existen una serie de estándares diferentes que permiten evaluar los resultados obtenidos y si estos son satisfactorios. No obstante, solo aquellos científicos que trabajan en cada uno de estos campos serán los que puedan juzgar si se ha progresado o no, ya que no contamos con estándares externos y generalizables (Munévar, 2005).

Por este motivo, nos vemos forzados a confiar en el criterio de los expertos. Sin embargo, esta situación no es la ideal, ya que confiar en los expertos no evita que el progreso se degenere y la Ciencia corre el riesgo de quedar controlada bajo el timón de una élite estancada (Lakatos, 1978).

Para Karl Polanyi (Viena, 25 de octubre de 1886 - Pickering, 23 de abril de 1964), la Ciencia es el fruto de la actividad científica llevada a cabo por individuos. El resultado de esta actividad es juzgada por una comunidad de científicos (Polanyi, 1962; Overington, 1977). Un hecho a destacar sobre la propuesta de Polanyi es el proceso de aprendizaje de un científico. Para el autor, los científicos aplican habilidades y conocimientos adquiridos mediante un proceso de aprendizaje en el cual las personas nóveles aprenden gracias a la tutela de otros científicos de mayor experiencia que pueden transferirle habilidades y conocimiento sobre las reglas culturales que rigen la actividad científica.

Polanyi sostiene que la comunidad científica es una forma de controlar la producción científica pero también supone un colectivo que tiene unas reglas intrínsecas de funcionamiento. De esta forma, es la comunidad científica la que, de alguna forma, *legisla* sobre la ciencia y la que decide quién puede practicarla.

La comunidad científica, sin embargo, está sometida a controversias, discusiones y cambios que provocan que nuevas ideas, normas y conocimiento pasen a formar parte del consenso de la comunidad científica. Este consenso depende fundamentalmente de tres áreas: El método científico, la educación científica que permite conocer dicho método y los patrones comunicativos existentes entre los científicos.

Finalmente, podemos mencionar a John Ziman (16 de mayo de 1925 - 2 de enero de 2005), la ciencia requiere un proceso de aprendizaje que se adquiere mediante la imitación y la experiencia (Ziman, 1974; Overington, 1977). El objetivo de la actividad científica no es la búsqueda de la verdad; sino del *consenso*, para el cual es necesario la existencia de una comunidad de científicos que acepten o refuten los postulados que se proponen por una parte de esta misma comunidad. Ziman coincide con Polanyi en el carácter altamente persuasivo del método científico como vía hacia el consenso, entendiendo el mismo como un mecanismo de retórica en el que se aportan argumentos de peso que resulten convincentes al resto de la comunidad científica.

En el proceso de aprendizaje de los nuevos científicos, estos adquieren relaciones con otros científicos, construyendo una red social de contactos que, a menudo, se encuentran estudiando los mismos problemas y con los que existe intercambio de información, problemas, resultados... Estos grupos de científicos forman las comunidades en las cuales el consenso científico se construye.

## 2.2. Revisión por pares

### 2.2.1. Evolución histórica del proceso de revisión por pares

La Revisión por pares (*peer review*, en inglés) consiste en la evaluación del trabajo realizado por una o más personas con competencias similares. Los orígenes de la revisión por pares no terminan de estar del todo claros. De acuerdo con (Spier, 2002), encontramos evidencias sobre un proceso de revisión por pares documentado en un libro escrito por *Ishap bin Ali Al Rahwi* donde se describe un código deontológico para registrar y revisar las actuaciones médicas sobre un paciente, de forma que un consejo de expertos pudiera comprobar estas actuaciones y verificar que se ajustan a un estándar concreto, pudiendo un paciente ser indemnizado si había evidencias de mala praxis. Tenemos que saltar al siglo XVII, concretamente a 1645, año en el que un grupo de académicos se reúnen para discutir y debatir opiniones y hallazgos. Este colectivo es el germen de la *Royal Society of London*, que en 1665 crearon la revista científica *Philosophical Transactions*, que aún existe a día de hoy [2]. Las primeras publicaciones en esta revista eran seleccionadas

[2] https://royalsocietypublishing.org/journal/rstl (Accedida el 16 de Abril del 2023).

y revisadas por el editor (o por personas seleccionadas por el editor), quien tomaba la decisión de la publicación en base a criterios estrictamente particulares. En 1731, la *Royal Society of Edinburgh* estableció un procedimiento para revisar obras científicas, que fue adoptado en 1752 por *Philosophical Transactions* (Spier, 2002; Burnham, 1990). Durante el siglo XIX el número de revistas fue creciendo, aunque el proceso de revisión de las obras publicadas en las mismas, generalmente, recaía en la opinión del editor, quien a veces podía solicitar el asesoramiento de comités de especialistas de la misma Sociedad Científica a la que pertenecía la revista. La dificultad de la época para replicar obras también era un factor limitante a la hora de establecer el número de personas que podían revisar una obra, ya que por lo general, solamente se producían entre tres y cinco copias de un manuscrito. Ya en esta época, la mayoría de las revisiones se realizaban sin que el autor pudiera conocer a las personas que habían revisado su trabajo. Si bien la revisión de doble ciego es una práctica más moderna.

Sin embargo, en aquel momento, la especialización y la diversidad de las temáticas a tratar estaban en considerable aumento, haciendo cada vez más complicada la tarea de evaluar una obra por parte de un editor. Esta situación fuerza a los editores a requerir asistencia especializada más allá de la misma Sociedad Científica. Este proceso, que se va adoptando de una manera lenta y desigual, se extiende hasta mediados del siglo XX, cuando el nacimiento de las fotocopiadoras facilitó considerablemente la tarea de replicar manuscritos (Kronick, 1990; Burnham, 1990).

También en este mismo siglo asistimos a un cambio en la composición de la comunidad científica: mientras en siglos anteriores solamente un reducido número de personas, generalmente en posiciones económicas o sociales privilegiadas, contribuían al avance científico; en el siglo XX el número de personas involucradas en la investigación científica se dispara a nivel global, lo que incrementa enormemente la producción de manuscritos y obliga a las revistas a ser más estrictas con aquello que se publica.

Esta revolución generada en el siglo XX se vuelve a repetir con la llegada de Internet. Ahora la distribución de contenido ya no depende, necesariamente, de formatos impresos y puede distribuirse a todo el planeta, lo que facilita la tarea de encontrar expertos que puedan revisar artículos (Spier, 2002).

### 2.2.2. La revisión por pares en la actualidad

La revisión por pares es un mecanismo para garantizar la calidad de un manuscrito. Generalmente, la revisión de una obra es responsabilidad de un grupo de investigadores independientes que tienen un conocimiento contrastado en el área de conocimiento del manuscrito a revisar. No es posible establecer un único proceso de revisión por pares, No obstante, y siempre hablando en términos generales, la figura Figura 2.2 ilustra de manera de-

tallada todas las etapas del proceso, que se pueden resumir en las siguientes (Adhikari, 2021; BioMed Central, 2023):

1. El autor (o autores) envía su manuscrito a una revista.

2. El editor de la revista comprueba que el manuscrito se ajusta al contenido de la revista y que cumple unos criterios mínimos de calidad. Aquí se toma una primera decisión, ya que el manuscrito puede ser rechazado o enviado a revisión. En ocasiones, el artículo puede ser transferido a otra revista donde tenga un mayor encaje.

3. Si el artículo es enviado a revisión, entonces se selecciona a un grupo de revisores que escriben un informe, cada uno, con comentarios y una valoración general sobre la calidad del artículo.

4. Con esta información, el editor puede tomar una decisión:

    - Aceptar el artículo para que sea publicado.
    - Rechazar el artículo.
    - Enviar los comentarios al autor y solicitarle que realice modificaciones o correcciones al mismo. En este caso, se vuelve a iniciar un proceso de revisión por pares, una vez que el autor realice las modificaciones solicitadas.

Generalmente, la tasa de rechazo de artículos es bastante elevada, no siendo raro que esta supere el 50 % en revistas reputadas (Garcia-Costa et al., 2022).

En cuanto a la revisión por pares, podemos destacar diferentes tipos (Emile et al., 2022) [3]:

- **Ciego sencillo** (*single-blind*, en inglés): El revisor sabe el nombre de los autores a los que está revisando, pero los autores no saben quién es su revisor.

- **Doble ciego** (*double-bind*): Ni los revisores conocen a los autores, ni los autores saben quién les está revisando su obra.

- **Revisión abierta** (*Open peer review*): Los revisores saben quiénes son los autores, y los autores conocen el nombre de los revisores.

[3] La lista de tipos de revisión por pares descrita en este trabajo no pretende ser absoluta, sino indicar algunas de las metodologías de revisión más comunes. Existen otras aproximaciones para realizar revisión por pares. Por ejemplo, se puede consultar la web https://authorservices.wiley.com/Reviewers/journal-reviewers/what-is-peer-review/types-of-peer-review.html (Accedida el 7 de abril del 2023) o (Emile et al., 2022) para ampliar esta información.

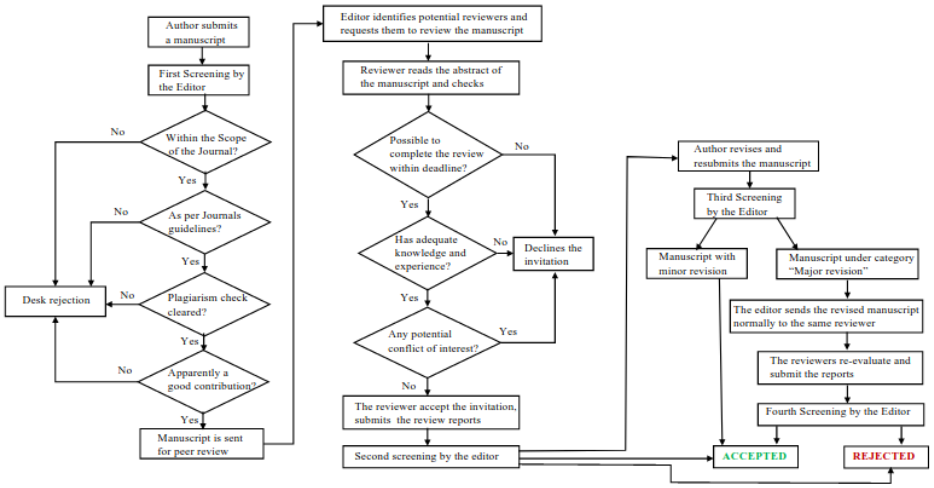



Figura 2.2: Diferentes etapas del proceso de revisión por pares. Fuente: (Adhikari, 2021)

Tabla 2.1: Diferentes tipos de sesgos que pueden afectar a la revisión por pares, junto con ejemplos de cada uno de ellos. Fuente: (Newton, 2010).

| Tipo de sesgo | Definición | Ejemplo |
|---|---|---|
| Situacional | Influenciado por la revista | Los articulos con un menor impacto *potencial* pueden ser más proclives a sufrir un rechazo en el proceso de revisión por pares, frente a trabajos que prevean atraer un mayor impacto a la revista. |
| Social | Normas sociales y culturales | Los revisores de países occidentales tienden a comportarse de manera más *persecutoria*, debido a factores culturales. |
| Personal | Contexto personal del revisor | Los revisores no siempre tienen el conocimiento adecuado requerido para realizar la revisión de un artículo. |
| Ético | Comportamientos adecuados o inadecuados | A veces, los revisores tienen conflictos de intereses que afectan al proceso de revisión. |

- **Revisión transparente** (*Transparent peer review*): Si el manuscrito es publicado, el nombre de los revisores es desvelado.
- **Colaborativa**: Bien los autores, bien los revisores, o bien autores conjuntamente con los revisores, trabajan colaborativamente en la revisión de un manuscrito.

Si bien el proceso de revisión por pares ha permitido a la comunidad científica garantizar unos criterios de calidad mínimos, no es un proceso infalible ni desprovisto de errores que, en ocasiones, dan lugar a serios problemas. La revisión por pares se realiza por humanos que no están exentos de sesgos que les alejan de realizar valoraciones objetivas. (Newton, 2010), en su trabajo, revisa diferentes causas que afectan al proceso de revisión. Algunos sesgos importantes que afectan a la revisión se han detallado en la Tabla 2.1.

En el año 2005, SCIgen (Stribling et al., 2005), un generador automático

de papers, fue protagonista de una gran polémica; ya que consiguió que varios artículos generados automáticamente por un ordenador superasen el filtro de calidad de algunas revistas y congresos. La revisión por pares evitó la publicación en muchas ocasiones, aunque algunos artículos que no tenían sentido alguno fueron recomendados para su publicación (Labbé y Labbé, 2013; Stribling et al., 2005).

Con el objetivo de incrementar la calidad del proceso de revisión de manuscritos, han surgido metodologías que tratan de añadir objetividad y establecer metodologías claras a la hora de abordar una revisión por pares (PRISMA, CARE, SQUIRE, CONSORT, etc) (Berman et al., 2017).

La revisión por pares es, por lo general, una actividad no remunerada de forma económica. Si bien los revisores pueden recibir certificados o reconocimientos de su labor. Esta situación genera cierto debate en la comunidad académica. Quienes se muestran a favor de que los revisores no sean remunerados argumentan que puede ponerse en peligro la *búsqueda del conocimiento* como razón de ser de la Academia, generar un incremento de costes en el proceso de investigación o que la remuneración de los revisores no soluciona los problemas reales de la revisión por pares (Moustafa, 2022; Arora y Arora, 2022; Vines y Muddit, 2021). En el lado contrario, hay académicos que sostienen que la revisión debe ser una actividad remunerada como cualquier otro proceso involucrado en la investigación a nivel profesional, y que una remuneración aceleraría el proceso de revisión de los artículos (Flaherty, 2022).

Mientras la comunidad debate sobre la remuneración de los revisores, algunos autores optan por aportar datos y estudiar el coste de la revisión por pares. Si bien es complicado dar una cifra precisa, podemos mencionar dos estudios diferentes: En primer lugar, Research Information Network (RIN) [4], una institución británica dedicada al análisis de la comunidad académica, estimó que el coste asumido por el tejido académico del Reino Unido en los procesos de revisión por pares ascendió a 165 millones de libras durante el año 2008. En el año del informe, el 7,1 % de los artículos publicados a nivel global habían sido revisados en el Reino Unido (Jubb, 2008). Por otro lado, tenemos un estudio más reciente, publicado en el año 2021, que cifra en 1.510 millones de dólares el coste de la revisión por pares en Estados Unidos durante el año 2020, mientras que ese mismo año, China asumió un gasto de 626 millones de dólares y el Reino Unido, 391 millones de dólares. Más allá de las cifras económicas, el año en el asistimos al despertar la pandemia del SARS-CoV-2, hubo cerca de 22 millones de revisiones que requirieron, en total, 130 millones de horas de trabajo (Aczel et al., 2021).

Sea como sea, hoy en día la revisión por pares tiene una importancia vital en la Academia, siendo un requisito fundamental para que una publicación

[4]La web de RIN, según figura en el informe citado, ya no existe. En la actualidad, debemos dirigirnos a https://www.researchinfonet.org/ (Accedida el 7 de Abril del 2023).

científica tenga credibilidad por parte de la comunidad científica. Este modelo de revisión por pares se pone en cuestión en las denominadas revistas parasitarias (*predatory journals*, en inglés). Entre otros factores, son revistas acusadas de no realizar revisión por pares o de realizar un proceso de revisión muy relajado, priorizando la publicación a toda costa frente a la calidad de un trabajo o su aportación a la ciencia. Generalmente, este tipo de revistas cobran una tasa de publicación al autor, siendo su modelo de negocio el conseguir ingresos a toda costa (Cobey et al., 2018) [5].

## 2.3. Las palabras clave: definición y ámbitos de estudio

Antes de sumergirnos en el mundo de las palabras clave en el ámbito de la ciencia es deseable, en primer lugar, realizar una breve introducción a las palabras clave desde una perspectiva más general.

La *International Organization for Standarization* (ISO), define el concepto de *palabra clave* de la siguiente manera (International Organization for Standardization, 1985):

> "A word or group of words, possibly in a lexicographically standardized form, taken out of a title or the text of a document characterizing its content and enabling its retrieval."

Así pues, las palabras clave permiten identificar conceptos esenciales. A pesar de que la ISO hace referencia a textos, las palabras clave pueden utilizarse también en obras de diversos formatos tales como imágenes, vídeos o audio [6].

Por su parte, la Real Academia de la Lengua Española hace también énfasis en la utilidad de las palabras clave como medio para almacenar y recuperar información almacenada en una base de datos, proponiendo la siguiente definición (Real Academia Española, 2021):

> "f. *Inform.* palabra significativa o informativa sobre el contenido de un documento, que se utiliza habitualmente para su localización y recuperación en una base de datos."

Si bien las palabras clave se utilizan en el ámbito de la recuperación de la información (como también explicaremos más adelante), el abanico de usos

[5] Existen iniciativas, como *Beall's list* https://beallslist.net/ (Accedida el 7 de Abril del 2023) que tratan de identificar las revistas parasitarias, para evitar que los autores envíen sus obras por error a este tipo de publicaciones.

[6] Parte de la aportación realizada en esta tesis consiste en la propuesta de un método que permite clasificar mediante palabras clave contenido, independientemente del formato en el que este se presente.

que podemos encontrar no termina aquí, y es que las palabras clave pueden ser estudiadas desde diferentes perspectivas.

Por ejemplo, las palabras clave también juegan un papel importante en el ámbito de la lingüística y, concretamente, en el Análisis del Discurso. Dentro de este ámbito, podemos ver cómo las palabras clave suponen un objeto de estudio que permite crear indicadores y delimitar el ámbito de una disciplina. Un buen ejemplo de ello lo podemos encontrar en (Sánchez-Saus Laserna, 2018), donde se utilizan las palabras clave como métricas para evaluar el Discurso generado en redes sociales en torno a la *Comunicación para el Desarrollo y el Cambio Social* (CDCS). Por su parte, (Duque, 2015) se vale de las palabras clave para realizar un análisis de contenido con el objetivo de perfilar los diferentes tipos de participantes representados en el discurso político. Este tipo de análisis se basan en la asunción de que existe una relación entre la frecuencia de una palabra determinada y su relevancia en la significatividad del conocimiento que se desea transmitir en un corpus. Bajo esta premisa se realiza una extracción de palabras clave del corpus, catalogando como tal a aquellas palabras cuya frecuencia de aparición son excepcionalmente más elevadas que lo que le correspondería según la distribución de palabras clave en un idioma o en comparación con otro corpus [7].

En este punto, es necesario realizar una puntualización con respecto a las palabras clave y la frecuencia de aparición de las mismas. Las palabras que aparecen en un corpus con más frecuencia no son, necesariamente, palabras clave. De hecho, generalmente serán palabras que aporten poca información y que frecuentemente se tienden a eliminar o a aplicar sobre ellas algún tipo de penalización a la hora de realizar análisis de un texto (Sarica y Luo, 2021). Las palabras que son verdaderamente informativas, y que por lo tanto son candidatas a ser consideradas como palabras clave, son aquellas que tienen una frecuencia de aparición que no está acorde con la distribución considerada normal y que pueden proporcionar información sobre la estructura de un documento. Esta idea es el fundamento de medidas ampliamente utilizadas por la comunidad científica en diferentes ámbitos del análisis de textos. Un ejemplo de ello es TF-IDF (Aizawa, 2003).

Frente a descripciones técnicas, también podemos hablar de palabras clave desde un punto de vista cultural. *Keywords: A Vocabulary of Culture and Society* (Williams, 1985) es un libro escrito por Raymond Williams y considerado uno de los trabajos clave a la hora de analizar el componente cultural de las palabras clave. De acuerdo con Williams, las palabras clave tienen en común, en primer lugar, ser palabras clave relevantes para una actividad concreta y, en segundo lugar, ser también importantes para definir

[7] Generalmente, se utilizan corpus con un vocabulario amplio y de ámbito general. Un ejemplo de ello puede ser el *Corpus de Referencia del Español Actual* (CREA), que puede consultarse en: https://corpus.rae.es/creanet.html (Accedida el 8 de Abril del 2023).

un sistema de pensamiento concreto. La obra de Williams estudia el idioma inglés, y analiza la evolución de las palabras clave desde sus orígenes en el latín y, posteriormente, el francés hasta llegar a la actualidad. En este tiempo, las palabras clave han tenido diferentes significados.

Como podemos ver, las palabras clave, pese a ser unidades elementales de información significativa, no están exentas de tener diferentes interpretaciones. De la misma manera que dos personas pueden interpretar una misma palabra clave con significados diferentes, estas dos personas también podrían asignar palabras clave diferentes a una misma obra (Strader, 2011; Whittaker, 1989). Y es que, de acuerdo con (Scott y Tribble, 2006), las palabras clave son un reflejo de los procesos de razonamiento internos de una persona.

Y pese a las diferencias interpretativas y culturales que acabamos de comentar, las palabras clave también pueden servir como puente que facilite la comunicación entre personas. Esto es lo que defiende la teoría del *Metalenguaje Semántico Natural* (en inglés: *Natural Semantic Metalanguage*, identificado normalmente con las siglas NSM) (Goddard y Wierzbicka, 2002; Wierzbicka, 1996), desarrollada por la lingüista *Anna Wierzbicka* a partir del trabajo de *Andrzej Boguslawski* y que defiende que es posible reducir el lexicón a un conjunto de vocabulario reducido llamado *primitivas semánticas*, que se caracterizan por ser universales al tener la misma traducción en cada idioma y por poder ser definidas utilizando diferentes palabras [8]. Estas primitivas universales permiten la comunicación entre personas, sin estar sometidas a sesgos culturales o lingüísticos (Wierzbicka, 2010).

En definitiva, el estudio de las palabras clave es muy diverso y puede ser abordado desde diferentes disciplinas. En esta tesis, nos centraremos en la relación entre las palabras claves y la producción de literatura científica por parte de la Academia.

## 2.4. Compendio de publicaciones

Como hemos comentado anteriormente, la tesis se publica en el formato de compendio de publicaciones. En los siguientes capítulos, se irán introduciendo, uno a uno, todos los artículos publicados amén de los repositorios de datos y otras obras generadas a raíz de la investigación que se ha ido de desarrollando estos años.

Junto con cada artículo, se incluye una pequeña sección que resume el contenido esencial del mismo, las contribuciones realizadas a la ciencia y algunas métricas sobre el artículo y la revista donde se encuentre publicado. Acto seguido, se presenta una copia del paper en formato prepublicación.

Comenzaremos, en primer lugar, por los artículos relacionados con la línea de investigación sobre revisión por pares, para continuar por la línea de

[8] Por ejemplo, puedo utilizar la palabra *crear* o *realizar* para referirme a la misma primitiva semántica (Minini, 2013).

las palabras clave. Finalmente, se incluye un capítulo dedicado a otro tipo de publicaciones.

A continuación, y antes de continuar con el siguiente capítulo de la tesis, proponemos un acercamiento a los conocimientos teóricos que hemos utilizado para las publicaciones de esta tesis.

## 2.5. Conocimientos previos

En esta sección se detalla la base teórica sobre la que se ha fundamentado este trabajo. Concretamente, introducimos aquí el modelo cuasiespecie (Subsección 2.5.1) y de la transformada wavelet (Subsección 2.5.2). En cuanto al modelo de cuasiespecie, hemos hecho uso de esta herramienta para caracterizar el comportamiento de autores, editores y revisores durante el proceso se revisión por pares. Por su parte, la transformada wavelet nos permite extraer información de documentos mediante la aplicación de principios del procesamiento de señales.

### 2.5.1. El modelo de cuasiespecie

El modelo de cuasiespecie (también denominado *modelo de cuasiespecies*) nace en el campo de la biología evolutiva. De acuerdo con (Eigen y Schuster, 1977), el modelo de cuasiespecie representa el proceso de la evolución darwiniana de un determinado tipo de entidades auto replicativas en un entorno caracterizado por un alto ratio de mutación. Podemos definir una cuasi-especie como una distribución de especies macro moleculares con secuencias fuertemente interrelacionadas y dominadas por una o varias copias maestras.

Los factores externos del entorno obligan a la selección de la distribución mejor adaptada, es decir, de la cuasiespecie más apta. La estabilidad interna depende de una serie de criterios que, si no se cumplen, conducen a la desintegración de la información de la secuencia de la copia maestra.

Más allá de su área de nacimiento, el modelo de cuasiespecie ha sido adaptado por autores de otros campos. Por ejemplo, la economista Friederike Mengel aplica este modelo en el ámbito de la Teoría de Juegos con el fin de estudiar el comportamiento evolutivo de la cultura humana y, concretamente, del razonamiento categórico (Mengel, 2012). De acuerdo con la autora, se puede entender una cultura como una partición de un conjunto de problemas de toma de decisiones ante diferentes situaciones. El comportamiento ante cada situación se transmite entre generaciones, pero esta transmisión de conocimiento está sujeta a cambios producidos por errores.

El modelo de cuasiespecie es el punto de partida de una de las líneas de investigación en la que se centra esta tesis: la revisión por pares.

### 2.5.2. La Transformada Wavelet

En el campo del procesado de señales, en 1984, *Jean Morlet* y *Alex Grossman* introdujeron el concepto de *wavelet* (traducido al castellano como *ondícula* u *ondita*) [9] [10]. Una transformada wavelet se caracteriza por representar la señal conservando la información sobre la escala y sobre la información espacial (Castro y Castro, 1995; Debnath, 2002). A diferencia de la célebre Transformada de Fourier, la Transformada Wavelet nos permite obtener información localizada a nivel espacial, mientras que Fourier es práctica cuando se pretende realizar un análisis de la señal a nivel global.

La idea intutiva tras la Teoría Wavelet consiste en realizar un análisis de la señal a diferentes escalas (o resoluciones). Cuando trabajamos con una señal a gran escala, no podemos observar detalles que sí veremos al reducir la escala. En el análisis Wavelet se parte de una función denominada wavelet madre. Existen multitud de wavelets madres bien conocidas. Tanto es así, que podemos agruparlas por familias (Por ejemplo: Haar, Daubechies, Biortogonales, Sombrero Mexicano, etc) (Sujatha y Devi, 2015). Esta función es sometida a diferentes transformaciones de traslación y dilatación para representar la señal que se quiere tratar (Debnath, 2002).

#### 2.5.2.1. Transformada Wavelet Continua

Antes de definir la Transformada Wavelet Continua (TWD) debemos introducir el concepto de transformada wavelet. Siendo $g(t)$ una función, para que pueda ser considerada wavelet madre, $\psi(t)$ ($g(t) = \psi(t)$) debe cumplir con dos propiedades básicas (Osorio-Sánchez, 2006):

$$\int_{-\infty}^{+\infty} g(t)dt = 0$$

$$\int_{-\infty}^{+\infty} g^2(t)dt = 1$$

Podemos representar el escalado o dilatación de $g$ por una constante $a$ de la siguiente manera:

$$g_a(t) = g(\frac{t}{a})$$

[9] Aunque, de forma general, en el conjunto de esta tesis se opta por utilizar nombres en castellano de manera preferente, en esta ocasión utilizaremos la denominación inglesa (wavelet) ya que es el concepto mayoritariamente utilizado en terminos académicos, incluso en trabajos escritos en lengua española.

[10] Si bien el concepto fue introducido en el año 1984, ya en 1909 Alfred Haar propuso la transformada Haar haciendo uso de la hoy llamada wavelet Haar. Sin embargo; en esa época, wavelet, como concepto, aún no existía (Haar, 1910).

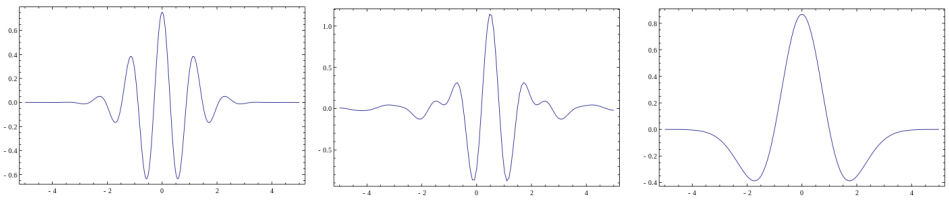


Figura 2.3: Diferentes ejemplos de funciones wavelet. De izquierda a derecha: *Morlet Wavelet, Meyer Wavelet* y wavelet de *Sombrero Mexicano*. Autor: Jon McLoone (Licencia: *Creative Commons Attribution-ShareAlike 3.0 Unported*).

Del mismo modo, podemos aplicar una traslación de $g$ por una constante $b$ mediante la expresión:

$$g^b(t) = g(t - b)$$

Podemos representar el escalado o dilatación de $g$ por una constante $a$, y la traslación por una constante $b$ aplicando la expresión:

$$g_a^b(t) = g(\frac{t-b}{a})$$

La transformada wavelet continua (CWT) se define mediante la función:

$$TWC(b,a) = \frac{1}{sqrt(|a|)} \int_{-\infty}^{+\infty} g(t)\psi^* \left(\frac{1}{a}(t-b)\right) dt$$

Esta función puede discretizarse, dando lugar a la Transformada Wavelet Discreta (TWD) (Osorio-Sánchez, 2006).

#### 2.5.2.2. Análisis multiescala utilizando la Transformada Wavelet Discreta

Sea $f(n)$ una función continua, definimos $A(f(n))$ y $D(f(n))$ de la siguiente manera [11]:

$$A(f(n)) = \frac{f(2n-1) + f(2n)}{\sqrt{(2)}}$$

$$D(f(n)) = \frac{f(2n-1) - f(2n)}{\sqrt{(2)}}$$

Siendo $A(f(n))$ la señal paso bajo y $D(f(n))$ la señal paso alto. Nótese que, a partir de este punto, tenemos dos funciones; cada una con la mitad de

[11] Por simplicidad, se describe en esta sección la transformada Haar (Daubechies n=2). Filtros de mayor complejidad (con n>2) afectarían a más elementos de $f(n)$.

valores que $f(n)$. Podemos reconstruir la función original mediante la suma de ambas funciones:

$$F(n) = A(f(n)) + D(f(n))$$

Podemos definir, a continuación, las matrices $V_n^1$ y $W_n^1$:

$$V_n^1 = \begin{pmatrix} \frac{1}{sqrt(2)} & \frac{1}{sqrt(2)} & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{sqrt(2)} & \frac{1}{sqrt(2)} & \dots & 0 & 0 \\ & & \ddots & & & \\ 0 & 0 & 0 & \dots & \frac{1}{sqrt(2)} & \frac{1}{sqrt(2)} \end{pmatrix}$$

$$W_n^1 = \begin{pmatrix} \frac{1}{sqrt(2)} & \frac{-1}{sqrt(2)} & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{sqrt(2)} & \frac{-1}{sqrt(2)} & \dots & 0 & 0 \\ & & \ddots & & & \\ 0 & 0 & 0 & \dots & \frac{1}{sqrt(2)} & \frac{-1}{sqrt(2)} \end{pmatrix}$$

Podemos obtener $A^1$ y $D^1$ de la siguiente manera:

$$A_n = \left[A^1\right]\left[V_n^1\right]$$

$$D_n = \left[D^1\right]\left[W_n^1\right]$$

Siendo $A^1 = A(f(n))$ y $D^1 = D(f(n))$.

A partir de aquí, podemos realizar un análisis en diferentes escalas separando la señal en altas y bajas frecuencias y realizando un muestreo modificando la escala.

# Capítulo 3

# Revisión por pares

## 3.1. An evolutionary explanation of assassins and zealots in peer review

### 3.1.1. Datos generales

1. **Autores**: Jorge Chamorro-Padial, Rosa Rodríguez-Sánchez, J. Fdez-Valdivia y J.A García.
2. **Revista**: Scientometrics.
3. **Datos sobre la publicación**:
   - **Referencia**: Chamorro-Padial et al. (2019).
   - **Volumen**: 120.
   - **Número**: 3.
   - **Páginas**: 1373–1385.
   - **Año**: 2019.
   - **Editorial**: Springer.
   - **DOI**: https://doi.org/10.1007/s11192-019-03171-3.
4. **Estado**: Publicado.
5. **Métricas**:
   - **Factor de impacto**: 3,801 (JCR, 2021).
   - **Ranking**:
     - *Social Science Citation Index (SSCI)*:
       - *Information Science & Library Science*: Q1 - 29/84 (año 2021).

  - *Information Science & Library Science*: Q1 - 21/87 (año 2019).
- *Social Science Citation Index Expanded (SSCIE)*:
  - *Computer Science, interdisciplinary applications*: Q2 - 54/112 (año 2021).
  - *Computer Science, interdisciplinary applications*: Q2 - 45/109 (año 2019).

### 3.1.2. Contribuciones principales

1. Definición y propuesta de un modelo experimental que permite explicar los diferentes tipos de comportamiento entre revisores y la estrategia que mejor conviene en cada situación.

2. Cuando un revisor es capaz de distinguir correctamente la calidad de un artículo, su recompensa será óptima si el ratio de errores es pequeño y la frecuencia de revisión de artículos de buena y baja calidad es similar (en torno al 50 % de cada tipo de artículo). En caso contrario, los revisores que no sepan distinguir artículos de buena calidad de artículos de baja calidad, tendrán mejor recompensa.

3. Si las revistas académicas siguen una política estricta de rechazar artículos antes de enviarlos a revisión (*desk reject*), se favorece llegar a un punto de equilibrio entre artículos de baja calidad y artículos de alta calidad que son sometidos a revisión. lo que reduce la aparición de revisores *asesinos* o *fanáticos*.

### 3.1.3. Resumen

Dentro del proceso de revisión por pares que hemos comentado en la Sección 2.2, (Siegelman, 1991) hace una distinción entre revisores según su perfil de comportamiento. En los casos más extremos, por un lado, encontramos a los revisores fanáticos (*zealots*), que generalmente tienden a favorecer la decisión de publicar un manuscrito. Y los revisores asesinos (*assassins*), que se muestran más proclives a proporcionar evaluaciones pobres de los documentos que pasan por sus manos. Entre ambos polos, nos encontramos al perfil mayoritario de revisores (*mainstreamers*).

Los autores escriben manuscritos que tienen un cierto nivel de calidad, podemos distinguir entre trabajos de baja calidad y trabajos de alta calidad. Del mismo modo, los revisores pueden ser capaces de distinguir un artículo de alta calidad de uno de baja calidad, en cuyo caso diremos que este revisor es capaz de hacer una partición fina de artículos (*fine partition*), o no ser capaz de distinguir entre artículos de alta y de baja calidad, siendo el revisor capaz de realizar particiones gruesas (*coarse partition*).

El modelo de cuasiespecie (Ver Subsección 2.5.1) expresa de una forma bastante natural el comportamiento de los revisores fanáticos y de los asesinos. En nuestro trabajo, mostramos que, dada una población concreta de revisores, y una cantidad de trabajos académicos cuyo reparto de calidades se encuentra en equilibrio entre manuscritos de alta y de baja calidad, los revisores que tienen una partición fina, obtienen mayor recompensa que aquellos que tienen una partición gruesa. Sin embargo, si el reparto de calidades de la colección de documentos a revisar se encuentra en desequilibrio, los revisores con una partición gruesa obtendrán una mayor recompensa.

Cada revisor tiene un perfil de recomendación. El perfil de recomendación de un revisor que no distinga la calidad de un artículo (esto es, que tenga una partición gruesa) estará formado por dos únicas acciones posibles (aceptar o rechazar un paper). A su vez, el perfil de recomendación de un revisor cuya capacidad de distinción de la calidad de un artículo se describa mediante la partición fina, tendrá cuatro posibilidades diferentes (por ejemplo: un revisor puede identificar correctamente un trabajo de baja calidad, pero recomendar su publicación y a la vez puede identificar artículos de alta calidad e, igualmente, recomendarlos) De acuerdo con el modelo de cuasi especie, y siguiendo el trabajo de (Mengel, 2012), los perfiles de recomendación son asimilables al concepto de entidades auto replicantes. Estas entidades se encuentra inmersas en un entorno donde se producen un gran número de mutaciones, ya que los revisores son humanos y, por ende, están sujetos a errores en sus acciones. Siempre tendremos un margen de error y no siempre un revisor que identifique correctamente, por regla general, la calidad de un artículo, acertará con cualquier manuscrito que deba revisar.

Estos errores humanos provocan mutaciones en los perfiles de recomendación de un revisor, el éxito de la supervivencia de estos nuevos perfiles de recomendación dependerán de la recompensa que sean capaces de obtener.

De acuerdo con los resultados arrojados por el desarrollo del modelo de cuasi especie, un perfil de recomendación basado en la partición fina puede conseguir la mejor recompensa si se cumplen dos condiciones:

- Un ratio de error suficientemente bajo.
- Existe equilibro entre el número de artículos de calidad alta y el número de artículos de calidad baja.

Si una de estas dos condiciones no se cumple, entonces un perfil de recomendación basado en la partición gruesa obtendrá, potencialmente, una mayor recompensa.

A la luz de estos resultados, resulta conveniente la estrategia que aplican la mayor parte de revistas académicas de imponer un filtro estricto en primera instancia, cuando un artículo es enviado a una revista, produciendo una gran cantidad de rechazos antes de enviar a revisión por pares (*desk*

*rejection*), ya que esta estrategia facilitará que los revisores con un perfil de recomendación basado en la partición fina obtengan mejores recompensas que aquellos con partición gruesa. Lo que contribuirá a reducir la aparición de revisores *asesinos* o *fanáticos*.

**Scientometrics manuscript No.**
(will be inserted by the editor)

# An evolutionary explanation of assassins and zealots in peer review

**Jorge Chamorro-Padial, Rosa Rodriguez-Sánchez, J. Fdez-Valdivia, and J.A. Garcia**

the date of receipt and acceptance should be inserted later

**Abstract** The peer review system aims to be effective in separating unacceptable from acceptable manuscripts. However, a reviewer can distinguish them or not. If reviewers distinguish unacceptable from acceptable manuscripts they use a fine partition of categories. But, if reviewers do not distinguish them they use a coarse partition in the evaluation of manuscripts. Most reviewers learned how to evaluate a manuscript from good and bad experiences, and they have been characterized as zealots (who uncritically favor a manuscript), assassins (who advise rejection much more frequently than the norm), and mainstream referees. In this paper we use the

J. A. Garcia, Rosa Rodriguez-Sánchez, and J. Fdez-Valdivia,
Departamento de Ciencias de la Computación e I. A., CITIC-UGR, Universidad de Granada, 18071 Granada, Spain.
Jorge Chamorro-Padial,
CITIC-UGR, Universidad de Granada, 18071 Granada, Spain.
Address correspondence to J. A. Garcia at jags@decsai.ugr.es
Jose A. Garcia ORCID iD `https://orcid.org/0000-0001-7742-7270`

quasi-species model to describe the evolution of recommendation profiles in peer review. A recommendation profile is composed of a reviewer recommendation for each manuscript category under a particular categorization of manuscripts (fine or coarse). We see the reviewer mind as being built up with recommendation profiles. Assassins, zealots and mainstream reviewers are "ecologically" interrelated species whose progeny tend to mutate through errors made in the process of reviewer training. We define the recommendation profile as replicator, and selection arises because different types of recommendation profiles tend to replicate at different rates. Our results help to explain why assassins and zealots evolutionary appear in peer review because of the evolutionary success of reviewers who do not distinguish acceptable and unacceptable manuscripts.

**Keywords**: Peer Review; Reviewers; Assassins; Zealots; Manuscript Categories; Quasi-species.

**1 Introduction**

The scientific community requires the review of manuscripts by peers, and it is assumed that they will be selected for publication based on their merits and suitability. After the research work has been received, the journal editor assigns a handling editor that is generally responsible for the review management. In some academic journals, multiple independent editors or editorial boards operate the journal and decide on the manuscript acceptance or rejection. In this process, the reviewers guarantee the journal quality by identifying the acceptable research, (Burnham, 1990; Campanario, 1998a,b).

Of course, peer review is a complex process in which several experts evaluate every submission on behalf of the editors. However, the role of the reviewers is advisory, and the editor of the journal usually has no formal obligation to accept their recommendations. In general, they look for novelty, quality and the suitability for publication. The mission of this peer review process is to seek promising manuscripts. Review processes become an institutionalized surveillance system of scientific communities and the system aims to be effective in separating acceptable manuscripts from unacceptable ones, (Merton, 1973). There the peer reviewer must think with the aid of manuscripts' categories (e.g., acceptable or unacceptable), and we cannot possibly avoid this categorical reasoning in the peer review process, (Burnham, 1990; Tenopir and King, 2007).

However, the peer review system is more a culture than a method that can be evaluated and the grade of efficacy of the peer reviewers depends on the relevance of peer-review categorization and why it is evolutionary successful to categorize in a particular way (Merton, 1973; Rodriguez-Sanchez et al., 2016). Most reviewers learned how to evaluate a manuscript from good and bad experiences, and the important role played by expert opinion in the manuscript evaluation may explain why peer review has generally not been taught, (Garcia et al., 2015b, 2016). As suggested by Souder (2011), "peer review is inherently ideological: no amount of scientific training will completely mask the human impulses to partisanship." In fact, many hypothesized forms of bias at the peer review stage are real (Burnham, 1990; Chubin & Hackett, 1990; Souder, 2011; Lee et al., 2013; Garcia et al., 2016).

A manuscript needs to be a good fit for the journal and so we think of peer review's culture as partitioning a set of manuscripts into categories (of acceptable or unacceptable) treated the same (accept or reject). Trying to publish a

manuscript in a peer-reviewed journal of high impact can be a difficult and frustrating experience for authors. Editors try to preserve the journal standards or even improve them so that the journal can rise in the rankings. Then, manuscripts that are relevant to the scope of the journal, are innovative, significantly advance the field, are well written have a higher probability of being accepted. (Garcia et al., 2015a) provides a formal study on manuscript quality control and shows that the effects of editors' bias on authors' satisfaction and motivation cause sorting in the authors who submit manuscripts to scholarly journals, and therefore, match authors and journals with similar quality standards. (Garcia et al., 2015a) also shows that some journals will be forced to lower the quality standards in order to be able to compete with journals of more biased editors. (Rodriguez-Sanchez et al., 2016) studies the evolution of manuscript quality control between authors and their editors, using evolutionary games. Within these games, with a certain probability, authors prefer to submit manuscripts of low- or high-quality, and editors prefer to accept low- or high-quality manuscripts. The frequency with which authors (editors) choose to submit (accept) high-quality or low-quality manuscripts change over time in response to the decisions made by all authors and editors in the respective populations. Using this dynamical structure, (Rodriguez-Sanchez et al., 2016) studies which strategies become extinct and which survive, as well as whether the system approaches some stable end-point. For instance, when there is a reduction in quality of reviewer's recommendations in the manuscript evaluation process, it is much less clear why the categorization of acceptable and unacceptable manuscripts should provide higher evolutionary reward.

Reviewers may favor a submission or find flaws in the methodology, results or discussion and conclude that the manuscript is invalid. In fact, reviewers have

been characterized as zealots (a referee that may uncritically favor a manuscript), assassins (a referee with stringent standards who advises rejection much more frequently than the norm), and mainstream referees (the mean), (Siegelman, 1991). Variations among referees in the perception of manuscript categories (acceptable or unacceptable) would make the review system unfair to authors whose manuscripts happened to be sent to an assassin reviewer. However, uncritical acceptance of a research work would also constitute unfairness (Siegelman, 1991).

Here we provide an evolutionary explanation for why assassins and zealots use a coarse partition of manuscript categories and their extreme recommendations are inevitable consequences of such a coarse categorization. For example, this coarse partition of manuscript categories could evolve in a learning process subject to errors through bad experiences, as well when unacceptable manuscripts are much more frequent under review than acceptable ones because journals send either all or nearly all of the manuscripts out for peer review without the option of desk-rejecting.

The quasi-species model was proposed to represent the process of the Darwinian evolution of certain self-replicating entities in an environment of high mutation rate, (Eigen and Schuster, 1979; Bull et al., 2005; Schuster and Swetina, 1988). Following Mengel (2012), we propose the quasi-species model to describe the evolution of recommendation profiles in peer review.

A recommendation profile is composed of a reviewer recommendation for each manuscript category under a particular categorization of manuscripts. For the fine partition of manuscript categories (if reviewers distinguish the two categories of unacceptable and acceptable manuscripts) there are four possible recommendation profiles: (1) = (reject, reject); (2) = (reject, accept); (3) = (accept, reject);

and (4)= (accept, accept). For example, (2) = (reject,accept) is the recommendation profile for mainstream reviewers who recommend 'reject' for unacceptable manuscripts and 'accept' for manuscripts that are relevant to the scope of the journal, are innovative, significantly advance the field, and well written (i.e., acceptable manuscripts). On the contrary, if reviewers do not distinguish unacceptable and acceptable manuscripts, a coarse partition of manuscript categories is used in the evaluation of manuscripts, and only one recommendation is learned since the two manuscript categories are not distinguished. For the coarse partition (if reviewers do not distinguish unacceptable and acceptable manuscripts) there are only two possible recommendations profiles:

- (1) = reject, the profile of reviewers who recommend 'reject' for both unacceptable and acceptable manuscripts. This is an assassin reviewer with stringent standards who advises rejection much more frequently than the norm.
- (2) = accept, the profile of reviewers who recommend 'accept' for both unacceptable and acceptable manuscripts. They are zealot reviewers that may uncritically favor a manuscript.

Assassins, zealots and mainstream reviewers are "ecologically" interrelated species whose progeny tend to mutate through errors made in the process of reviewer training. Further, selection arises because different types of recommendation profiles tend to replicate at different rates. We see the reviewer mind as being built up with recommendation profiles, and the recommendation profile as replicator. Selection arises because different types of recommendation profiles tend to replicate at different rates. The evolutionary success of a particular recommendation profile depends not only on its own replication rate, but also on the replication

rates of the mutant recommendation profiles it produces, and on the replication rates of the recommendation profiles of which it is a mutant.

In the following, we first present a simple mathematical model for a quasi-species in our application. Then, we also study the conditions for the existence of assassins and zealots in peer review. Finally, we conclude by suggesting some implications of our analysis.

## 2 Basic Model

Let us consider the set of manuscript categories $S = \{s_1, s_2\}$ that the reviewer could face in different peer review situations. Let $s_1$ be the category of unacceptable manuscripts (i.e., using inappropriate research methods, poor data analysis and presentation, inadequacy of data to justify the conclusions, failure to follow the journal's styles and guidelines and failure or unwilling to revise manuscripts as per reviewers' suggestions). Besides, let $s_2$ be the category of acceptable manuscripts (i.e., relevant to the scope of the journal, innovative, significantly advance the field, and well written).

For each manuscript category $s_j$ in $S$ there is a unique optimal reviewer recommendation denoted by $i^*(s_j)$, and such that

$$i^*(s_1) = \text{reject}$$

with $s_1$ being the situation in which submission is an unacceptable manuscript, and

$$i^*(s_2) = \text{accept}$$

with $s_2$ being the situation in which submission is an acceptable manuscript.

Each recommendation is optimal for one peer review situation and all other recommendations $i \neq i^*(s_j)$ are sub-optimal, as given by the following reward function $\pi_i(s_j)$:

$$\pi_i(s_j) = \left\{ \begin{array}{ll} 1 & \text{if } i = i^*(s_j) \\ 0 & \text{otherwise} \end{array} \right\} \quad (1)$$

where, for example, $\pi_{\text{reject}}(s_1) = 1$, and $\pi_{\text{accept}}(s_1) = 0$, since $s_1$ is the situation in which submission is an unacceptable manuscript.

That is, any reviewer recommendation $i$ chosen in peer review situation $s_j$ yields rewards of 1 if it is optimal (e.g., $i$ = reject $= i^*(s_1)$) and rewards of 0 otherwise (e.g., $i =$ accept $\neq i^*(s_1)$).

In each review process, referees face a manuscript randomly drawn from $S = \{s_1, s_2\}$, i.e., an unacceptable or an acceptable submission. However, we suppose that, at the peer review stage, a reviewer can either distinguish the two manuscript categories $s_1$ and $s_2$ or not distinguish them. If reviewers distinguish the two manuscript categories $s_1$ and $s_2$ then they use a fine partition of categories in the manuscript evaluation, which we denote by $K_F$. But, if reviewers do not distinguish them, then they use the coarse partition of manuscript categories in the evaluation, which we denote by $K_C$.

In the peer review process for a given journal, we assume that category $s_1$ (i.e., unacceptable manuscripts) occurs with frequency $f$, while category $s_2$ (i.e., acceptable manuscripts) occurs with frequency $1 - f$. We suppose that reviewers can learn which recommendation to choose in each manuscript category from good and bad experiences as reader, author, and referee or from training courses. Hence there could be some errors (noise) in this training process. If reviewers distinguish the two manuscript categories $s_1$ and $s_2$, they use the fine partition $K_F$ and

a recommendation profile consisting of one recommendation for each manuscript category $s_1$ and $s_2$ is learned using experience or some other ways (e.g., peer reviewer training courses). On the contrary, if reviewers do not distinguish them, the coarse partition $K_C$ is used in the evaluation of manuscripts, and only one recommendation is learned since the two manuscript categories are not distinguished.

For the coarse partition of manuscript categories $K_C$ (if reviewers do not distinguish them) there are only two possible recommendations profiles, i.e., (1) = reject; (2)= accept. For example, (1) = reject is the recommendation profile of assassins, while (2)=accept is the recommendation profile of zealots.

For the fine partition $K_F$ (if reviewers distinguish the two manuscript categories $s_1$ and $s_2$) there are four possible recommendation profiles, i.e., (1) = (reject, reject); (2) = (reject, accept); (3) = (accept, reject); (4)= (accept, accept). For example, (2) = (reject,accept) is the recommendation profile for mainstream reviewers who reject in situation $s_1$ (i.e., unacceptable manuscripts) and accept in situation $s_2$ (i.e., acceptable manuscripts).

We are interested in reviewer recommendations when there are errors (noise) in the peer review training process. Reviewers are human beings with weaknesses and they have blind spots. Peer review evolution in such error-prone environment can be modeled using evolutionary game theory as follows.

Let $\epsilon_{ij}$ be the probability that the reviewer training process of recommendation profile $(i)$ results in recommendation profile $(j)$. We assume that $\epsilon_{ij} = \epsilon$ for all $i \neq j$, and $\epsilon_{ii} = 1 - (n-1)\epsilon$, with $n$ being the number of recommendation profiles (e.g., $n = 2$ for the coarse partition, while $n = 4$ for the fine partition). Therefore, we assume that $\epsilon_{ij} < \epsilon_{ii}$, and so, each recommendation profile is more likely to be accurately replicated than to mutate into any other given recommendation profile.

In our problem, following (Mengel, 2012), a simple mathematical model for a quasi-species is as follows.

For a given partition $K$ of the categories of unacceptable and acceptable manuscripts $s_1$ and $s_2$ (i.e., $K_F$ or $K_C$, if reviewer can either distinguish the two manuscript categories–$s_1$ and $s_2$–or not distinguish them), we denote by $p_{(i)}(K)$ the frequency of recommendation profile $(i)$ in the population of peer reviewers using partition $K$.

Recall that for the fine partition $K_F$ there are four possible recommendation profiles, i.e., (1) = (reject, reject); (2) = (reject,accept); (3) = (accept,reject); (4)= (accept,accept). For the coarse partition, there are only two possible recommendations profiles, i.e., (1) = reject; (2)= accept. Therefore, we denote by

$$p(K) = (p_{(1)}(K), \ldots, p_{(n)}(K))$$

the vector of frequencies of the different recommendation profiles in peer review using partition $K$.

Besides, let $\pi_{(i)}(K)$ be the reward of recommendation profile $(i)$ under partition $K$ given by

$$\pi_{(i)}(K) = \sum_{s \in S} f_s \pi_i(s) \tag{2}$$

with $S = \{s_1, s_2\}$ being the set of manuscript categories; $f_s$ being the frequency of manuscript category $s$ in peer review; and $\pi_i(s)$ being the reward function under recommendation profile $(i)$ for manuscript category $s$, as defined in equation (1). Then, we denote by

$$\pi(K) = (\pi_{(1)}(K), \ldots, \pi_{(n)}(K))$$

the reward vector of recommendation profiles under partition $K$.

The average reward of the reviewers' population using partition $K$ is therefore given by the inner product of the vector of frequencies of recommendation profiles $p(K)$ and the reward vector $\pi(K)$:

$$\hat{\pi}(K) = \pi(K) \cdot p(K) \tag{3}$$

Then, the quasi-species formulation may be expressed as a linear differential equation, where recommendation profile $(i)$ is obtained by replicating any recommendation profile $(j)$ at rate $\pi_{(j)}(K)$ times the probability $\epsilon_{ji}$ that training process of recommendation profile $(j)$ results in profile $(i)$, (Mengel, 2012):

$$\dot{p}_{(i)}(K) = \sum_{j=1}^{n} p_{(j)}(K)\pi_{(j)}(K)\epsilon_{ji} - \hat{\pi}(K)p_{(i)}(K) \tag{4}$$

where each recommendation profile is removed at rate $\hat{\pi}(K)$ to ensure that the total size of reviewers' population remains constant.

To solve such an equation we combine rewards and errors in reviewer training to derive the mutation selection matrix $W(K) = [w_{ji}] = [\pi_{(j)}(K)\epsilon_{ji}]$ for our problem of peer review, (Mengel, 2012).

Then the quasi-species equations for the peer review problem can be rewritten as

$$\dot{p}(K) = p(K)W(K) - \hat{\pi}(K)p(K) \tag{5}$$

with equilibrium (i.e., $\dot{p}(K) = 0$) satisfying

$$p(K)W(K) = \hat{\pi}(K)p(K) \tag{6}$$

Hence the average reward of the reviewers' population using partition of manuscript categories $K$ is the largest eigenvalue of the matrix $W(K)$, (Mengel, 2012). With the proper normalization $\sum p_{(i)}(K) = 1$, the left-hand eigenvector associated with

this average reward of the reviewers' population provides the equilibrium structure of the quasi-species for recommendation profiles in peer review. Therefore, the reward of a partition of manuscript categories $K$ is determined by the largest eigenvalue of the matrix $W(K)$.

In the following we will compare this average reward of the reviewers' population with a partition $K$ of manuscript categories across different partitions (i.e., the fine partition $K_F$ and the coarse partition $K_C$).

**3 Conditions for the existence of assassins and zealots in peer review**

One would expect the fine partition of manuscript categories $K_F$ (when reviewers can distinguish the two categories–$s_1$ and $s_2$– of unacceptable and acceptable manuscripts) to be optimal since it allows the reviewer to choose the optimal recommendation for each manuscript category that could appear in any review situation. That is, recommendation profile (2) = (reject, accept) for mainstream reviewers who reject in situation $s_1$ (unacceptable manuscripts) and accept in situation $s_2$ (acceptable manuscripts).

Besides, the advantage of using $K_F$ is biggest whenever both manuscript categories $s_1$ and $s_2$ occur equally often in peer review processes.

However, when unacceptable manuscripts ($s_1$) are much more frequent than others ($s_2$) because journals send either all or nearly all of the manuscripts out for peer review without the option of desk-rejecting, the coarse partition of categories $K_C$ (i.e., reviewers do not distinguish manuscript categories $s_1$ and $s_2$) can provide higher reward to the reviewers' population–if the recommendation which is optimal

in the very frequent situation is chosen– whenever errors in the reviewer training process are sufficiently frequent. This is illustrated in the following example.

Let us consider that manuscript category $s_1$ (i.e., unacceptable manuscripts) occurs under review with frequency $f = \frac{3}{4}$, while manuscript category $s_2$ (i.e., acceptable manuscripts) occurs under review with frequency $(1-f) = \frac{1}{4}$.

For the coarse partition $K_C$, there are only two possible recommendations profiles, i.e., (1) = reject; (2)= accept. For each manuscript category, the optimal reviewer recommendation denoted by $i^*(s_j)$ is such that $i^*(s_1)$= reject, while $i^*(s_2)$=accept.

Therefore, the reward of recommendation '(1) = reject' in the coarse partition $K_C$ is

$$\pi_{(1)}(K_C) = f\pi_{reject}(s_1) + (1-f)\pi_{reject}(s_2) = f \times 1 + (1-f) \times 0 = \frac{3}{4}$$

and similarly, the reward of recommendation '(2) = accept' in the coarse partition $K_C$ is

$$\pi_{(2)}(K_C) = f\pi_{accept}(s_1) + (1-f)\pi_{accept}(s_2) = f \times 0 + (1-f) \times 1 = \frac{1}{4}.$$

For the fine partition $K_F$ there are four possible recommendation profiles, i.e., (1) = (reject, reject); (2) = (reject,accept); (3) = (accept,reject); (4)= (accept,accept).

Therefore, the reward of recommendation '(1) =(reject,reject)' in the fine partition $K_F$ is

$$\pi_{(1)}(K_F) = f \times \pi_{reject}(s_1) + (1-f) \times \pi_{reject}(s_2) = f \times 1 + (1-f) \times 0 = \frac{3}{4}$$

the reward of recommendation '(2) = (reject,accept)' is

$$\pi_{(2)}(K_F) = f \times \pi_{reject}(s_1) + (1-f) \times \pi_{accept}(s_2) = f \times 1 + (1-f) \times 1 = 1$$

Similarly, the reward of recommendation '(3) = (accept,reject)' is

$$\pi_{(3)}(K_F) = f \times \pi_{accept}(s_1) + (1-f) \times \pi_{reject}(s_2) = f \times 0 + (1-f) \times 0 = 0$$

and the reward of recommendation '(4) = (accept,accept)' is

$$\pi_{(4)}(K_F) = f \times \pi_{accept}(s_1) + (1-f) \times \pi_{accept}(s_2) = f \times 0 + (1-f) \times 1 = \frac{1}{4}.$$

Recall that $\epsilon_{ij}$ denotes the probability that the reviewer training process of recommendation profile $(i)$ results in recommendation profile $(j)$, and that $\epsilon_{ij} = \epsilon$ for all $i \neq j$, and $\epsilon_{ii} = 1-(n-1)\epsilon$, with $n$ being the number of recommendation profiles (e.g., $n = 2$ for the coarse partition, while $n = 4$ for the fine partition).

Following Mengel (2012), to solve the quasi-species equation we combine rewards and errors in reviewer training to derive the mutation selection matrix for the coarse partition $W(K_C) = [\pi_{(j)}(K_C)\epsilon_{ji}]$ which is given by

$$W(K_C) = \begin{pmatrix} \frac{3(1-\epsilon)}{4} & \frac{3\epsilon}{4} \\ \frac{\epsilon}{4} & \frac{(1-\epsilon)}{4} \end{pmatrix}$$

Therefore the average reward of the reviewers' population using the partition $K_C$ is the largest eigenvalue of the matrix $W(K_C)$ which is given by $\frac{1-\epsilon}{2} + \frac{1}{4}\sqrt{1-2\epsilon+4\epsilon^2}$. With the proper normalization $\sum p_{(i)}(K) = 1$, the left-hand eigenvector associated with this average reward of the reviewers' population provides the equilibrium structure of the quasi-species for recommendation profiles in peer review. This eigenvector is the vector of frequencies of the different recommendation profiles in peer review using partition $K_C$

$$p(K_C) = (p_{(1)}(K_C), p_{(2)}(K_C)) = (p_{reject}(K_C), p_{accept}(K_C))$$

where $p_{accept}(K_C) = 1 - p_{reject}(K_C)$.

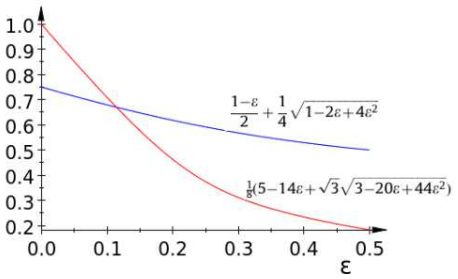



**Fig. 1** Comparison of the average reward of the reviewers' population with the coarse partition $K_C$ (blue) and with the fine partition $K_F$ (red).

To solve the quasi-species equation for the reviewers' population using the fine partition $K_F$ with four possible recommendation profiles, i.e., (1) = (reject, reject), (2) = (reject,accept), (3) = (accept,reject), and (4)= (accept,accept), we again combine rewards and errors in reviewer training to derive the corresponding mutation selection matrix $W(K_F) = [\pi_{(j)}(K_F)\epsilon_{ji}]$

$$W(K_F) = \begin{pmatrix} \frac{3(1-3\epsilon)}{4} & \frac{3\epsilon}{4} & \frac{3\epsilon}{4} & \frac{3\epsilon}{4} \\ \epsilon & (1-3\epsilon) & \epsilon & \epsilon \\ 0 & 0 & 0 & 0 \\ \frac{\epsilon}{4} & \frac{\epsilon}{4} & \frac{\epsilon}{4} & \frac{(1-3\epsilon)}{4} \end{pmatrix}$$

with the average reward of the reviewers' population using the partition $K_F$ being the largest eigenvalue of the matrix $W(K_F)$ which is bound above by $\frac{1}{8}(5-14\epsilon+\sqrt{3}\sqrt{3-20\epsilon+44\epsilon^2})$.

By the comparison of the average reward of the reviewers' population with the coarse partition $K_C$ and with the fine partition $K_F$ we obtain that whenever $\epsilon > 0.114$, the population of reviewers using the coarse partition has higher evolutionary reward (see Figure 1).

Therefore, if reviewers do not distinguish categories of unacceptable and acceptable manuscripts, then they can achieve higher reward whenever errors in the training process of reviewers are sufficiently frequent (in this example around 10% which is not very high).

At this value of error, the frequencies of the different recommendation profiles in peer review are

$$(p_{reject}(K_C), p_{accept}(K_C) = (0.81, 0.19).$$

Hence, under the coarse partition the recommendation with the highest reward (reject) also has the highest share of reviewers' population (81%).

As a result, given the higher average reward using a coarse partition of manuscript categories, we see that the quasi-species equation conveys in a natural way the conditions under which assassins and zealots evolutionary appear as follows. If one of the two manuscript categories is much more frequent at the peer review stage than the other ($f$ close to zero or one), then the coarse partition of manuscript categories usually does better than the fine partition whenever errors in the reviewer training process are sufficiently frequent. This explains the evolutionary success of zealots and assassins in peer review, since they use a coarse partition of manuscripts, and therefore, do not distinguish acceptable from unacceptable.

The previous example provides a numerical illustration that the required error rate is not very high (around 10%). A mathematical result states this idea in a more formal way as follows.

**Proposition.** *Suppose that category $s_1$ (i.e., unacceptable manuscripts) occurs under review with frequency $f$, while category $s_2$ (i.e., acceptable manuscripts) occurs under review with frequency $1 - f$.*

*Then, there is an error threshold $\hat{\epsilon}$ decreasing as $|f-1/2|$ rises, such that, whenever errors in the reviewer training are sufficiently frequent ($\epsilon > \hat{\epsilon}$), the coarse partition of manuscript categories (under which reviewers do not distinguish unacceptable from acceptable) yields higher average reward for a population of reviewers than the fine partition of manuscript categories.*

Proof: See Appendix A

Figure 2 illustrates the comparison of the average reward of the reviewers' population with the coarse partition of manuscript categories $K_C$ (blue) and with the fine partition $K_F$ (red), when the category $s_1$ (i.e., unacceptable manuscripts) occurs with frequency $f = 0.5, 0.6, 0.7, 0.8$, and 0.9. Recall that manuscript category $s_2$ (i.e., acceptable manuscripts) occurs with frequency $1 - f$.

This figure shows that if manuscript category $s_1$ of unacceptable manuscripts is much more frequent than $s_2$ (i.e., $f$ close to one) then the coarse partition of manuscript categories usually does better than the fine partition whenever reviewer errors $\epsilon$ increase. Something similar happens when manuscript category $s_2$ of acceptable manuscripts is much more frequent than $s_1$ (i.e., $f$ close to zero).

Figure 2 also shows that the fine partition does best when both manuscript categories occurs under review with similar frequency (e.g., $f = 0.5$), whenever errors $\epsilon$ are small.

## 4 Conclusion

In each review process referees face a manuscript randomly drawn from $S = \{s_1, s_2\}$, i.e., one unacceptable or acceptable submission. However a reviewer can either distinguish the two manuscript categories $s_1$ and $s_2$ or not distinguish them.

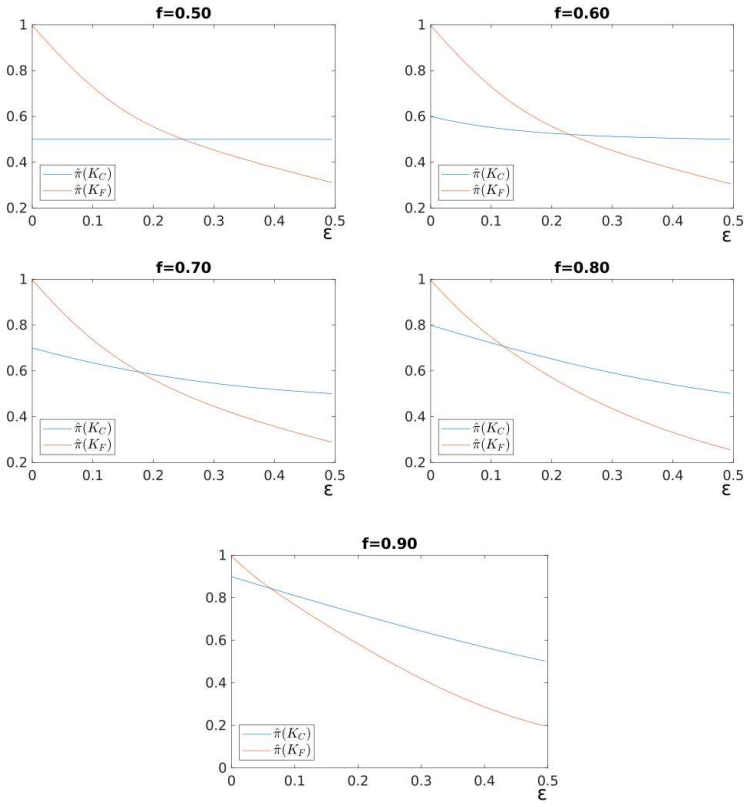



**Fig. 2** Comparison of the average reward of the reviewers' population with the coarse partition $K_C$ (blue) and with the fine partition $K_F$ (red). Manuscript category $s_1$ (i.e., unacceptable manuscripts) occurs with frequency $f = 0.5, 0.6, 0.7, 0.8, 0.9$.

If reviewers distinguish the two manuscript categories $s_1$ and $s_2$ then they use a fine partition in the manuscript evaluation. But, if reviewers do not distinguish them, then they use the coarse partition in the evaluation of manuscripts.

In the peer review process for a given journal, we assume that category $s_1$ (i.e., unacceptable manuscripts) occurs with frequency $f$, while category $s_2$ (i.e.,

acceptable manuscripts) occurs with frequency $1 - f$. Then, reviewers can learn which recommendation to choose in each manuscript category from good and bad experiences as reader, author, and referee or from training courses. There could be some errors (noise) in this training process. If reviewers distinguish the two manuscript categories, a recommendation profile consisting of one recommendation for each manuscript category is learned. On the contrary, if reviewers do not distinguish them, only one recommendation is learned for both acceptable and unacceptable manuscripts and zealots and assassins appear.

Here we propose the quasi-species model to describe the evolution of recommendation profiles in peer review. In our application, the self-replicating entities are recommendation profiles under a given partition of manuscript categories. Mutations occur through errors made in the process of peer-review training. The quasi-species equation conveys in a natural way the conditions under which assassins and zealots evolutionary appear because of the higher reward using a coarse partition of manuscript categories: if one of the two manuscript categories is much more frequent at the peer review stage than the other ($f$ close to zero or one), then the coarse partition usually does better than the fine partition whenever errors in the training process of reviewers are sufficiently frequent.

The fine partition does best when both manuscript categories are equally important in peer review ($f = 0.5$). In terms of errors, the fine partition does best when errors are small. Hence, in order that reviewers distinguish acceptable from unacceptable manuscripts (by using a fine partition), only similar percentages of them should sent out to peer review. Besides, editors should assess unfairness in reviewers' recommendations to recognize and exclude errors from the peer review training processes.

Our results explain why most academic journals should follow a strict desk rejection policy. Manuscripts should be desk rejected when they do not fit the journal's scope or are too underdeveloped or flawed to benefit from the review process. This helps to send similar percentages of different categories of manuscripts to external review, which avoids the conditions under which assassins and zealots evolutionary appear in peer review.

Of course, journals should also clarify and communicate criteria for desk rejection during the editorial pre-screening process. Moreover, authors should bear in mind that their works are subject to desk rejections.

On the other hand, our results promote standardization in the reviewer training processes to reduce the incidence of inappropriate evaluation. This standardization possibly connects peer review to credentialing that makes sure that all reviewers meet the threshold criteria to filter out troublesome referees.

**Acknowledgments.** This research was sponsored by the Spanish Board for Science, Technology, and Innovation under grant TIN2017-85542-P, and co-financed with European FEDER funds. Sincere thanks are due to the reviewers for their constructive suggestions.

**References**

Eigen M, Schuster P (1979). The Hypercycle: A Principle of Natural Self-Organization. Berlin: Springer-Verlag. ISBN 978-0-387-09293-5.

Bull JJ, Meyers LA, Lachmann M (2005). Quasispecies made simple. PLoS Computational Biology. 1 (6): e61. https://doi.org/10.1371/journal.pcbi.0010061

Burnham, J. C. (1990). The Evolution of Editorial Peer Review. JAMA, 263(10), 1323-1329.

Campanario, J.M. (1998a). Peer review for journals as it stands today - Part 1. Science Communication, 19(3), 181-211.

Campanario, J.M. (1998b). Peer review for journals as it stands today - Part 2. Science Communication, 19(4), 277-306.

Chubin, D.E., & Hackett, E.J. (1990). Peerless science: Peer review and U.S. science policy. Stony Brook, NY: State University of New York Press.

Garcia, J.A., Rodriguez-Sanchez, Rosa, Fdez-Valdivia J., (2015a). The author–editor game. Scientometrics, 104(1). https://doi.org/10.1007/s11192-015-1566-x

Garcia, J.A., Rodriguez-Sanchez, R., and Fdez-Valdivia, J., (2015b). Adverse selection of reviewers. Journal of the Association For Information Science and Technology, 66(6), 1252-1262, https://doi.org/10.1002/asi.23249

Garcia, J.A., Rodriguez-Sanchez, R., and Fdez-Valdivia, J., (2016). Why the referees' reports I receive as an editor are so much better than the reports I receive as an author?. Scientometrics, 106(3), 967-986, https://doi.org/10.1007/s11192-015-1827-8

Lee, Carole J., Sugimoto, Cassidy R., Zhang, Guo, and Cronin, Blaise, (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2-17.

Mengel, F (2012). On the evolution of coarse categories. Journal of Theoretical Biology 307(21). 117-124. https://doi.org/10.1016/j.jtbi.2012.05.016

Merton, R.K. (1973). The sociology of science: Theoretical and empirical investigations. Chicago: University of Chicago Press.

Rodriguez-Sanchez, Rosa, Garcia, J.A., Fdez-Valdivia J., (2016). Evolutionary games between authors and their editors. Applied Mathematics and Computation, 273(15), pp. 645-655. https://doi.org/10.1016/j.amc.2015.10.034

Schuster P, Swetina J (1988). Stationary mutant distributions and evolutionary optimization. Bulletin of Mathematical Biology. 50 (6): 635-660. https://doi.org/10.1007/BF02460094

Siegelman, S S (1991), Assassins and zealots: Variations in peer review. Special report. Radiology, 178(3), 637-642. https://doi.org/10.1148/radiology.178.3.1994394

Souder, L. (2011). The ethics of scholarly peer review: a review of the literature. Learned Publishing, 24(1), 55-72.

Tenopir, C. and King, D.W. (2007), Perceptions of value and value beyond perceptions: measuring the quality and value of journal article readings, Serials, 20(3), pp. 199-207.

**A Proof**

We have to prove that there is an error threshold $\hat{\epsilon}$ decreasing in $|f - 1/2|$ such that whenever errors in the reviewer training are sufficiently frequent ($\epsilon > \hat{\epsilon}$), then the coarse partition (under which reviewers do not distinguish categories of unacceptable and acceptable manuscripts) yields higher average reward for a population of reviewers than the fine partition.

To this aim we follow the proof of result 1 in Mengel (2012). So, given the average reward of the reviewers' population using the coarse partition $K_C$, $\hat{\pi}(K_C)$, and

that using the fine partition $K_F$, $\hat{\pi}(K_F)$, we show that $\hat{\pi}(K_F) - \hat{\pi}(K_C)$ decreases in $\epsilon$ for all $\epsilon < 1/2$.

The average reward of the reviewers' population using the partition $K_C$ is the largest eigenvalue of the matrix

$$W(K_C) = \begin{pmatrix} (1-\epsilon)f & \epsilon f \\ \epsilon(1-f) & (1-\epsilon)(1-f) \end{pmatrix}$$

which is given by

$$\hat{\pi}(K_C) = \frac{1-\epsilon}{2} + \frac{1}{2}\sqrt{(1-2f)^2 - \epsilon(2-8f+8f^2) + \epsilon^2}$$

and therefore, taking derivatives we find

$$\frac{\partial \hat{\pi}(K_C)}{\partial \epsilon} = \frac{1}{2}\left(\frac{\epsilon - (1-2f)^2}{\sqrt{(1-2f)^2 - 2\epsilon(1-2f)^2 + \epsilon^2}} - 1\right)$$

hence, $-1 \leq \frac{\partial \hat{\pi}(K_C)}{\partial \epsilon} \leq 0$.

Similarly, the average reward of the reviewers' population using the partition $K_F$ is the largest eigenvalue of the matrix

$$W(K_F) = \begin{pmatrix} (1-3\epsilon)f & \epsilon f & \epsilon f & \epsilon f \\ \epsilon & (1-3\epsilon) & \epsilon & \epsilon \\ 0 & 0 & 0 & 0 \\ \epsilon(1-f) & \epsilon(1-f) & \epsilon(1-f) & (1-3\epsilon)(1-f) \end{pmatrix}$$

which solves the equilibrium of the quasi-species equations

$$p(K_F)W(K_F) = \hat{\pi}(K_F)p(K_F).$$

From this equilibrium we get

$$p_{(2)}(\hat{\pi}(K_F) - (1-4\epsilon)) = p_{(4)}(\hat{\pi}(K_F) - (1-4\epsilon)(1-f))$$

where we denote by $p_{(i)}$ the frequency of recommendation profile $(i)$ in the population of peer reviewers using partition $K_F$, and there are four possible recommendation profiles, i.e., (1) = (reject, reject); (2) = (reject,accept); (3) = (accept,reject); (4)= (accept,accept). Therefore it follows that

$$\hat{\pi}(K_F) = \frac{(1-4\epsilon)(p_{(2)} - (1-f)p_{(4)})}{p_{(2)} - p_{(4)}}$$

We observe that taking differences between $\hat{\pi}(K_F)$ and $\hat{\pi}(K_C)$ we find (with $f \neq 0.5$)

$$\hat{\pi}(K_F) - \hat{\pi}(K_C) = \left\{ \begin{array}{ll} (1-f) > 0 & \text{at } \epsilon = 0 \\ < 0 & \text{at } \epsilon = \frac{1}{4} \end{array} \right\}$$

Now taking derivatives in $\hat{\pi}(K_F)$ we get

$$\frac{\partial \hat{\pi}(K_F)}{\partial \epsilon} = \frac{\left( \begin{array}{c} -4(p_{(2)} - p_{(4)}) \left[ (p_{(2)} - (1-f)p_{(4)}) + \epsilon \left( \frac{\partial p_{(2)}}{\partial \epsilon} - (1-f)\frac{\partial p_{(4)}}{\partial \epsilon} \right) \right] \\ -(1-4\epsilon) \left( \frac{\partial (p_{(2)} - p_{(4)})}{\partial \epsilon} (p_{(2)} - (1-f)p_{(4)}) \right) \end{array} \right)}{(p_{(2)} - p_{(4)})^2}$$

Therefore, taking differences between $\frac{\partial \hat{\pi}(K_F)}{\partial \epsilon}$ and $\frac{\partial \hat{\pi}(K_C)}{\partial \epsilon}$ we find

$$\frac{\partial \hat{\pi}(K_F)}{\partial \epsilon} - \frac{\partial \hat{\pi}(K_C)}{\partial \epsilon} = \left( -4 \left( 1 + \epsilon \left( \frac{\partial p_{(2)}}{\partial \epsilon} - (1-f)\frac{\partial p_{(4)}}{\partial \epsilon} \right) \right) - \frac{\partial (p_{(2)} - p_{(4)})}{\partial \epsilon} \right) - (-1) < 0$$

Hence, given $f$, both $\frac{\partial \hat{\pi}(K_F)}{\partial \epsilon}$ and $\frac{\partial \hat{\pi}(K_C)}{\partial \epsilon}$ are negative for all values of $\epsilon$ (and continuous). Therefore, there is a $\hat{\epsilon}$, with $0 < \hat{\epsilon} < \frac{1}{4}$, such that

$$\hat{\pi}(K_F) < \hat{\pi}(K_C), \text{ for all } \epsilon > \hat{\epsilon}.$$

Also, by Lemma 2 in Mengel (2012), we have that, for any $\epsilon > 0$, $\hat{\pi}(K_F) - \hat{\pi}(K_C)$ is maximized at $f = 1/2$. Hence, following (Mengel, 2012), the upper bound on $\hat{\epsilon}(f)$ can be found by looking at the uniform case. Therefore, the set of the eigenvalues for $W(K_F)$ is

$$\left\{ 0, \frac{1}{2}(1-4\epsilon), \frac{1}{4}(3 - 8\epsilon \pm \sqrt{1 - 8\epsilon + 32\epsilon^2}) \right\}$$

To complete the proof we only have to observe that the maximal eigenvalue for the coarse partition $W(K_C)$ is given by $\lambda = 1/2$ which exceeds the maximal element of the set of eigenvalues for $W(K_F)$, as $\epsilon > 1/4$.

## 3.2. The author's ignorance on the publication fees is a source of power for publishers

### 3.2.1. Datos generales

1. **Autores**: Jose A. Garcia, Rosa Rodríguez-Sánchez, J.Fdez-Valdivia y Jorge Chamorro-Padial.
2. **Revista**: Scientometrics.
3. **Datos sobre la publicación**:
   - **Referencia**: García et al. (2019).
   - **Volumen**: 121.
   - **Número**: 3.
   - **Páginas**: 1435–1445.
   - **Año**: 2019.
   - **Editorial**: Springer.
   - **DOI**: https://doi.org/10.1007/s11192-019-03231-8.
4. **Estado**: Publicado.
5. **Métricas**:
   - **Factor de impacto**: 3,801 (JCR, 2021).
   - **Ranking**:
     - *Social Science Citation Index (SSCI)*:
       - *Information Science & Library Science*: Q1 - 29/84 (año 2021).
       - *Information Science & Library Science*: Q1 - 21/87 (año 2019).
     - *Social Science Citation Index Expanded (SSCIE)*:
       - *Computer Science, interdisciplinary applications*: Q2 - 54/112 (año 2021).
       - *Computer Science, interdisciplinary applications*: Q2 - 45/109 (año 2019).

### 3.2.2. Contribuciones principales

1. Propuesta de un modelo que permite explicar la competitividad existente entre revistas académicas para captar la atención de los autores de la Comunidad Científica. En este modelo, se relaciona la complejidad del sistema de tarifas de publicación de un artículo con respecto al número de revistas competidoras entre sí en un área de investigación.

2. Cuando crece la competitividad en un área, aumenta la complejidad del sistema tarifario de los artículos.

3. Si por parte de la Comunidad científica o de un grupo de interés se emprenden campañas informativas, los autores que comprenden el modelo tarifario también se incrementa, incluso si crece la competitividad entre revistas.

### 3.2.3. Resumen

Las revistas científicas, generalmente, utilizan las tarifas de publicación (*article processing charges*, APC) como fuente de financiación. Estas tarifas varían en cuantía y complejidad en cada revista (Björk y Solomon, 2012). El entramado de cargos es, frecuentemente, demasiado complejo y logra escapar al entendimiento de los autores.

En nuestro trabajo, estudiamos cómo las estrategias que adquieren las revistas académicas en cuanto a política tarifaria varía en función del número de competidores existentes en un área de la ciencia. Para ello, definimos un modelo (*publication cost game*) basado en (Carlin, 2009). Debemos distinguir entre autores informados (entienden el modelo de costes de una revista) y no informados (**no** entienden el modelo de costes de una revista) Bajo nuestro modelo, podemos extraer las siguientes proposiciones:

1. En una situación de equilibrio entre número de revistas y número de autores. Si una revista opta por un modelo de bajo coste para el autor, la estrategia óptima consiste en minimizar la complejidad de su modelo de tarifas incrementando así el número de autores informados. Del mismo modo, si una revista ha adoptado un modelo de coste elevado para el autor, la estrategia óptima consiste en maximizar su complejidad tarifaria para incrementar el número de autores no informados.

2. Cuando el número de revistas de un impacto similar se incrementa en un área, la probabilidad de que cada revista añada complejidad en su sistema de tarifas también aumenta.

3. Cuando el número de revistas de un impacto similar se incrementa en un área, el número de autores no informados crece, siempre y cuando no exista ninguna regulación en el área. Contrariamente, si se realizan campañas educativas por parte de la Academia o de grupos de interés en el área, el número de autores informados puede incrementarse conforme el nivel de competitividad entre revistas crezca.

**Scientometrics manuscript No.**
(will be inserted by the editor)

## The author's ignorance is a source of power for publishers

**J. A. García, Rosa Rodriguez-Sánchez,**

**J. Fdez-Valdivia, and Jorge**

**Chamorro-Padial**

the date of receipt and acceptance should be inserted later

**Abstract** Over the last few years, the former editorial board of the Journal of Informetrics has grown increasingly dissatisfied with Elsevier's actions and policies, such as Elsevier's refusal to participate in the Initiative for Open Citation, its restrictive open access policies and prohibitive subscription costs. The problem is that even when new journals enter the research field, the publication costs for authors often do not decrease and may in fact rise. In this context, publishers of scientific journals may create author's ignorance by making the costs for authors more complex, thereby gaining power to increase their

J. A. García, Rosa Rodriguez-Sánchez, and J. Fdez-Valdivia,

Departamento de Ciencias de la Computación e I. A., CITIC-UGR, Universidad de Granada, 18071 Granada, Spain.

Jorge Chamorro-Padial,

CITIC-UGR, Universidad de Granada, 18071 Granada, Spain.

Address correspondence to J. A. García at jags@decsai.ugr.es

Jose A. Garcia ORCID iD `https://orcid.org/0000-0001-7742-7270`

profits. In this paper, we present a model of competition between academic journals to publish authors' manuscripts. Since the low-fees journals want authors to know that they have the lowest costs to publish their research works, these journals want fees to be reasonably clear. Adding clarity allows them to undercut their competitors and gain percentage of authors in the field. On the contrary, high-fees journals desire more complexity in the disclosure of publication costs for authors. We also show that a higher proportion of journals will add complexity to their costs when there is greater journal competition in the research field. In our model, as journal competition increases, the expected number of authors informed about the publication costs in the field may decrease without optimal regulation. On the contrary, if the academy or interest groups in the field add educational initiatives, the expected number of informed authors may still increase as the number of competitor journals grows.

**Keywords**: Journal Competition; Publication Costs; Complexity; Informed Authors; Educational Initiatives.

## 1 Introduction

Recently we have heard about the resignation of the editorial board of the Journal of Informetrics. For example, (Larivière, 2018) explained it in the following way: "Over the last few years, the editorial board of the Journal of Informetrics (JOI) has grown increasingly dissatisfied with Elsevier's actions and policies. While some of those have specific effects on our field—such as El-

sevier's refusal to participate in the Initiative for Open Citation (I4OC)–others are affecting all fields of science–such as its restrictive open access policies and prohibitive subscription costs."

The problem is that despite the large number of academic journals in each research field, even when new journals enter the field, the author costs often do not decrease and may in fact rise. In this context, publishers of scientific journals may create author's ignorance by making the publication costs more complex, thereby gaining power to increase their profits. So, how much does it cost to publish a research article in a scholarly journal? For example, a research work can consume thousands of hours of author's life before its publication. Other hidden costs of authoring an article in a journal are the number of days before a final decision on the manuscript submission. Both subscription-based and open access journals may charge a fee at the time of manuscript submission to help to fund editorial and peer review administration, (Panter, 2019). In some journals, authors can also expedite the peer-review and publishing process by paying an optional fee.

Academic journals may add complexity to their publication costs for authors in several ways, e.g., nominal submission fees payable on submission, and/or article processing fees payable only in case of acceptance. There, to cover the cost of printing, and particularly color printing, certain traditional journals charge per page and/or per color figure. In rare cases, supplementary materials may also incur a flat charge or a charge per item or page. Publication fees are charged by certain open access journals post-acceptance, are

also known as author publishing charges or article processing charges, (Panter, 2019). Page/color-independent fees may also be billed by traditional journals without unrestricted access and/or reuse provisions. So, a number of journals charge fees to authors of one kind or another. Pre-publication fees, such as a submission fee or membership fee, are less common, (Panter, 2019). Researchers are more likely to encounter post-publications fees, such as an article processing charge or page fee. Complexity in publication costs may also involve that predatory journals may take advantage of the model of article processing charges to receive payment in return for minimal peer review and processing.

Therefore, it is easy to get overwhelmed by the diversity of potential author fees, (Panter, 2019): What are all of these types of fees? Which types of journals generally charge them? When? Why? Many authors understand open access journals in terms of article processing charges, while they think of subscription journals just like no author charge. However, as an author, you may have to pay for submission to and/or publication in a subscription-based journal and may not have to do so for an open access one, (Panter, 2019). The latter is possible by alternative sources of revenue that cover the costs of the editorial, peer review, and publication processes, such as advertising, or subsidy by a journal's affiliated foundation or society, (Panter, 2019). Note also that for both traditional and open access publications that do entail so-called author charges, authors may not have to pay these fees in full because of discounts related to membership programs, waivers of service, country of au-

thor's economic status or due to coverage by author's institution, department, or funder/grant, (Panter, 2019).

In this paper, we present the publication cost game in which academic journals of similar citation impact publish papers in some research field. In a first period, journals choose the cost for authors (e.g., submission fees, publication fees, and so on) and the complexity of the cost structure (e.g., disclosure of publication costs for authors). Based on the complexity choices of the journals, a fraction of expert authors become informed about publication costs, while the rest remain uninformed about the diversity of potential author fees. In a second period, expert authors choose to submit their manuscripts to the journal of low publication costs, while the uninformed authors submit randomly to any journal. Of course, journals compete strategically for both types of authors. In equilibrium, based in each journal's strategy about publication costs for authors and complexity, it arises a dispersion in those costs to publish a research article in a scholarly journal. The journal with the lowest publication costs captures the expert authors. However, all journals receive some demand to publish manuscripts from the uninformed authors. Then we study several questions: How do these optimal strategies over publication costs and complexity change as the number of journals increases? What is the potential effect of the journals' strategies on author's ignorance about publication costs in the field?

## 2 The publication cost game

Following Carlin (2009), we present a basic model of competition between academic journals to publish authors' manuscripts. Consider a research field in which $n$ academic journals publish papers on a particular topic. We assume that they have a similar journal citation impact of $q$. Therefore, their only potential difference is the publication costs that they may charge to authors and the complexity that the journals add to their strategy of publication costs.

In the research field, there are $M$ authors who each wish to publish a paper in one of the journals. The author $i$ maximizes the expected payment from the manuscript publication. The utility of author $i$ is

$$U_i = q - c_i \tag{1}$$

where $q$ is the journal citation impact and $c_i$ is the cost to publish the manuscript in the scholarly journal. The journals' impact factor $q$ is a measure of the frequency with which the average article in a journal has been cited in a period of time. The author is only able to choose the citation impact $q$ of the journal where the paper is published. Maximizing the utility $U_i$ in the research field is equivalent to minimizing the publication costs that the authors have to pay (hidden costs, submission fees, publication fees, and so on).

Journals may add complexity to their publication costs which affect how educated the authors are about the differences between journals in the publication costs. Expert authors (fraction $\eta$) are those researchers who become fully informed about the publication costs in the research field. They choose

to publish at the lowest publication cost $c_{min}$ available assuming a similar journal citation impact of $q$:

$$c_{min} = \min\{c_k\}_{k=1}^{n} \tag{2}$$

Besides, there is a fraction $1 - \eta$ of uninformed authors who are unaware of the differences between the costs to publish a paper in the different journals. They submit manuscripts to a randomly chosen journal with a citation impact of $q$. Then, the probability that an uninformed author submits the manuscript to any one journal is $1/n$ and the expected publication cost they will pay is

$$\overline{c} = \frac{1}{n}\sum_{k=1}^{n} c_k \tag{3}$$

*In the first period of the publication cost game*, each scholarly journal $j$ generates strategic choices for the costs $c_j \in [0, q]$ that charges to the author, and for the complexity $k_j \in [\underline{k}, \overline{k}]$ that measures how difficult it is to screen hidden costs and pre- and post-publication fees for the journal. As a result, $\sigma_j = (c_j, k_j)$ defines the journal j's mixed strategy over publication costs and complexity, with $\sigma_j \in [0, q] \times [\underline{k}, \overline{k}]$.

Here, the publication cost game describes the natural competition between academic journals to be chosen by authors who wish to publish their manuscripts. But, how do journals optimally add complexity to their publication costs to maximize profits? How do these optimal strategies over publication costs and complexity change as the number of journals increases? What is the potential effect of the journals' strategies on author's ignorance about publication costs? To answer these questions, in what follows, we study the

mixed-strategy Nash equilibrium for the publication cost game. In any Nash equilibrium, the mixed strategies of the journals are given by the vector

$$\sigma^* = (\sigma_1^*, \ldots, \sigma_n^*)$$

where $\sigma_j^* = (c_j^*, k_j^*)$, with $j = 1, \ldots, n$.

In game theory, the Nash equilibrium is a solution concept of a non-cooperative game involving two or more players, (Nash, 1951). In the publication cost game involving two or more academic journals, if each journal has chosen a mixed strategy $\sigma_j^* = (c_j^*, k_j^*)$ over publication costs and complexity and no journal can benefit by changing strategies while the other journals keep theirs unchanged, then the current set of strategy choices constitutes a Nash equilibrium. Stated simply, scholarly journals $i$ and $j$ are in a Nash equilibrium if $i$'s journal is making the best decision over publication costs and complexity, taking into account $j$'s journal decision while $j$'s journal decision remains unchanged, and $j$'s journal is making the best decision over publication costs and complexity, taking into account $i$'s journal decision while $i$'s journal decision remains unchanged.

In order to capture the idea that complexity choices by competing journals not only make it difficult to screen the publication costs for one journal, but also may make it more difficult to compare publication costs among journals, the fraction of educated authors $\eta$ is formulated as follows:

$$\eta : [\underline{k}, \overline{k}]^n \to (0, 1) \tag{4}$$

such that $\delta\eta/\delta k_j < 0$ for all $j$ which implies that as any one journal $j$ makes their publication costs more difficult to screen, it makes the journals in the research field harder to evaluate, and thereby lowers the fraction $\eta$ of informed authors; and also, the multivariate map $\eta$ is such that $\delta^2\eta/\delta k_j \delta k_l = 0$ for all $l \neq j$ which implies that the complexity $k_j$ in evaluating the costs to publish in journal $j$'s does not affect that of competing journal $l$'s.

*In the second period of the publication cost game*, each researcher submits their manuscript to one scholarly journal based on their knowledge of publication costs in the research field (i.e., expert or uninformed authors). The academic journals of low publication costs receive the submissions from the entire mass of expert authors and a fraction $(1/n)$ of the uninformed authors. The rest of journals of higher publication costs receive only a fraction $1/n$ of uninformed authors. Therefore, the journals may influence (through the complexity of publication costs $k_j$) how informed the authors are by affecting the information that they are given. Note that to choose one particular journal, the author should compare every journal's publication costs to all of the competitors. Then, every journal's complexity $k_j$ choice affects the cost of the entire analysis and therefore the fraction of expert authors.

## 3 Optimal journal strategy in the publication cost game

Following Carlin (2009), we now present the optimization problem faced by the academic journals of creating a publication costs structure in the field.

We derive the strategic complexity choices that the journals will employ when setting their publication costs.

Define $J^*$ to be the set of journals who quote the lowest publication costs in equilibrium. Let $n_{J^*}$ be the number of journals in $J^*$, so that the $n_{J^*}$ academic journals in $J^*$ split the demand from the informed authors equally. Each journal $j$ chooses the publication cost $c_j$ and complexity $k_j$ to maximize its expected profit $\Pi_j(c_j, k_j)$ as follows:

$$\max_{c_j, k_j} \Pi_j(c_j, k_j) = c_j Q_j \tag{5}$$

where the expected demand to publish a manuscript from informed and uninformed authors, noted as $Q_j$, is defined as

$$\begin{aligned} Q_j = \ & (\text{the demand from the fraction } \eta \text{ of informed authors}) + \\ & (\text{the demand from the fraction } 1-\eta \text{ of uninformed authors}) \\ = \ & \tfrac{\eta \times 1_{j \in J^*}}{n_{J^*}} + \tfrac{1-\eta}{n}. \end{aligned} \tag{6}$$

where $1_{j \in J^*}$ is a function having the value 1 for all journals of $J^*$ and the value 0 for all journals of $J$ not in $J^*$ and therefore journal $j$ receives $1/n_{J^*}$ of the demand to publish a manuscript from the fraction $\eta$ of informed authors if $j$ is one of the $n_{J^*}$ journals that quote the lowest publication costs in the field; additionally journal $j$ also receives $1/n$ of the demand to publish a manuscript from the fraction $1-\eta$ of uninformed authors.

The following proposition explains how the journals choose their complexity to disclosure the publication costs given their expected relative cost rank-

ing (i.e., the publication costs they choose and the distribution that competing journals use when they set their publication costs).

**Proposition 1.** *In equilibrium, if a scholarly journal has a relatively low publication cost for authors (below a threshold level $\hat{p}$), the optimal strategy will be to choose minimal complexity in the disclosure of publication costs (i.e., $\underline{k}$) to maximize the fraction $\eta$ of informed authors.*

*On the contrary, if a journal has a relatively high publication cost for authors (above level $\hat{p}$), the optimal strategy will be to choose maximal complexity ($\overline{k}$) to maximize the number of uninformed authors.*

Proof: See Appendix A

Academic journals may add complexity to their costs to publish a paper in several ways. First, they can make it more difficult for authors to become informed by partitioning costs into nominal submission fees payable on submission and article processing fees payable only in case of acceptance. This practice makes understanding publication costs more challenging as it places the responsibility on the author to appreciate all of the key costs components and compute the actual costs to publish the manuscript. Second, complexity may be added when journals devise new technical language for their disclosures of the publication costs. If journals in the field use different methods of disclosure, this makes it more difficult for authors to compare cost to publish their work.

From Proposition 1, we have that complexity in the disclosure of publication costs for authors is determined through strategic interaction between

the academic journals. In equilibrium, all journals enjoy a positive rent from having some degree of complexity and preventing some authors from becoming informed about their costs to publish a research work. However, low-fees journals desire less complexity than high-fees journals.

Since the low-fees journals want authors to know that they have the lowest costs to publish their research works, they want fees in the field to be reasonably clear. Adding clarity allows them to undercut their competitors and gain percentage of authors in the field. On the contrary, high-fees journals desire more complexity in the disclosure of publication costs. As fees in the research field becomes more difficult to screen, the fraction of uninformed authors rises, thereby increasing the percentage of authors that high-fees journals receive. According to our model, decreasing publication cost transparency is the way high-fees journals gain percentage of authors in the field.

## 4 Strategic complexity choices when journal competition increases

Again following (Carlin, 2009) we now explore how increasing journal competition affects the complexity in the disclosure of publication costs for authors. Based on the publication cost game as given above, it follows that as more journals compete in the field, the probability that each journal adds complexity in the disclosure of costs rises. The basic idea is that as more scholarly journals compete for the percentage of informed authors who submit manuscripts in the field, any journal's chance of winning new manuscript's submissions decreases. As a result, each academic journal tends to add more complexity in

the disclosure of publication costs, in an attempt to increase the fraction of authors who are uninformed about them. In such a way, they increase their profits in the case that they do not win demand to publish new manuscripts from the group of informed authors. Therefore, the journals may use complexity to preserve their gains in the face of higher competition between scholarly journals in the research field.

Now, in the following, we present the mathematical result that shows as more journals compete, the probability that each journal adds complexity rises.

**Proposition 2.** *In the publication cost game, as more journals compete in the research field, the probability that each journal adds complexity in the disclosure of publication costs for authors rises.*

*As the number of scholarly journals n converges to infinity, all journals choose maximal complexity* $\overline{k}$.

Proof: Following the same arguments that Proposition 1 in Carlin (2009) we have that the ex ante probability that a journal chooses high complexity ($\overline{k}$) in the disclosure of publications costs is uniquely determined in equilibrium and set at

$$\left[\frac{1}{n}\right]^{1/(n-1)}$$

which only depends on the number of journals $n$. For a more general problem, this result is demonstrated in its entirety in the Appendix of Carlin (2009).

Since this probability is increasing in $n$, it follows that each journal's probability of choosing high complexity in the disclosure of their costs is monotonically increasing in the number of journals $n$. In the same way, as the number

of scholarly journals $n$ converges to infinity, all journals choose maximal complexity $\overline{k}$ since the probability $\left[\frac{1}{n}\right]^{1/(n-1)}$ converges to 1.

From Proposition 2 it follows that a higher proportion of journals will add complexity to their costs for authors when there is greater journal competition in the research field. Thus, we can draw a prediction from this result of the publication cost game: Journal concentration in the field and complexity in the disclosure of publication costs should be negatively correlated. As a result, it will be better for authors to have fewer academic journals (and therefore more concentration) than to have more journals (and consequently more complexity).

## 5 Rising journal competition and author knowledge on publication costs

It is important to note that the previous results in Proposition 2 are based on the map of informed authors $\eta$ in equation 4 that is given exogenously in the publication cost game. That is, while the proportion of informed authors may suffer as the result of rising journal competition in the field and complexity in the disclosure of publications costs, this result is not guaranteed to evolve. Analytically, this depends on the particular map of informed authors $\eta$ that we consider.

In particular, when the the fraction of informed authors $\eta$ is a function of the average of the complexity choices for the academic journals in the research field, we can demonstrate the following result.

**Proposition 3.** *In the publication cost game, as journal competition increases, the expected number of informed authors may decrease without optimal regulation in the research field.*

*On the contrary, if the academy or interest groups in the field add educational initiatives, the expected number of informed authors may increase as the number of competitor journals grows.*

Proof: See Appendix B

From Proposition 3, we obtain that the expected number of informed authors is a function of both the complexity choices of the academic journals in the field, as well as factors outside of the model analyzed in the publication cost game. Therefore, rising journal competition may have disparate effects on author's ignorance. Such other factors outside of the model might be educational initiatives or interest groups that help to increase the number of informed authors.

For instance, the Initiative for Open Citations I4OC is a collaboration between scholarly publishers, researchers, and other interested parties to promote the unrestricted availability of scholarly citation data, (I4OC, 2019). As suggested by Proposition 3, part 2, I4OC is one example of the educational initiatives such that help to increase the number of informed authors in all fields of science. Recall that Elsevier's refusal to participate in the Initiative for Open Citation was one of the reasons that led to the resignation of the editorial board of the Journal of Informetrics (JOI), (Larivière, 2018). The former editorial board of the JOI unanimously decided to redirect its labor

to a newly created journal, Quantitative Science Studies (QSS), published by MIT Press and owned by the International Society for Scientometrics and Informetrics, (Larivière, 2018). As an interest group, they invite all authors in the field to join them in this new endeavor and help them demonstrate how the research community can take control back of its means for disseminating knowledge, in a fair, open, and transparent way, (Larivière, 2018).

## 6 Conclusion

Based on the analysis in this paper, we have learned that as journal competition increases, the expected number of informed authors may decrease without optimal regulation in the research field. In fact, we have obtained that as more journals compete in the field, the probability that each journal adds complexity in the disclosure of publication costs for authors rises.

However, using the same model, we demonstrated that the relationship between journal competition and the number of informed authors remains equivocal. That is, while the number of uninformed authors may increase as the result of rising journal competition and complexity in the disclosure of publication costs, this result depends on the particular map of informed authors that we consider.

For instance, when the the fraction of informed authors is a function of the average of the complexity choices for the academic journals in the research field, we have demonstrated that if the academy or interest groups in the field add educational initiatives, the expected number of informed authors can still

increase as the number of competitor journals grows. Therefore, rising journal competition in the field may have disparate effects on the ability of authors to become knowledgeable about the costs to publish a paper.

This key result has important implications given the large number of academic journals in all fields of science. In particular, it highlights the importance of initiatives such as that of the former editorial board of the Journal of Informetrics, an Elsevier journal. The members of the editorial board of JOI, have unanimously resigned and have moved to the new journal Quantitative Science Studies (QSS), published by MIT Press. An important reason for the resignation is Elsevier's lack of support for the Initiative for Open Citations as well as disagreements about open access policies. From Proposition 3, part 2, we can conclude that this is a major step for the field of quantitative science studies.

**Acknowledgments.** This research was sponsored by the Spanish Board for Science, Technology, and Innovation under grant TIN2017-85542-P, and co-financed with European FEDER funds.

## References

Carlin, Bruce I. (2009). Strategic price complexity in retail financial markets. Journal of Financial Economics, 91(3), pp. 278-287. https://doi.org/10.1016/j.jfineco.2008.05.002

Dasgupta, P., and Maskin, E., (1986). The existence of equilibrium in discontinuous economic games, I and II: Theory and applications. Review

of Economic Studies, 53(1), pp. 1-41. https://doi.org/10.2307/2297588 https://doi.org/10.2307/2297589

Initiative for Open Citations, (2019). https://i4oc.org/

Larivière, Vicent (2018). Blog post by Vincent Larivière (Interim Editor, QSS). http://issi-society.org/blog/posts/2019/january/resignation-of-the-editorial-board-of-the-journal-of-informetrics/

Nash, John (1951). Non-Cooperative Games. The Annals of Mathematics, 54(2), pp. 286-295

Panter, Michaela, (2019). Understanding Submission and Publication Fees. AJE Scholar, https://www.aje.com/en/arc/understanding-submission-and-publication-fees/?cv=1

## A Proof of Proposition 1

The proof of Proposition 1 follows from results in (Carlin, 2009). The outline of the arguments used there is as follows. From equation (5), the payoff function for each academic journal is continuous, except when its cost to publish a paper is the lowest and the same to one of the other journals. Then, the journal can increase this payoff by lowering their cost to publish a paper. It is possible to show, however, that each journal's payoff function is indeed weakly lower semi-continuous when its cost is the lowest and equal to at least one of its competitors, (Carlin, 2009).

Additionally, since the sum of the payoffs to all of the journals in $J$ is a continuous function of any one journal's cost, the publication cost game

satisfies the conditions that are required for the existence of a symmetric mixed-strategy Nash equilibrium as outlined by (Dasgupta and Maskin, 1986).

It is then possible to show that the optimal complexity choice for each journal only depends on its own publication cost and the distribution of costs that competing journals use to mix over publication costs.

To sum up, the proof is similar to that of Prop. 1 in Carlin (2009). So we skip it in this communication since it is given in its entirety in the Appendix of Carlin (2009).

**B Proof of Proposition 3**

Again we follow (Carlin, 2009) to prove Part 1 in the proposition. So, consider that

$$\eta = 1 - \frac{1}{n}\sum_{j=1}^{n} k_j \tag{7}$$

with $k_j \in [\underline{k}, \overline{k}]$, and where $0 < \underline{k} < \overline{k} < 1$. It is straightforward to verify that $\eta$ satisfies the conditions defined in Section 2. Let $\alpha = \overline{k} - \underline{k}$. It follows then from equation (7) that $\eta_{max} - \eta_{min} = \alpha$.

Since each journal's complexity choice is binary and the probability of adding high complexity is $\left[\frac{1}{n}\right]^{1/(n-1)}$, it follows that the expected fraction of informed authors in the research field is computed as:

$$E[\eta] = \eta_{max} - \sum_{m=0}^{n} \frac{n!}{m!(n-m)!}\frac{m\alpha}{n}\left(\left[\frac{1}{n}\right]^{1/(n-1)}\right)^{m}\left(1 - \left[\frac{1}{n}\right]^{1/(n-1)}\right)^{n-m} \tag{8}$$

with $m$ being the number of journals choosing to add high complexity to the disclosure of production costs. Since $E[\eta]$ in equation (8) is $\eta_{max}$ minus the

expectation of a binomial variable, it follows that

$$E[\eta] = \eta_{max} - \frac{\alpha}{n} n \left[\frac{1}{n}\right]^{1/(n-1)} = \eta_{max} - \alpha \left[\frac{1}{n}\right]^{1/(n-1)} \tag{9}$$

which is decreasing in $n$. Consequently, as journal competition in the research field $n$ rises, the expected number of informed authors decreases.

We also follow (Carlin, 2009) to prove Part 2 in the proposition. Now, consider that

$$\eta = 1 - \frac{1}{n^2} \sum_{j=1}^{n} k_j \tag{10}$$

with $k_j \in [\underline{k}, \overline{k}]$, and where $0 < \underline{k} < \overline{k} < 1$. Again $\eta$ satisfies the conditions defined in Section 2, and consider $\alpha = \overline{k} - \underline{k}$. In this case, we have that $\eta$ is more sensitive to changes in $n$ than before, because this construction of $\eta$ captures the idea that the academy or interest groups in the field add educational initiatives for disseminating knowledge, in a fair, open, and transparent way about the publication costs, as the number of journals grows in size.

As before, each journal's complexity choice is binary and the probability of adding high complexity is $\left[\frac{1}{n}\right]^{1/(n-1)}$. Then, the expected fraction of informed authors in the research field is computed as:

$$E[\eta] = \eta_{max} - \sum_{m=0}^{n} \frac{n!}{m!(n-m)!} \frac{m\alpha}{n^2} \left(\left[\frac{1}{n}\right]^{1/(n-1)}\right)^m \left(1 - \left[\frac{1}{n}\right]^{1/(n-1)}\right)^{n-m} \tag{11}$$

where again $m$ is the number of journals choosing to add high complexity to the disclosure of production costs. Therefore, in this case, it follows that the expected fraction of informed authors is

$$E[\eta] = \eta_{max} - \frac{\alpha}{n^2} n \left[\frac{1}{n}\right]^{1/(n-1)} = \eta_{max} - \alpha \left[\frac{1}{n}\right]^{n/(n-1)} \tag{12}$$

which is now increasing in $n$. Consequently, as journal competition in the research field $n$ rises, the expected number of informed authors increases, despite the increased tendency for academic journals to add complexity to the disclosure of publication costs.

## 3.3. What is the sensitivity and specificity of the peer review process?

### 3.3.1. Datos generales

1. **Autores**: Jose A. Garcia, Jorge Chamorro-Padial, Rosa Rodríguez-Sánchez y J.Fdez-Valdivia.
2. **Revista**: Accountability in Research.
3. **Datos sobre la publicación**:
   - **Referencia**: García et al. (2022).
   - **Páginas**: 1–22.
   - **Año**: 2022.
   - **Editorial**: Taylor & Francis.
   - **DOI**: https://doi.org/10.1080/08989621.2022.2122817.
4. **Estado**: Publicado.
5. **Métricas**:
   - **Factor de impacto**: 3,057 (JCR, 2021).
   - **Ranking**:
     - *Social Science Citation Index Expanded (SSCIE)* [1]:
       - *Medical Ethics*: Q2 - 5/16 (año 2021).

### 3.3.2. Contribuciones principales

1. Se definen los conceptos de sensibilidad y especificidad aplicados al proceso de revisión por pares.
2. Mediante la aplicación de la probabilidad Bayesiana, se propone un modelo de comportamiento de un editor, con un sesgo de comportamiento inicial que varía en base a los comentarios recibidos por los revisores.

### 3.3.3. Resumen

En el proceso de revisión por pares, definimos la sensibilidad (*sensitivity*) como la probabilidad de que un manuscrito reciba una recomendación favorable por parte de un revisor, cumpliendo el manuscrito, además, con

[1] A fecha de depósito de esta tesis, aún no se disponen de datos del año 2022.

los criterios de calidad necesarios para ser publicado en una revista determinada. Del mismo modo, definimos la especificidad (*specificity*) como la probabilidad de que un manuscrito reciba una recomendación negativa tras una revisión, siendo el manuscrito de calidad no aceptable para una revista en concreto.

Cuando un artículo es enviado a una revista, el editor tiene una idea preconcebida sobre la calidad del artículo (en ocasiones, esta idea inicial puede provocar una decisión sobre el artículo sin que este sea enviado a revisión por pares). Un proceso de revisión por pares donde exista un nivel muy elevado de especificidad, por ejemplo, permitirá a los editores aceptar un artículo que reciba una recomendación favorable por parte de los revisores, ya que podemos tener la *seguridad* de que el proceso de revisión tiende de forma general a rechazar artículos que no cumplan con los umbrales mínimos de calidad deseada.

En este artículo, presentamos un modelo que permite inferir el nivel de sensibilidad y de especificidad de un proceso de revisión por pares. Para este trabajo, nos basamos en tres factores clave que tienen lugar durante una revisión por pares:

1. El editor no sabe, de antemano, cuál será el resultado del proceso de revisión por pares, pero tiene una creencia inicial sobre la calidad de un artículo.

2. En base a esta creencia, es posible que el editor necesite obtener más información sobre un artículo, recurriendo al proceso de revisión por pares. Este proceso supone un *coste* adicional.

3. Es posible que la información obtenida tras la revisión, siga siendo incompleta para tomar una decisión adecuada.

Nuestra propuesta recurre al modelo de probabilidad Bayesiana. Por un lado, $X$ hace referencia a la calidad del manuscrito, siendo X = 1 una situación en la cual el manuscrito reúne los requisitos de calidad de la revista a la que ha sido sometido, y X = 0 si no reúne estos requisitos de calidad. El editor tiene una idea preconcebida sobre la validez de un artículo, esta creencia supone una probabilidad *a priori* de que el artículo sea, aceptado ($q = P(X = 1)$), o no ($1 - q = P(X = 0)$). Tras el proceso de revisión por pares, el editor recibe una señal, S, por parte de los revisores. Esta señal es binaria ($S \in \{0, 1\}$) siendo S = 1 una señal de aceptación, y S = 0 una señal de rechazo al manuscrito que está siendo sometido.

Finalmente, la sensibilidad del proceso de revisión se expresa como $\delta_1 = P(S = 1|X = 1)$, mientras que la especificidad se describe como $\delta_0 = P(S = 0|X = 0)$.

Los revisores, normalmente, presentan un sesgo hacia los comentarios: bien favorables, o bien desfavorables. Cuando un editor presta más atención

a los comentarios favorables, es más probable que la decisión final sobre un artículo sea más acertada cuando la sensibilidad es mayor que la especificidad. Del mismo modo, si un editor da más importancia a los comentarios negativos, la decisión sobre la publicación de un manuscrito será más adecuada si la especificidad es mayor que la sensibilidad.

Los comentarios de los revisores, además, cambian la creencia original que un editor tenía sobre la calidad de un artículo. Pudiendo rebajar o elevar las expectativas sobre el mismo.

# What is the sensitivity and specificity of the peer review process?

J. A. García, *
Jorge Chamorro-Padial, †
Rosa Rodriguez-Sánchez, ‡
J. Fdez-Valdivia. §

September 21, 2022

**Abstract**

In this paper, we introduce the concepts of sensitivity and specificity to mathematically describe the accuracy of the peer review process. Sensitivity refers to the probability that the final decision for a manuscript would be acceptance, provided the manuscript meets the journal standards required for publication (i.e., true positive rate). Specificity refers to the probability that the final decision would be rejection, provided the work does not meet the standards required for publication (i.e., true negative rate). Therefore, in the peer review process, sensitivity measures the ability to correctly accept manuscripts that meet the required standards (true positives) and specificity measures the ability to correctly reject manuscripts that do not meet those

*Departamento de Ciencias de la Computación e I. A., CITIC-UGR, Universidad de Granada, 18071 Granada, Spain. Address correspondence to J. A. Garcia at jags@decsai.ugr.es. Jose A. Garcia ORCID iD https://orcid.org/0000-0001-7742-7270

†CITIC-UGR, Universidad de Granada, 18071 Granada, Spain. Address correspondence at jorgechp@correo.ugr.es

‡Departamento de Ciencias de la Computación e I. A., CITIC-UGR, Universidad de Granada, 18071 Granada, Spain. Address correspondence at rosa@decsai.ugr.es.

§Departamento de Ciencias de la Computación e I. A., CITIC-UGR, Universidad de Granada, 18071 Granada, Spain. Address correspondence at jfv@decsai.ugr.es.

quality standards required for publication (true negatives). Sensitivity and specificity values can inform the editor under what conditions the outcome of a peer review process becomes more precise and, therefore, if this does not occur, when the editor must improve the analysis involved in processing the information received from reviewers' reports. Sensitivity and specificity understood in this way can promote the ethical conduct of peer review processes and improve the validity of manuscript editorial decisions.
*Keywords: Peer Review; Sensitivity; Specificity; Bayesian Inference; Mutual Information.*

# 1 Introduction

The existence of editorial biases is undeniable in academic publishing. For instance, publication bias refers to those situations in which the outcome of a research study presented in the manuscript has a bearing on the editorial decision of acceptance or rejection (Dickersin et al., 1987, 2002). Several studies found that more manuscripts with positive results are accepted compared to those with negative results just because they are positive or negative, e.g., (von Elm et al., 2008; Stern and Simes, 1997). Looking beyond the results presented in the manuscript, prior studies have shown editorial biases associated with factors such as the origin or characteristics of the authors, the country of origin, the institution that produces it, the previous articles published by the returning authors, or the authors' native language (Yousefi-Nooraie et al., 2006; Patel and Sumathipala, 2001; Mendis et al., 2003; Keiser et al., 2004; Garfunkel et al., 1994; Figg et al., 2006; Okike et

al., 2008). Therefore, to address the important scientific problem posed by these types of biases, there is a need to study and develop methods and systems that can promote the ethical conduct of peer review processes and improve the validity of editorial decisions.

In a different situation, diagnostic and screening tests are two kinds of medical tests that can be used to correctly detect a target disease or condition in patients, (Geraint et al., 2008). However, while the presence or absence of a condition is definitively determined using diagnostic tests, screening tests only serve to identify those people who have or may develop a target disease, (Murad et al., 2017a,b). Although screening tests are therefore imperfect and can produce ambiguous results in certain cases, they are generally less invasive and dangerous, less expensive, faster, and definitely less uncomfortable for people.

Screening tests are therefore less reliable tests but also simpler to perform both for the medical system and for the person themselves (Geraint et al., 2008; Murad et al., 2017b). Nevertheless, the problem is how to determine the accuracy of screening tests when used to identify the probable presence or absence of a target disease. Their findings can only be guaranteed when such information is available. To this end, in a screening situation, a test's sensitivity represents its ability to correctly identify those people who really are sick, (Yerushalmy, 1947). Furthermore, a test's specificity represents its

ability to correctly identify people who really are healthy, (Parikh et al., 2008; Altman and Bland, 1994).

In our situation, a journal editor wants to make a final decision on a manuscript submitted for publication (i.e., acceptance or rejection). This editorial decision must be based on whether the research meets the quality standards required for publication (Burnham, 1990; Chubin & Hackett, 1990; Bornmann, 2008, 2011). However, the journal editor does not know about this condition in advance, although they could have an initial belief about the quality of the research work. When the editor's prior belief that the manuscript meets the required journal standards for publication is very low or very high, the editor would likely make a desk decision without external review. Otherwise, when the editor's prior belief that the manuscript meets the journal standards required for publication is neither too high nor too low, the manuscript will be sent for external review. In this situation, based on the information gathered from the peer review process and after reading the reviewers' reports, and perhaps discussing them with other experts, the editor in charge receives a binary signal, (Burnham, 1990; Chubin & Hackett, 1990; Garcia et al., 2015). The signal received by the editor from the peer review process changes the editor's belief from an initial probability to a posterior probability. More precisely, if the editor receives an 'accept' signal from the peer review process, they increase their belief that the manuscript

meets the journal standards. On the contrary, if the editor receives a 'reject' signal from the peer review process, they decrease their initial belief. Using this signal that represents the aggregation of favorable and unfavorable information about the manuscript, the editor can then update their knowledge about the match between the manuscript and the journal standards required for publication (Garcia et al., 2015, 2019).

As suggested in Abby et al. (1994), this peer review processing represents a screening situation performed on a large number of manuscripts to identify those that are likely to meet the journal standards required for publication. For instance, (Abby et al., 1994) studied "the effectiveness of peer review as a screening process to evaluate medical manuscripts." Therefore, it makes sense to introduce the concepts of sensitivity and specificity to determine the extent to which a peer review process (interpreted as a screening test) is able to identify whether a manuscript does or does not meet the journal standards required for publication. In this scenario, sensitivity and specificity represent the accuracy of the peer review outcome (i.e., an 'accept' or 'reject' signal). On one hand, sensitivity is the probability of an 'accept' outcome from the peer review process, provided the manuscript meets the journal standards required for publication. On the other hand, specificity is the probability of a 'reject' outcome from the peer review process, provided the research work does not meet the standards required for publication.

Therefore, a very high level of sensitivity (e.g., sensitivity $> 0.95$) permits manuscripts to be confidently rejected by the journal editor if the peer review process yields a 'reject' outcome. This is so because a highly sensitive peer review process is unlikely to reject manuscripts that meet the journal standards. Furthermore, a very high level of specificity (e.g., specificity $> 0.95$) permits manuscripts to be confidently accepted by the editor if the review process yields an 'accept' outcome. This is so due to the fact that a peer review process with a high specificity is unlikely to accept manuscripts that do not meet the journal standards required for publication.

When the journal editor pays more (or less) attention to favorable comments from reviewers than to unfavorable ones, they are choosing the level of sensitivity and specificity of the peer review process. For instance, if the editor allocates more attention to unfavorable comments than favorable ones, the peer review outcome would be more accurate when the research work does not meet the journal standards required for publication than when it does. Due to the fact that they may ignore some positive attributes of the manuscript when making a final decision. In this situation, it follows that specificity is greater than sensitivity. On the contrary, when the editor pays more attention to positive comments from reviewers than negative ones, the peer review outcome is relatively more accurate when the manuscript meets the journal standards required for publication than when it does not (i.e.,

sensitivity is greater than specificity). This is so because they may ignore some negative attributes of the research when making a final decision.

In the following, we present a more formal definition of sensitivity and specificity in a peer review process. Then, we discuss how an editor can find the optimal sensitivity and specificity of the peer review outcome by trading off the value of reviewers' positive and negative comments with the cost of analyzing and understanding those pieces of information. The three key features in our model are as follows: (i) editors have prior beliefs that a manuscript does or does not meet the quality standards required for publication in the journal; (ii) some costly information acquisition from reviewers' reports may be necessary before an editorial decision is made; and (iii) the information acquired about the match between the manuscript and the quality standards required for publication is likely to be incomplete. Finally, in this paper, we also present a computational tool that helps to calculate the sensitivity and specificity of a peer review process (see Figure 1). Using this computational tool we will then discuss the results of our model.

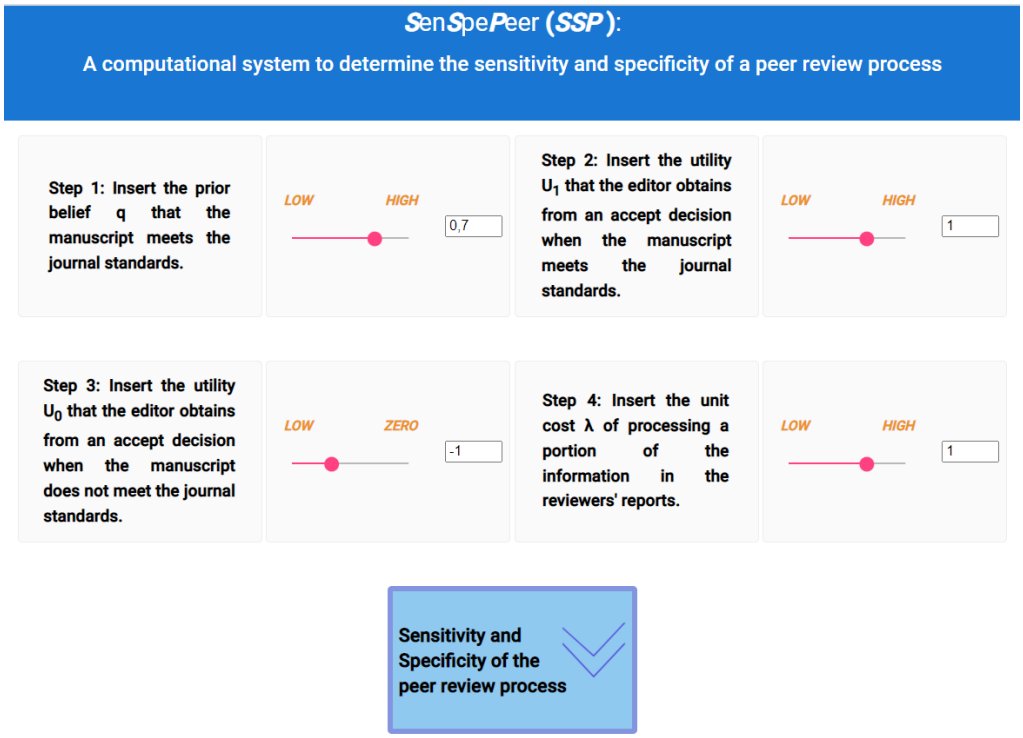



Figure 1: A computational tool that helps to calculate the sensitivity and specificity of a peer review process.

## 2 The sensitivity and specificity of a peer review process

An editor wants to make a final decision on a manuscript, based on whether the research meets the journal standards required for publication. Let $X$ be a state variable which only takes two actual values, i.e., $X$ is either 1 or 0. If $X = 1$, the manuscript meets the journal standards, and the editor gets a utility of $U_1$ from an 'accept' decision. Otherwise, if $X = 0$, the manuscript does not meet the quality standards required for publication, and the editor gets a utility of $U_0$ from an 'accept' decision, with $U_0 < 0 < U_1$. We assume that a 'reject' decision on the manuscript has a utility of 0.

The problem is that the journal editor does not know in advance the true value of $X$. However, an editor can have an initial hunch about the manuscript's quality so they would have a prior belief $q$ about the real value of $X$, which is given by a baseline probability $q = P(X = 1)$. Only after the manuscript is published will the journal editor learn the true value of $X$.

We should state that our Bayesian analysis requires a prior probability, and these are sometimes difficult to formulate. This means that the analysis is personal to one editor, anyone else observing the same data has to form personal conclusions. Therefore, Bayesian inference is logically coherent but only conditional on the assumed probability model for a given editor. When

making decisions under uncertainty in a situation like ours with repeated, clear, relatively immediate feedback and strong incentives to be right, humans used to be excellent Bayesians. However, when making decisions in other situations, humans can be poor Bayesians. So, are there any alternatives to Bayesian optimization? Yes, there are, for example, Gradient Descent but there we need to define a differentiable objective function, Genetic Algorithms where we then need to define some objective function, or Spectral methods as Hidden Markov models. In this paper, we will use Bayesian inference and, therefore, the implicit assumption is that, some initial distribution exists from which conditional probabilities can be derived using the rules of probability theory.

Based on the information gathered from the peer review process, the editor in charge receives a binary signal $S \in \{0, 1\}$. If $S = 1$, the editor obtains an 'accept' signal from the peer review process, and they increase their belief that the manuscript meets the journal standards from prior $q = P(X = 1)$ to posterior $P(X = 1|S = 1) > q$, with $P(X = 1|S = 1)$ denoting the probability that the actual value of $X$ is 1, conditional on the signal $S$ obtained being 1.

On the contrary, if $S = 0$, the editor obtains a 'reject' signal from the peer review process, and they decrease their belief that the manuscript meets the journal standards from prior $q = P(X = 1)$ to posterior $P(X = 1|S = 0) < q$,

with $P(X = 1|S = 0)$ denoting the probability that the actual value of $X$ is 1, conditional on the signal $S$ obtained being 0. Therefore, the signal $S$ received by the editor from the peer review process changes the editor's belief from an initial probability to a posterior probability. Using this signal $S$ that represents the aggregation of favorable and unfavorable information about the manuscript, the editor can then update their knowledge about the state variable $X$ (i.e., the match between the manuscript and the journal standards required for publication).

In our model, the sensitivity and specificity of a peer review process refer to the accuracy of the review outcome (i.e., signal $S$). On one hand, a peer-review's sensitivity $\delta_1$ is given by the probability of an 'accept' signal from the peer review, $S = 1$, provided that the manuscript meets the journal standards required for publication, i.e., $X = 1$,

$$\delta_1 = P(S = 1|X = 1).$$

On the other hand, a peer-review's specificity $\delta_0$ is given by the probability of a 'reject' signal from the peer review process, $S = 0$, provided that the research work does not meet the standards required for publication, i.e., $X = 0$,

$$\delta_0 = P(S = 0|X = 0).$$

Editors need to spend time and cognitive effort to be able to analyze and understand the information contained in the reviewers' reports. Therefore, editors incur non-trivial costs during the peer review process and must optimize the time and effort they spend in gathering and processing information about the manuscript's quality. From Cover and Thomas (2006); Garcia et al. (2019), it follows that the cost of information (about that quality) becomes increasingly more expensive as this information gets more precise. Therefore, editors could choose to be only partially informed when making a decision on the manuscript.

Since spending a great deal of time and effort to collect all the information about a manuscript's quality is never optimal, the information acquired to make an editorial decision on the manuscript is not likely to be complete. Hence, we allow the journal editor to pay more (or less) attention to favorable comments from reviewers than to unfavorable ones. In this context, we define favorable (or unfavorable) comments from reviewers as those that are likely to inform aspects of the manuscript that increase (or decrease) the editor's belief that the work meets the standards required by the journal for publication. Thus, as a consequence, if the editor allocates more attention to unfavorable comments than favorable ones, the peer review outcome (signal $S$) would be more accurate when the research work does not meet the journal standards required for publication ($X = 0$) than when it does ($X = 1$). Under this

scenario, it follows that

$$\delta_0 = P(S = 0|X = 0) > P(S = 1|X = 1) = \delta_1$$

and we have that specificity ($\delta_0$) is greater than sensitivity ($\delta_1$). This is so because, if a journal editor spends less time and effort analyzing and understanding favorable comments from reviewers than unfavorable ones, they may ignore some positive attributes of the manuscript when making a final decision.

On the contrary, when the editor pays more attention to positive comments from reviewers than negative ones, the peer review outcome (signal $S$) is relatively more accurate when the manuscript meets the quality standards required for publication ($X = 1$) than when it does not ($X = 0$). Under this scenario, it follows that

$$\delta_1 = P(S = 1|X = 1) > P(S = 0|X = 0) = \delta_0$$

and we have that sensitivity ($\delta_1$) is greater than specificity ($\delta_0$). This is because, if a journal editor spends less time and effort analyzing and understanding unfavorable comments than favorable ones, they may ignore some negative attributes of the research when making a final decision. Then, the editor's choice of attention allocation determines the sensitivity and speci-

ficity of the peer review outcome $S$.

Therefore, in our problem, paying more attention to positive comments from reviewers than negative ones means spending more time and efforts analyzing and understanding favorable comments than unfavorable ones, which produces a peer review outcome (signal $S$) for which sensitivity ($\delta_1$) is greater than specificity ($\delta_0$). Hence, the editor chooses $\delta_1 > \delta_0$ (i.e., sensitivity greater than specificity) if they pay more attention to positive comments from reviewers than negative ones, or, conversely, chooses $\delta_1 < \delta_0$ (i.e., specificity greater than sensitivity) when the attention is reversed.

# 3 The optimal sensitivity and specificity of the peer review process

In this section, we discuss how an editor can find the optimal sensitivity and specificity of a peer review outcome, $S$. To this end, the journal editor trades-off the value of reviewers' positive and negative comments with the cost of analyzing and understanding that information.

## 3.1 The value of reviewers' positive and negative comments

Upon receipt of the reviewers' reports, the editor evaluates them to obtain a signal $S$ and decides the future of the manuscript. More precisely, an 'accept' signal ($S = 1$) changes the belief of the journal editor from the initial probability $q = P(X = 1)$ to a posterior probability $P(X = 1|S = 1)$. This posterior probability can be formulated in terms of the sensitivity and specificity of $S$ (i.e., $\delta_1$ and $\delta_0$) as follows

$$P(X = 1|S = 1) = \frac{q\delta_1}{q\delta_1 + (1 - q)(1 - \delta_0)}$$

by using Bayes' Theorem. Similarly, a 'reject' signal received by the editor from the reviewers' comments ($S = 0$) changes the belief of the editor from the initial probability $q = P(X = 1)$ to a posterior probability

$$P(X = 1|S = 0) = \frac{q(1 - \delta_1)}{q(1 - \delta_1) + (1 - q)\delta_0}$$

by using Bayes' Theorem. In this framework, the editor can choose the values of sensitivity and specificity for the review outcome $S$ by processing different amounts of favorable and unfavorable comments from reviewers. However, the chosen values of sensitivity and specificity must verify that $\delta_0 + \delta_1 > 1$

in order to get

$$P(X = 1|S = 0) < q < P(X = 1|S = 1).$$

This is so because, as defined, if the journal editor receives an 'accept' signal, $S = 1$, they increase their belief that the manuscript meets the journal standards, and when the editor receives a 'reject' signal, $S = 0$, they decrease this belief. If we allow that $\delta_0 + \delta_1 = 1$, we obtain

$$P(X = 1|S = 0) = P(X = 1|S = 1) = q,$$

and therefore the editor does not receive any additional information from the peer review process. Furthermore, if $\delta_0 + \delta_1 < 1$, we obtain

$$P(X = 1|S = 0) > q > P(X = 1|S = 1),$$

which is not possible given the definitions of 'accept' and 'reject' signals.

In our model, we assume that the editor makes a final decision to accept the manuscript when they receive an accept signal ($S = 1$) from the reviewers' reports. Otherwise, the editor decides to reject the research work when they receive a reject signal ($S = 0$). Hence, the editor receives an accept signal from the peer review process and makes a final decision of acceptance with

probability $P(S = 1)$. In this scenario, the editor gets a posterior utility of $U_1P(X = 1|S = 1) + U_0P(X = 0|S = 1)$, with utilities $U_0$ and $U_1$ as defined above. Otherwise, the editor receives a reject signal from the peer review process and makes a final decision of rejection with probability $P(S = 0)$. We assume that a reject decision on the manuscript has a utility of 0.

Following this analysis, the value of reviewers' positive and negative comments ($EV$) is given by the expected utility that a journal editor obtains by using the signals received from those comments (i.e., either $S = 1$ or $S = 0$)

$$EV = q\delta_1 U_1 + (1 - q)(1 - \delta_0)U_0 \tag{1}$$

as shown in Appendix A.

From this equation, we get that the value of reviewer's comments increases as the sensitivity and specificity of signal $S$ increase (i.e., the peer review outcome is more accurate). Please see Appendix A for further details.

If the journal editor makes a desk decision without external review (i.e., the manuscript is not sent out to expert reviewers for evaluation and therefore the editor does not receive any $S$ signal from external reviewers), then they make a final decision of acceptance if and only if

$$qU_1 + (1 - q)U_0 \geq 0$$

where the editor's belief that the manuscript meets the journal standards does not change from the prior probability $q$.

## 3.2 The cost of analyzing and understanding reviewers' comments

Now again, suppose that the editor receives a signal $S$ when they process the reviewers' reports. Based on this peer review outcome $S$, the editor updates their belief that the manuscript meets the journal standards for publication from a baseline probability $p(X)$ to $p(X|S)$. It then reduces the level of uncertainty with regard to the actual value of $X$ using the peer review outcome $S$.

Following (Cover and Thomas, 2006; Garcia et al., 2020), "the mutual information $I(X, S)$ can be used to measure this uncertainty reduction that comes from (the peer review outcome) $S$." The more the editor wants to reduce the uncertainty about $X$, the more reviewers' comments will have to be analyzed and understood. The mutual information $I(X, S)$ quantifies the exact amount of information learned about $X$ that is contained in $S$, or, more intuitively, how many reviewers' comments the editor has to process to update their belief from $p(X)$ to $p(X|S)$.

In our model, we follow several studies that adopt the mutual information "to quantify the cost of processing information" (e.g., Jerath and Ren (2021);

Sims (2003, 2006)). More precisely, the cost of analyzing and understanding reviewers' comments to receive a signal $S$ is given by the cost of updating the editor's belief from an initial probability $p(X)$ to the posterior probability $p(X|S)$ as follows

$$\text{Cost} = \lambda I(X, S)$$

with $\lambda > 0$ being the unit cost of analyzing and understanding a piece of the information in the reviewers' reports, and $I(X, S)$ denoting the mutual information between $X$ and $S$ as given above (see Cover and Thomas (2006) for further details). The cost of analyzing the information in the reviewers' reports ($\lambda$) increases with the level of noise in the signal received from the reviewers. Furthermore, it increases with the complexity of the manuscript review process. Moreover, this cost also increases as a consequence of the choice of weights that the editor would have to assign to the comments provided by the different types of reviewers (e.g., based on the prestige or expertise of the scholar).

A potential issue with this approach is that $I(X, S)$ represents the most efficient processing case, but editors may not process information from reviewers' reports in the most efficient possible manner. However, in our model, if an editor has greater limitations on time and resources and, therefore, they do not process information from reviewers' reports in the most efficient manner, then the unit cost $\lambda$ becomes higher to be able to represent the in-

formation that editors actually process. Furthermore, the cost of processing information in an editor setting is not necessarily proportional to the reduction of entropy and, therefore, other meaningful metrics (e.g., the reduction of the variance of the belief distribution) can also be used to quantify the cost of processing information from reviewers' reports. However, the cost function in our model, which is based on mutual information has several interesting properties as shown in Appendix B:

- If the editor wants to increase the sensitivity or specificity of the peer review process, it will become progressively costlier (and thus require more and more effort).
- If the editor processes a larger number of positive comments from reviewers and thus the sensitivity is higher, the marginal cost required to analyze and understand the next negative comment also increases, and vice versa.
- If the editor has increased the sensitivity of the peer review process significantly by processing a large number of positive comments from reviewers, then it is marginally less costly to increase the specificity than the sensitivity because understanding a negative comment next requires relatively less effort than a positive comment would.

## 3.3 The optimal sensitivity and specificity of a peer review process

An editor's choice of how much time and effort they spend processing favorable and unfavorable comments determines their decision on the sensitivity and specificity of a peer review, i.e., $\delta_0$ and $\delta_1$.

The editor might prefer to make a desk decision without external review. However, when the editor decides to send the manuscript out to external review, they find the optimal sensitivity and specificity, $\delta_1$ and $\delta_0$, by trading-off the value of reviewers' positive and negative comments, $EV$, with the cost of their processing, $Cost$, as follows

$$\text{Maximize the value of } (EV - Cost) \tag{2}$$

subject to $\delta_0 + \delta_1 > 1$, with $EV$ and $Cost$ as given above.

When the editor's prior belief that the manuscript meets the journal standards required for publication, $q$, is neither too high nor too low, the optimal sensitivity and specificity of the peer review process ($\delta_1^*$ and $\delta_0^*$) are found by maximizing equation (2) as shown in Appendix B. Furthermore, the journal editor accepts the manuscript upon receiving the peer review outcome $S = 1$ with probability $\Pr(S = 1)$. Please see Appendix B for further details.

However, the editor could exhibit constraints or stickiness in terms of

editor's beliefs that would effect their ability to change from an initial probability to a posterior probability. In this situation, the editor's prior belief that the manuscript meets the journal standards is likely to be either very low or very high, and our model finds that the editor makes a desk decision without external review. The proof for this mathematical result is also shown Appendix B.

Given the importance of determining the accuracy of a peer review process, we developed a computational tool that is available at `https://blackcat.ugr.es/senspepeer/`. It can be used to calculate both sensitivity and specificity for the review process (see Figure 1). When using this online tool, the editor inserts their initial belief $q$ that the manuscript meets the journal standards required for publication. For instance, in Figure 2, $q = 0.54$. They then insert the utility $U_1$ and $U_0$ that the editor gets from an accept decision when the manuscript does and does not meet the journal standards for publication, respectively (e.g., $U_1 = 1.3$ and $U_0 = -1.5$ in Figure 2). Finally, the editor inserts the unit cost $\lambda$ of processing a portion of the information in the reviewers' reports (e.g., $\lambda = 0.6$ in Figure 2).

The online tool then calculates the sensitivity $\delta_1^*$ and the specificity $\delta_0^*$ of the peer review process. For the example in Figure 2, we obtain $\delta_1^* = 0.91$ and $\delta_0^* = 0.92$ as shown in Figure 3. For this setting, the online tool predicts that the peer review outcome is relatively more accurate when the manuscript does


Figure 2: (i) Editor inserts their prior belief $q = 0.54$; (ii) they also insert utility $U_1 = 1.3$; (iii) utility $U_0 = -1.5$; and (iv) the unit cost $\lambda = 0.6$.

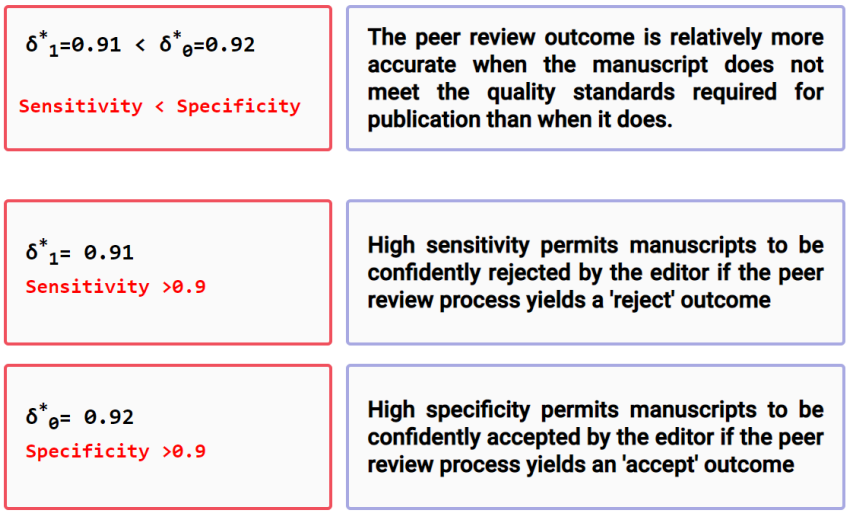



Figure 3: Sensitivity and specificity for the example in Figure 2

not meet the quality standards required for publication than when it does (i.e., specificity is greater than sensitivity). This is illustrated in Figure 3. It also predicts that the high level of sensitivity ($\delta_1^* > 0.9$) permits manuscripts to be confidently rejected by the journal editor if the peer review process yields a reject signal (see Figure 3). Due to the fact that a peer review process with a high level of sensitivity is unlikely to reject manuscripts that meet the quality standards required for publication. Similarly, it predicts that the high level of specificity ($\delta_0^* > 0.9$) permits manuscripts to be confidently accepted by the editor if the peer review process yields an accept signal (see Figure 3). This is so because a peer review process with a high level of specificity

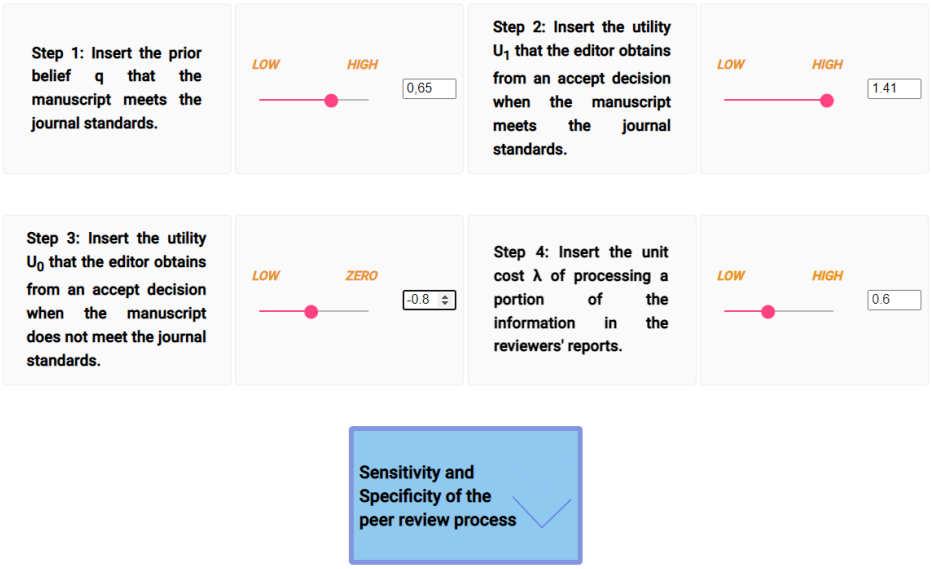



Figure 4: (i) Editor inserts their prior belief $q = 0.65$; (ii) they also insert utility $U_1 = 1.41$; (iii) utility $U_0 = -0.8$; and (iv) the unit cost $\lambda = 0.6$.

is unlikely to accept manuscripts that do not meet the journal standards for publication.

Figures 4 and 5 show the performance of the SenSpePeer tool for a different example. In Figure 4, the initial belief that the manuscript meets the journal standards required for publication is $q = 0.65$. The utility that the editor obtains from an accept decision when the manuscript does and does not meet the journal standards is $U_1 = 1.41$ and $U_0 = -0.8$, respectively. Furthermore, the unit cost of processing a portion of the information in the reviewers' reports is $\lambda = 0.6$.

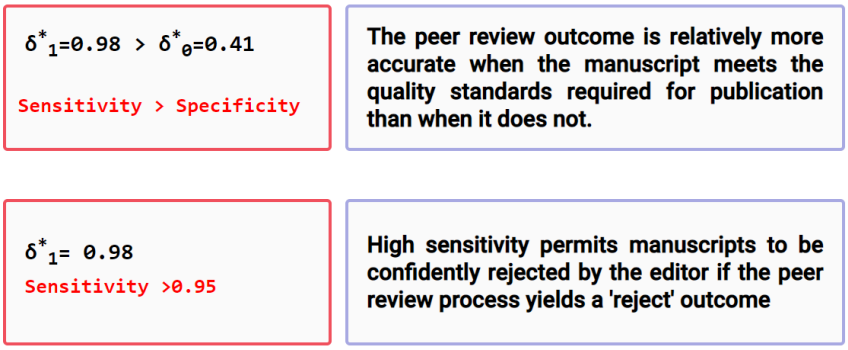



Figure 5: Sensitivity and specificity for the example in Figure 4

Figure 5 illustrates that the sensitivity and the specificity of the peer review process are $\delta_1^* = 0.98$ and $\delta_0^* = 0.41$, respectively. Therefore, the free online tool now predicts that the peer review outcome is relatively more accurate when the manuscript meets the quality standards required for publication than when it does not (i.e., sensitivity is greater than specificity). This is illustrated in Figure 5. It also predicts that the high level of sensitivity ($\delta_1^* > 0.95$) allows for manuscripts to be confidently rejected by the journal editor if the peer review process yields a reject signal (see Figure 5). Again, due to the fact that a review process with a very high level of sensitivity is unlikely to reject research works that meet the journal standards.

# 4 Discussion

In order to analyze the behavior and utility of the sensitivity and specificity of a peer review process, we also performed an experiment varying the values of $q$, $U_1$, $U_0$, and the unit cost $\lambda$. In this computational simulation, we set $U_0 \in \{-0.8, -0.6\}$, $U_1 \in \{1, 1.2\}$, and $\lambda \in \{0.2, 0.6\}$. The results of this experiment are shown in Figure 6.

More precisely, this figure illustrates that the optimal sensitivity $\delta_1^*$ increases as initial belief $q$ increases. A peer-review's sensitivity $\delta_1^*$ also increases in $U_1$, and $U_0$, where $U_0 < 0 < U_1$. Furthermore, we find that sensitivity is greater than specificity ($\delta_0^* < \delta_1^*$) when the prior belief ($q$) is large enough (see Fig. 6). This effect increases as unit cost of processing a piece of information ($\lambda$) decreases. This is also illustrated in Fig. 6: (top) the unit cost is set at $\lambda = 0.2$; and (bottom) the unit cost is set at $\lambda = 0.6$. Note that the cost of analyzing the information in the reviewers' reports ($\lambda$) increases with the level of noise in the signal received from the reviewers and with the complexity of the manuscript review process.

Therefore, when the editor has a higher initial belief that the manuscript meets the journal standards, they will have a greater motivation to invest significantly more time and effort into analyzing and understanding favorable comments from reviewers than unfavorable ones. In this situation, the computational tool predicts that sensitivity is greater than specificity and

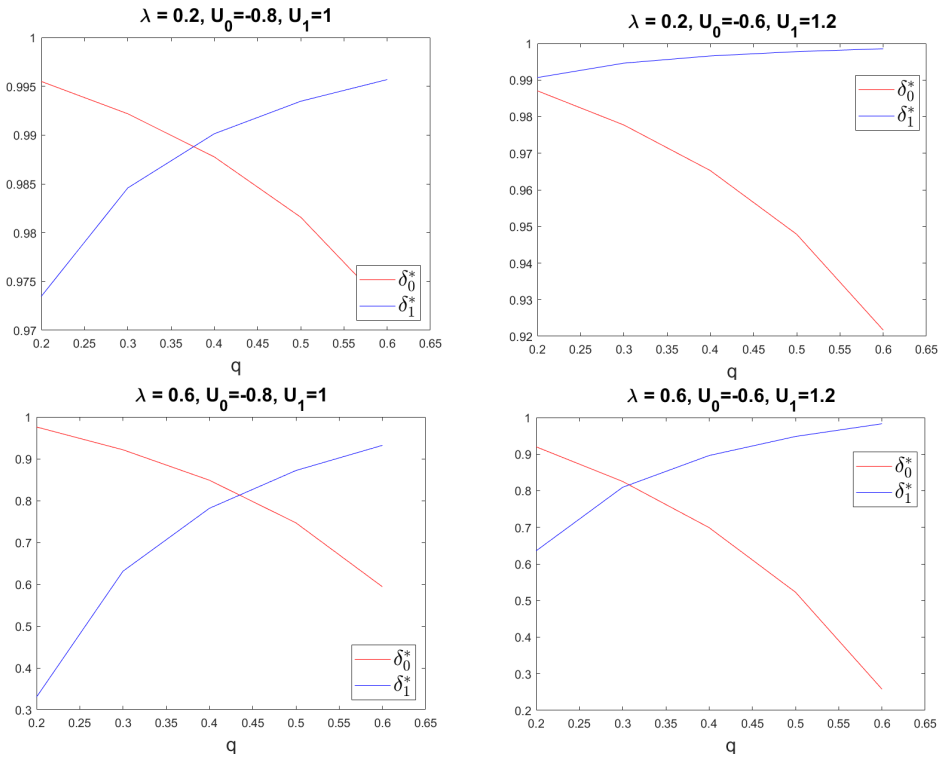



Figure 6: The optimal sensitivity $\delta_1^*$ increases as initial belief $q$ increases. A peer-review's sensitivity also increases in $U_1$, and $U_0$, where $U_0 < 0 < U_1$. On the contrary, a peer-review's specificity $\delta_0^*$ decreases in $q$, $U_1$, and $U_0$.

the optimal sensitivity increases its value. Using these values of sensitivity and specificity, the editor learns that the peer review outcome becomes more accurate when the manuscript meets the journal standards required for publication. It helps the editor understand the importance of improving the analysis performed in review processes involving manuscripts that might not meet the standards required by the journal.

The same is true when $|U_0|$ (i.e., the loss of accepting a manuscript that does not meet the journal standards) becomes relatively smaller as compared to $U_1$ (i.e., the gain of accepting a manuscript that meets the journal standards). As shown in Fig. 6(right), when the loss of accepting a low-quality manuscript decreases and the gain of accepting a high-quality manuscript increases, sensitivity and specificity inform the editor that the review outcome becomes more accurate when the paper meets the standards required for publication and, therefore, they must improve the analysis involved in review processes of low-quality manuscripts. Sensitivity and specificity values thus understood can promote the ethical conduct of peer review processes and improve the validity of editorial decisions for manuscripts that do not meet the standards required for publication by the journal.

Figure 6 also shows that a peer-review's specificity $\delta_0^*$ decreases in $q$, $U_1$, and $U_0$. In addition, we find that specificity is greater than sensitivity ($\delta_0^* > \delta_1^*$) when the editor has a lower initial belief that the manuscript meets

the standards required for publication ($q$), or when the loss of accepting a manuscript that does not meet the journal standards ($|U_0|$) becomes relatively larger as compared to the gain of accepting a manuscript that meets the journal standards ($U_1$). Therefore, under this scenario, the editor has a greater motivation to invest significantly more time and effort into understanding unfavorable comments and less to positive ones. As a result, sensitivity and specificity values inform the editor that a peer review outcome becomes more accurate when the manuscript does not meet the standards required for publication and, therefore, they must improve the analysis involved in review processes of quality manuscripts. In this situation, sensitivity and specificity values can help to improve the validity of editorial decisions for manuscripts that actually meet those standards required for publication.

# 5 Conclusion

Our paper contributes to the literature by being the first to introduce the concepts of sensitivity and specificity in regard to peer review. Sensitivity is the probability of an 'accept' outcome from the peer review process, provided that the manuscript meets the journal standards required for publication. While specificity is the probability of a 'reject' outcome from the peer review process, provided that the work does not meet the standards required for

publication.

In this paper, we have shown that a very high level of sensitivity permits research works to be confidently rejected by the journal editor when the peer review process yields a 'reject' outcome. Similarly, a very high level of specificity permits manuscripts to be confidently accepted by the editor when the review process yields an 'accept' outcome. We have also shown how the journal editor may choose the sensitivity and specificity of a peer review process by allocating more (or less) attention to favorable comments from reviewers than to unfavorable ones. Spending a great deal of time and effort collecting all the information about a manuscript's quality is never optimal, and therefore the information acquired is likely to be incomplete. If the editor then allocates more attention to unfavorable comments than favorable ones, the peer review outcome would be more accurate when the research work does not meet the journal standards required for publication than when it does (i.e., specificity is greater than sensitivity). On the contrary, when the editor pays more attention to favorable comments from reviewers than unfavorable ones, the peer review outcome is relatively more accurate when the manuscript meets the journal standards required for publication than when it does not (i.e., sensitivity is greater than specificity).

In this paper, we have also presented a free online tool that helped us to study the sensitivity and specificity of the peer review process. When

using the computational tool, the editor inserts: (1) their initial belief that the manuscript meets the journal standards required for publication, (2) the utilities that the editor gets from an accept decision when the manuscript does or does not meet the journal standards for publication, and (3) the unit cost of processing a portion of the information in the reviewers' reports.

The computational tool then calculates the sensitivity and the specificity of the peer review process. Using this information, the online tool can predict under what conditions the peer review outcome is relatively more accurate. For peer review processes with a very high sensitivity, the online tool predicts that it is unlikely to reject manuscripts that meet the journal standards. Similarly, it predicts that a very high specificity allows for manuscripts to be confidently accepted by the editor if the peer review process yields an accept signal.

Additionally, we have performed an experiment to study the behavior of both sensitivity and specificity. We found that sensitivity increases as initial belief that the manuscript meets the journal standards increases. A peer-review's sensitivity also increases as utilities that the editor gets from an accept decision increase. We have also found that sensitivity is greater than specificity when the prior belief is large enough. Therefore, in this situation, the journal editor has a greater motivation to invest significantly more time and effort into understanding favorable comments from reviewers

than unfavorable ones.

Similarly, using this experiment, we have found that specificity decreases as initial belief (that the manuscript meets the journal standards) increases and utilities (that the editor gets from an accept decision) increase. When the editor has a lower initial belief, specificity is greater than sensitivity, and thus, the editor has a greater motivation to invest significantly more time into understanding unfavorable comments and less to favorable ones.

Future research could extend our work by examining how the use of cost functions that differ from the mutual information can change the model's results. There are also disadvantages to using Bayesian analyses. This is because it does not tell us how to select an editor's prior belief. Bayesian inference requires skills to translate a subjective prior belief into the baseline probability. The analysis is logically coherent but personal to one editor. Furthermore, it should be noted that the idea of an editor doing a self-analysis of their own decisions can be difficult to put into practice, especially considering the usual position of editors, who tend to be more concerned with issues related to editorial independence or the inappropriate behavior of the authors than with a critical assessment of the decisions they make. Therefore, before experimenting with our approach in practice in a peer-reviewed journal, we need to publish a formal model capable of convincing the scientific community made up of authors, reviewers and journal editors. A different

solution for this operational problem –and one that opens an alternative line of research– could also come from the development of the concept of a "satisficing" sensitivity (specificity) instead of an "optimal" sensitivity (specificity) following the Nobel Prize winning work by Herb Simon.

**Acknowledgments.** We would like to thank the reviewers and editors for their thoughtful comments and efforts towards improving our manuscript.

## References

Abby, M., Massey, M.D., Galandiuk, S., Polk, H.C. Jr. (1994). Peer review is an effective screening process to evaluate medical manuscripts. JAMA, 272(2), 105-107.

Altman, D. G., Bland, J. M. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. BMJ, 308:1552, https://doi.org/10.1136/bmj.308.6943.1552

Bornmann, L. (2008). Scientific peer review: An analysis of the peer review process from the perspective of sociology of science theories. Human Architecture: Journal of the Sociology of Self-Knowledge, 6(2), 23-38.

Bornmann, L. (2011). Scientific peer review. Annual Review of Information Science and Technology, 45(1), 197-245.

Burnham, J. C. (1990). The Evolution of Editorial Peer Review. JAMA, 263(10), 1323-1329.

Chubin, D.E., & Hackett, E.J. (1990). Peerless science: Peer review and U.S. science policy. Stony Brook, NY: State University of New York Press.

Cover, T. M. and Thomas, J. A. (2006). Elements of Information Theory. Wiley, Hoboken, NJ.

Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith HJ., (1987). Publication bias and clinical trials. Control Clin Trials. 8, 343-353.

Dickersin, K., et al., (2002). Association between time interval to publication and statistical significance. JAMA. 287, 2829-2831.

Figg WD, et al., (2006). Scientific collaboration results in higher citation rates of published articles. Pharmacotherapy. 26, 759-767.

Garcia, J.A., Rodriguez-Sanchez, R., and Fdez-Valdivia, J., (2015). The author-editor game. Scientometrics, 104, 361-380, https://doi.org/10.1007/s11192-015-1566-x

Garcia, J.A., Rodriguez-Sanchez, R., Fdez-Valdivia, J. (2019). The optimal amount of information to provide in an academic manuscript. Scientometrics, 121, 1685-1705. https://doi.org/10.1007/s11192-019-03270-1

Garcia, J.A., Rodriguez-Sanchez, R., Fdez-Valdivia, J. (2020). Confirmatory bias in peer review. Scientometrics, 123, 517-533. https://doi.org/10.1007/s11192-020-03357-0

Garfunkel JM, Ulshen MH, Hamrick HJ, Lawson EE., (1994). Effect of institutional prestige on reviewers' recommendations and editorial decisions. JAMA. 272,137-138.

Geraint, H. L., Sheringham, J., Kalim, K., Crayford, T. (2008). Mastering Public Health: A guide to examinations and revalidation. CRC Press, London, p. 616, ISBN 1853157813.

Jerath, K. and Ren, Q., (2021). Consumer Rational (In)Attention Allocation to Favorable and Unfavorable Product Information, and Firm Information Design. Journal of Marketing Research, 58(2), 343-362, https://doi.org/10.1177/0022243720977830

Keiser J, Utzinger J, Tanner M, Singer BH., (2004). Representation of authors and editors from countries with different human development indexes in the leading literature on tropical medicine: survey of current evidence. BMJ. 328, 1229-1232.

Mendis S, Yach D, Bengoa R, Narvaez D, Zhang X., (2003). Research gap in cardiovascular disease in developing countries. Lancet. 361, 2246-2247.

Murad, R., Morgan, O., Mackenzie, K. (2017a). Diagnosis and Screening. Health Knowledge, Public Health Action Support Team (PHAST), https://www.healthknowledge.org.uk/public-health-textbook/disease-causation-diagnostic/2c-diagnosis-screening

Murad, R., Morgan, O., Mackenzie, K. (2017b). Differences between screening and diagnostic tests and case finding. Health Knowledge, Public Health Action Support Team (PHAST), https://www.healthknowledge.org.uk/public-health-textbook/disease-causation-diagnostic/2c-diagnosis-screening/screening-diagnostic-case-finding

Okike K, Kocher MS, Mehlman CT, Heckman JD, Bhandari M., (2008). Nonscientific factors associated with acceptance for publication in The Journal of Bone and Joint Surgery. J Bone Joint Surg Am. 90, 2432-2437.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. Indian journal of ophthalmology, 56(1), 45–50.

Patel V, Sumathipala A., (2001).International representation in psychiatric literature: survey of six leading journals. Br J Psychiatry. 178, 409.

Sims, Christopher A. (2003), Implications of Rational Inattention. Journal of Monetary Economics, 50(3), 665–690.

Sims, Christopher A. (2006), Rational Inattention: Beyond the Linear-Quadratic Case. American Economic Review, 96(2), 158-163.

Stern, J.M., Simes, R.J., (1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ. 315, 640-645.

von Elm, E., Röllin, A., Blümle, A., Huwiler, K., Witschi, M., Egger, M., (2008). Publication and non-publication of clinical trials: longitudinal study of applications submitted to a research ethics committee. Swiss Med Wkly. 138, 197-203.

Yerushalmy, J. (1947). Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques. Public Health Reports (1896-1970), 62(40), 1432–1449. https://doi.org/10.2307/4586294

Yousefi-Nooraie R, Shakiba B, Mortaz-Hejri S., (2006). Country development and manuscript selection bias: a review of published studies. BMC Med Res Methodol. 6, 37.

# Appendix

## A The value of reviewers' positive and negative comments

The value of reviewers' positive and negative comments ($EV$) is given by the expected utility that a journal editor obtains by using the signals received from those comments (i.e., either $S = 1$ or $S = 0$):

$$EV = P(S = 1) \cdot [U_1 P(X = 1|S = 1) + U_0 P(X = 0|S = 1)] + P(S = 0) \cdot 0,$$

and therefore

$$EV = q\delta_1 U_1 + (1 - q)(1 - \delta_0)U_0 \tag{3}$$

by using Bayes' Theorem. This is so because

$$P(S = 1)P(X = 1|S = 1) = P(X = 1)P(S = 1|X = 1) = q\delta_1,$$

and similarly,

$$P(S = 1)P(X = 0|S = 1) = P(X = 0)P(S = 1|X = 0) = (1 - q)(1 - \delta_0).$$

From equation (3), we can see that the value of reviewer's comments increases as the sensitivity and specificity of signal $S$ increase (i.e., the peer review outcome is more accurate) since

$$\frac{\partial EV}{\partial \delta_1} > 0, \text{ and, } \frac{\partial EV}{\partial \delta_0} > 0.$$

# B Optimal sensitivity and specificity

In order to obtain the optimal values of sensitivity $\delta_1$ and specificity $\delta_0$ for a peer review process, the journal editor trades-off the value of reviewers' positive and negative comments $EV$ with the cost of analyzing and understanding those pieces of information $Cost$:

$$\sup_{\delta_0,\delta_1}(EV - Cost)$$

subject to $\delta_0 + \delta_1 > 1$, and where

$$EV = q\delta_1 U_1 + (1-q)(1-\delta_0)U_0$$

and

$$Cost = \lambda I(X, S).$$

Following (Cover and Thomas, 2006), we have that

$$\begin{aligned} I(X,S) &= \left[-\sum_s p(s)\log(p(s))\right] - \left[-\sum_x p(x)\sum_s p(s|x)\log(p(s|x))\right] \\ &= -[q\delta_1 + (1-q)(1-\delta_0)]\log[q\delta_1 + (1-q)(1-\delta_0)] \\ &\quad -[q(1-\delta_1) + (1-q)\delta_0]\log[q(1-\delta_1) + (1-q)\delta_0] \\ &\quad +q[\delta_1\log\delta_1 + (1-\delta_1)\log(1-\delta_1)] \\ &\quad +(1-q)[\delta_0\log\delta_0 + (1-\delta_0)\log(1-\delta_0)]. \end{aligned}$$

Therefore,

- the cost function $Cost = \lambda I(X,S)$ verifies that:

$$\frac{\partial^2 Cost}{\partial\delta_0^2} = \lambda\frac{\partial^2 I(X,S)}{\partial\delta_0^2} > 0, \text{ and, } \frac{\partial^2 Cost}{\partial\delta_1^2} = \lambda\frac{\partial^2 I(X,S)}{\partial\delta_1^2} > 0$$

  and so, it follows that if the editor wants to increase the sensitivity $\delta_1$ or specificity $\delta_0$ of the peer review process, it will become progressively costlier.

- Furthermore,

$$\frac{\partial^2 Cost}{\partial\delta_0\partial\delta_1} = \lambda\frac{\partial^2 I(X,S)}{\partial\delta_0\partial\delta_1} > 0, \text{ and, } \frac{\partial^2 Cost}{\partial\delta_1\partial\delta_0} = \lambda\frac{\partial^2 I(X,S)}{\partial\delta_1\partial\delta_0} > 0$$

  which implies that if the editor processes a larger number of positive

comments from reviewers and the sensitivity $\delta_1$ is higher, the marginal cost requi to analyze and understand the next negative comment also increases, and vice versa.

- There exists a constant $c \geq 1$ such that

$$\text{if } \delta_0 > c\delta_1, \text{ then } \frac{\partial I(X,S)}{\partial \delta_0} > \frac{\partial I(X,S)}{\partial \delta_1},$$

and similarly,

$$\text{if } \delta_1 > c\delta_0, \text{ then } \frac{\partial I(X,S)}{\partial \delta_1} > \frac{\partial I(X,S)}{\partial \delta_0}.$$

Therefore, if the editor has increased the sensitivity of a peer review process $\delta_1$ significantly by processing a large number of positive comments from reviewers, then it is marginally less costly to increase the specificity $\delta_0$ than the sensitivity $\delta_1$ because understanding a negative comment next requires relatively less effort than a positive comment would, and vice versa.

Since the $(EV - Cost)$ function is strictly concave, we obtain the optimal solutions $\delta_0^*$ and $\delta_1^*$ by using the first order condition for optimization. This

condition states that the optimal value of the specificity $\delta_0^*$ verifies

$$0 = \frac{\partial(EV - Cost)}{\partial \delta_0}$$

or equivalently

$$e^{\frac{-U_0}{\lambda}} = \frac{1 - q + q\frac{\delta_1}{1-\delta_0}}{1 - q + q\frac{1-\delta_1}{\delta_0}}.$$

Similarly, using again the first order condition, the optimal value of the sensitivity $\delta_1^*$ verifies

$$0 = \frac{\partial(EV - Cost)}{\partial \delta_1}$$

or equivalently

$$e^{\frac{U_1}{\lambda}} = \frac{q + (1-q)\frac{\delta_0}{1-\delta_1}}{q + (1-q)\frac{1-\delta_0}{\delta_1}}.$$

Solving the equations we get

$$\delta_0^* = \frac{1}{1 - e^{-\frac{U_1 - U_0}{\lambda}}} \left[ 1 - \frac{q}{1-q} \frac{e^{\frac{U_1 - U_0}{\lambda}} - e^{\frac{-U_0}{\lambda}}}{e^{\frac{U_1 - U_0}{\lambda}} \left( e^{\frac{-U_0}{\lambda}} - 1 \right)} \right]$$

and

$$\delta_1^* = \frac{1}{1 - e^{-\frac{U_1 - U_0}{\lambda}}} \left[ 1 - \frac{1-q}{q} \frac{e^{\frac{-U_0}{\lambda}} - 1}{e^{\frac{U_1 - U_0}{\lambda}} - e^{\frac{-U_0}{\lambda}}} \right]$$

Then, considering the constraints $\delta_0^* + \delta_1^* > 1$, $0 < \delta_1^* < 1$, and $0 < \delta_0^* < 1$,

it follows that

$$\frac{e^{\frac{-U_0}{\lambda}}-1}{e^{\frac{U_1-U_0}{\lambda}}-1} < q < \frac{1-e^{\frac{U_0}{\lambda}}}{1-e^{-\frac{U_1-U_0}{\lambda}}}$$

and

$$\begin{aligned}\Pr(S=1) &= q\delta_1^* + (1-q)(1-\delta_0^*) \\ &= \frac{q}{1-e^{-\frac{-U_0}{\lambda}}} - \frac{1-q}{e^{\frac{U_1}{\lambda}}-1} \\ &= \frac{q}{1-e^{-k}} + \frac{1-q}{1-e^{l-k}}\end{aligned}$$

where $k = \frac{-U_0}{\lambda}$, $l = \frac{U_1-U_0}{\lambda}$.

Therefore, we obtain that

$$\delta_1^* = \frac{1-\frac{1-q}{q}\frac{e^k-1}{e^l-e^k}}{1-e^{-l}}$$

and

$$\delta_0^* = \frac{1-\frac{q}{1-q}\frac{e^l-e^k}{e^l(e^k-1)}}{1-e^{-l}}$$

and the journal editor accepts the manuscript upon receiving the peer review outcome $S = 1$ with probability

$$\Pr(S=1) = \frac{q}{1-e^{-k}} + \frac{1-q}{1-e^{l-k}}.$$

However, when the editor's prior belief that the manuscript meets the journal

standards required for publication is either very low, i.e., $q \leq \frac{1-e^{k}}{1-e^{l}}$, or very high, i.e., $q \geq \frac{1-e^{-k}}{1-e^{-l}}$, the editor makes a desk decision without external review.

## 3.4. The Relevance of Title, Abstract, and Keywords for scientific paper quality and potential impact

### 3.4.1. Datos generales

1. **Autores**: Jorge Chamorro-Padial y Rosa Rodríguez-Sánchez.
2. **Revista**: Multimedia Tools and Applications.
3. **Datos sobre la publicación**:
   - **Referencia**: Chamorro-Padial y Rodríguez-Sánchez (2023b).
   - **Año**: 2023.
   - **Editorial**: Springer.
   - **DOI**: https://doi.org/10.1007/s11042-023-14451-9.
4. **Estado**: Publicado.
5. **Métricas**:
   - **Factor de impacto**: 2,577 (JCR, 2021).
   - **Ranking**:
     - *Social Science Citation Index Expanded (SSCIE)* [2]:
       - *Computer Science, Software Engineering*: Q2 - 48/110 (año 2021).
       - *Computer Science, Information Systems*: Q3 - 107/164 (año 2021).
       - *Computer Science, Theory & Methods*: Q2 - 42/110 (año 2021).
       - *Engineering, Electrica & Electronic*: Q3 - 144/276 (año 2021).
       - *Computer Science, software, graphics, programming*: NA (año 2021).

### 3.4.2. Contribuciones principales

1. Mediante el modelo cuasi especies, obtenemos el mejor perfil de sumisión de un autor de manuscritos científicos. El mejor perfil es aquel que ofrece el mayor beneficio que resulta entre los perfiles definidos por la capacidad de distinción del autor entre artículos de bajo, medio y alto impacto.

[2] A fecha de depósito de esta tesis, aún no se disponen de datos del año 2022.

Tabla 3.1: Particiones y perfiles de envío.

| **Partición** | **Número de perfiles de envío** |
|---|---|
| $K_F = \{\{S_1\}, \{S_2\}\{S_3\}\}$ | 27 |
| $K_{F_1} = \{\{S_1\}, \{S_2, S_3\}\}$ | 9 |
| $K_{F_2} = \{\{S_2\}, \{S_1, S_3\}\}$ | 9 |
| $K_{F_3} = \{\{S_3\}, \{S_1, S_2\}\}$ | 9 |
| $K_C = \{S_1, S_2, S_3\}$ | 3 |

2. Desarrollo de una aplicación web que permite evaluar la cantidad de información mínima y tiempo necesario para identificar la calidad de un manuscrito.

3. El título, abstract y palabras clave de un artículo han sido suficientes para que una muestra de personas cualificadas puedan, en aproximadamente 30 segundos, tomar una decisión acertada sobre la calidad de un trabajo.

4. Proporcionar a los autores información sobre su partición y la estrategia más óptima de envío de trabajos a revistas.

### 3.4.3. Resumen

En este trabajo, continuamos con la aplicación del modelo de cuasi especie que ya habíamos utilizado en otros manuscritos previamente comentados. Esta vez, nos centramos en la perspectiva de los autores. Dentro del proceso de publicación de artículos académicos, los autores deben enviar artículos a revistas académicas. Para una misma área de conocimiento, es típico que exista varias revistas, y que cada una de ellas tenga un nivel determinado de impacto u otras métricas. De esta forma, existirán revistas de mayor prestigio que otras.

Las revistas de mayor prestigio, por norma general, tienden a aceptar artículos de muy alta calidad y siguen un proceso de revisión por pares más estricto que otras revistas de menor prestigio. A su vez, los autores crean manuscritos que tienen diferentes niveles de calidad. Si el proceso de revisión por pares fuese perfecto, los artículos de mayor calidad serían publicados en las revistas con más prestigio y los de menor calidad, en revistas de menor reputación.

Pero los autores tienen diferentes niveles de percepción de la calidad de un artículo, pudiendo sobrestimar o subestimar la calidad de un manuscrito o, por ejemplo, no siendo capaz de diferenciar entre artículos de diferentes calidades.

Definimos la partición de un autor como su capacidad para distinguir la calidad de un trabajo académico. Un autor puede ser capaz de diferenciar entre artículos de calidad baja, media y alta, en cuyo caso tendrá una partición fina $K_F$, o puede ser incapaz de diferenciar la calidad de ningún artículo, teniendo por tanto una partición gruesa, $K_C$. Entre $K_F$ y $K_C$ existe todo un abanico de posibilidades, como se muestra en la Tabla 3.1. En esta misma tabla, podemos ver el número de perfiles de envío de cada partición. El perfil de envío es la respuesta del autor tras haber juzgado la calidad de un artículo. Por ejemplo: un autor puede ser incapaz de identificar la calidad de un manuscrito (partición gruesa, $K_C$) y optar por someter siempre todos los artículos en revistas de un impacto medio. En este caso, el perfil de envío de este autor se identificaría como **(M-I)** y es uno de los tres perfiles posibles para la partición gruesa (los otros dos serían **(L-I)** para bajo impacto, y **(H-I)** para alto impacto ).

Con la decisión de enviar un artículo a una revista, el autor consigue una recompensa. Si bien en otros artículos, de índole más teórica, entramos de lleno en esta cuestión, analizando la mejor estrategia para cada partición. En este artículo el objetivo era analizar la cantidad de información y el tiempo que un autor necesita para tomar una decisión sobre la calidad de un artículo. Para ello, hemos lanzado una aplicación Web [3] y un experimento donde los participantes tenían que leer el título, el abstract y las palabras clave de un artículo, y decidir la calidad del mismo. Este conjunto de información no es aleatorio, sino que es el que, habitualmente, está disponible para los autores en los repositorios académicos, bases de datos y portales de revistas académicas dentro del área de las Ciencias de la Computación. El tiempo promedio empleado por artículo fue de 29,47 segundos, mientras que la partición más frecuente fue $K_F$. Se debe tener en cuenta que la población que formó parte de este experimento estaba formada, mayoritariamente, por investigadores o bien personas con un nivel de formación equivalente a máster y experiencia laboral en el sector TIC.

Con los resultados experimentales en la mano, ahora sí, se puede calcular la mejor estrategia de envío para cada autor, de acuerdo con su partición. La aplicación web desarrollada, además de su parte meramente experimental, indica al autor cuál su partición y cuál es su perfil de envío (y con qué perfil de envío obtendría una mejor puntuación).

[3] En el Capítulo 6 analizamos más en detalle las herramientas y programas que han sido generados a raíz de esta tesis.

# The Relevance of Title, Abstract, and Keywords for scientific paper quality and potential impact

Jorge Chamorro-Padial * [1] [2]

Rosa Rodríguez-Sánchez [3]

**Keywords: Peer-review process; Informed authors; Authors skills; Quasi-species; Systematic review; Recommendation System**

**Abstract**

Authors, editors, and reviewers need to have a good perception regarding the quality of a manuscript in order to improve their skills, save effort, and prevent errors that can affect the submission procedure. In this paper, we compared the author's perception of a manuscript's quality with the manuscript's actual impact. In addition, we analyzed the uncertainty of the author's perception of the manuscript's quality. From there, we defined 'partition' as the author's ability to perceive the actual quality. We did this by launching a website for the use of the scientific community. This webpage provided a tool to help improve an investigator's skill in understanding and recognizing the quality of a manuscript so as to help researchers improve and maximize their works' potential

[1] CITIC-UGR. Universidad de Granada. 18071 Granada, Spain. ORCID: 0000-0002-6334-3786

[2] Corresponding author: Jorge Chamorro Padial. jorgechp@correo.ugr.es

[3] Departamento de Ciencias de la Computación e I.A. CITIC-UGR. Universidad de Granada, 18071 Granada, Spain. ORCID: 0000-0001-7886-9329

impact. We carried out the experiment with 106 experienced users who tested our webpage. We found that the Abstract, the Title, and the Keywords were enough to perform a substantially decent evaluation of a manuscript. Most of the researchers were able to determine the quality of a paper in less than a minute from this small amount of information.

**1 Introduction**

In the academic field, authors need to publish their results in order to make them available to the public, and manuscripts are one of the most important ways to achieve those needs. We can simplify the manuscript publishing process into the following steps: 1) Authors write a manuscript and send it to a journal, 2) The journal reviews the manuscript to check if their quality standards are met, and 3) After the review process, the manuscript is either published or rejected. We can consider manuscripts as potential papers once they pass a review process and various corrections or modifications are made (if required).

Usually, there are plenty of candidate journals where a manuscript could suitably fit according to the topic(s) and the manuscript's field of knowledge. This situation forces authors to select a journal where they would prefer to have their work published. Typically, every candidate journal has a different level of impact. The impact is defined by [1] as *one of academia's strongest currencies*. For a journal, impact is an important

asset which ensures that the scientific community, libraries, and academic researchers all continue to pay attention to its publications. Low impact journals face the risk of being removed from scientific indexes and losing interest from the community. For authors, the impact directly affects their visibility and prestige, which can, in turn, directly affect their research career.

Articles published in high-impact journals will presumably have a more significant impact than articles published in low-impact journals. Quality is another critical asset along with impact. We consider a manuscript's quality in terms of originality, importance, soundness of theory, and verified conclusions. From the point of view of a journal, high-quality articles have a greater probability of attracting the scientific community's interest. Thus, journals tend to define measures in order to select articles of the highest possible quality [7]. When a journal receives an article, the manuscript is often initially checked to determine if a minimum quality is met. If not, the article would be promptly rejected as a desk decision. If the article fits the journal's minimum quality standards then it is usually sent to the next step, a peer-review [11].

In this paper, we want to offer support to the peer-review process by providing a tool that can be useful for training and improving the skills of authors, reviewers, and editors while imparting valuable knowledge:

- For authors: We propose that our tool can help authors recognize their ability when distinguishing the quality of a manuscript. This knowledge can be useful

when choosing a journal for their manuscripts. Also, authors tend to suffer from confirmation bias, which leads to overconfidence as authors believe their manuscripts are of a higher quality regardless of what the actual, objective quality is [18]. The proposed tool can help authors correct this bias by comparing their choices about the manuscript's quality (Low, Medium, High quality) with its actual quality.

- For reviewers: We want to give them feedback about their ability to identify an article's quality. For example, a reviewer who incorrectly matches the quality of a manuscript and the quality standard of a journal would incur a cost to the journal [6].
- For editors: For editors, knowing their skill-level when properly identifying the quality of a manuscript can be also useful. For example, a desk decision can save effort and time for reviewers and authors if the editor believes that the manuscript's quality does not match the journal's quality standards [8,17].

In order to do all of this we built a web training system[1] where users can review information about different real papers and decide what their quality is while receiving direct feedback about their decisions. On our website, authors are presented with only the key elements of an article (abstract, title, and keywords). During the training process,

[1] https://blackcat.ugr.es/quasispecies/

their time per response is measured. From this we wanted to answer the following questions:

1. What is the level of uncertainty that authors have about the quality of a manuscript with respect to its actual quality?
2. Do the Title, the Abstract, and the Keywords contain enough information to determine the quality of a manuscript?
3. What is the average time that an author spends reviewing the key information of a manuscript?

We assumed that an article's actual quality matched the impact category (Low, Medium, or High) of the journal where the manuscript was published.

Our work is structured in the following manner:

- **State of the Art and related works**: Explores the current state of the art information and research in this field in terms of scientific literature.
- **Model:** Describes our model.
- **Methodology:** Explains the methodology of our experiment, the website, and the dataset used.
- **Results:** Reports the results obtained from users who used our website.
- **Discussion**: In this section, we perform a more detailed analysis of the results achieved.

- **Conclusions:** The last section of our work presents the critical information from our paper.

## 2 State of the Art and related works

Peer review is a standard quality control procedure that is part of a consensus-seeking scientific discussion on quality assurance [13]. During the peer review stage, an editor selects reviewers who are expected to read the candidate manuscript and give a critical assessment regarding the work's quality. Peer review acts as an editor's source of knowledge to help them decide if a manuscript should be accepted, rejected, or returned to the author(s) with corrections [3]. Manuscripts relevant to the journal's scope, which are innovative and well written, have a higher probability of being accepted [6].

In this context, authors will want to send their manuscript to as high an impact journal as possible, while journals will want to publish articles of the highest quality possible. From the perspective of an author, if he or she sends a high-quality manuscript to a low-impact journal there will be a substantial cost in terms of lack of visibility. It is important to mention that journal impact factor is a strong predictor of the number of citations [2]. Nevertheless, sending a low-quality manuscript to a high-impact journal increases the risk of rejection, which would affect the time and effort spent trying to publish, as well as

the author's motivation. Writing a manuscript for a high-impact peer-reviewed journal can be a challenging and frustrating experience. For example, [22] concludes that “the authors do few manuscript submissions prior to journal acceptance, most commonly by lower impact factor journals”.

In [18] the authors analyzed the evolutionary game derived from journal quality controls. An author produces low or high-quality manuscripts which are then submitted to journals who accept manuscripts of different qualities with a certain probability. The authors also identified different strategies and their survival chances according to evolutionary games. These strategies are based on the concept of authors' and editors' quality profiles. An author's profile is based on the probability of an author submitting articles of a certain quality (low or high). In contrast, an editor's profile is based on the frequency with which an editor accepts articles with a specific quality (low or high).

A vast majority of authors still feel the need to enhance their skills in popular science writing [16]. Nowadays, the author can use different tools that can help in the process of writing a manuscript. Among these tools, we can distinguish Jasper[1] and Hemingway Editor[2]. Jasper uses AI to help write different parts of the manuscript. For its part, Hemingway helps the author highlight problems with their writing. Its goal is to make

[1] https://www.jasper.ai

[2] https://hemingwayapp.com/

complex sentences easier. However, while those tools help to write a manuscript, the author must have the ability to recognize the quality of the manuscript.

Different works focus on the author’s perception of the manuscript’s quality. This topic is important to analyze since according to [15] absolute impact factor of the journal, match between perceived "quality" of their study, and journal impact factor were considered to be the three most important factors by the authors when they have to submit a manuscript.

In [19] concrete suggestions for improving the perception of a paper in the reader's minds is presented. Also, [23] proposed a pilot study to evaluate a method of teaching neurology residents the basic concepts of biostatistics, research methodology, and review of scholarly literature by employing a program of peer-reviewed scientific manuscripts.

Selecting a journal is not always without problems, as authors can suffer from having a flawed perception about their article's quality. Additionally, reviewers can have imperfect knowledge or bias when determining the quality of a reviewed work. If authors can distinguish the actual quality or impact of a manuscript, they have a fine partition. Conversely, if they cannot distinguish the actual quality of an article, they have a coarse partition. Here, a partition is defined as a map between the author's perception of the manuscript's quality and the actual impact of the manuscript. If the author's perception

coincides with reality, we can say that they have a fine partition. The actual impact of a manuscript can be measured by the impact of the journal where it was published. We also used the author's profile as one of the possible indicators of the quality of a manuscript given the author's partition. For example, suppose an author cannot distinguish between a low, medium, and high impact manuscript (they have a coarse partition) when the author has to evaluate a manuscript's impact. In that case, he or she would have three profiles: low, medium, or high impact. On the contrary, if an author has a fine partition, they could have 27 possible profiles. The perception of the quality of a manuscript depends on the author's partition and the distribution of articles over three different categories (High impact, Medium impact, and Low impact).

The same concept is applied to reviewers [4]. The quasi-species model inspires our work to determine the evolution of an authors' profiles after the peer-review process. This model was intended to represent the Darwinian evolution of self-replicating entities when a high mutation rate occurs [12, 20]. According to this model, a quasi-species is a big group, or *cloud*, of genotypes in an environment where their descendants will have a high probability of mutation. The evolutionary success of a quasi-species strongly depends on the replication rates of clouds. In [4] the authors adapted the quasi-species model from biology to the author-editor game's evolutionary environment. Self-replicating entities are submission profiles under a given partition of manuscript categories. Errors produce profile mutations, and only submission profiles with high replication rates survive.

Peer review is not exempt from criticism and deficiencies [10–11], but nowadays it is one of the scientific community's essential tools to validate and improve the quality of science. Every year, about 13.7 million reviews are done in the academic ecosystem for a total of 3 million scientific articles [9, 21]. Additionally, peer-review is an indicator of prestige and confidence for journals and authors [2, 14]. Ultimately, we can say that peer-review is a crucial element of the science of today, and it is necessary to continue to improve it by raising the skill of all actors involved in the process.

**3 Model**

As described in the introduction, an author submits a manuscript that can have different levels of quality. In this paper, we define three different manuscript categories: $S=\{s_1, s_2, s_3\}$, with $s_1$ being a low-quality manuscript, $s_2$ a medium-quality manuscript, and $s_3$ a high-quality manuscript. Likewise, we define three different journal impacts, $I=\{Low-impact, Medium-impact, High-impact\}$. The action of sending an article to a journal can be seen as optimal or non-optimal. For example, if an author sends a low-quality article to a high-impact journal, it is very likely to get a rejection. In that case, the author has lost time and effort, so it is considered a non-optimal action. Additionally, if an author sends a high-quality article to a low-impact journal, the author is paying the

price in terms of visibility, prestige, and impact, which is also a non-optimal action. For $s_j \in S$ we can define an optimal action as follows:

$$i^{*}(s_1) = Low - impact$$

With $i^{*}(s_1)$ being the optimal action of sending a low-quality article to a low-impact journal.

$$i^{*}(s_2) = Medium - impact$$

With $i^{*}(s_2)$ being the optimal action of sending a medium-quality article to a medium-impact journal.

$$i^{*}(s_3) = High - Impact$$

With $i^{*}(s_3)$ being the optimal action of sending a high-quality article to a high-impact journal.

Every action gives a score to the author. In our model, the result for non-optimal actions is 0, while the optimal action score is 1. We define the reward function, $\pi_i(s_j)$ for $i \in I$ and $j \in S$, as follows:

$$\pi_i(s_j) = \begin{Bmatrix} 1 & i = i^{*}(s_j) \\ 0 & i \neq i^{*}(s_j) \end{Bmatrix}$$

Eq 1

Every author has a different ability to identify the quality of an article. We formally represent the distinctive capabilities of authors as *partitions*. Every author uses a particular partition. If an author can distinguish between low, medium, and high-quality articles, then the author has a fine partition, $K_F=\{\{s_1\},\{s_2\},\{s_3\}\}$. If an author does not distinguish between any type of quality, then the author uses a coarse partition $K_C=\{s_1,s_2,s_3\}$. Among these polarized partitions, we can also identify other ones:

- $K_{F_1}=\{\{s_1\},\{s_2,s_3\}\}$: The author can identify low-quality articles but cannot identify medium and high-quality articles.
- $K_{F_2}=\{\{s_2\},\{s_1,s_3\}\}$: The author can identify medium-quality articles but cannot identify between low and high-quality articles.
- $K_{F_3}=\{\{s_3\},\{s_1,s_2\}\}$: The author can identify high-quality articles but cannot identify low and medium-quality articles.

Using a partition is the basic knowledge that an author has to decide what the potential (impact) of a manuscript would be. Knowledge is also gained from good and bad experiences when submitting manuscripts to different journals. This additional knowledge allows them to have informed opinions about where to submit an article with a certain level of quality. This extra information makes up part of the *submission profile* of an author. For every category in an author partition, there is a corresponding submission pattern. For example, an author who uses a fine partition has a submission profile consisting of three different submission patterns. An example of a submission profile for a fine partition is $(L-I,M-I,H-I)$ where an author can identify low, medium,

and high-quality articles. However, low and medium-quality manuscripts are sent to low-impact journals, while high-impact manuscripts are sent to high-impact journals. Table 1 summarizes the number of submission profiles per partition, Table 2 and Table 3 describe the submission profiles for partitions $K_F=\{\{s_1\},\{s_2\},\{s_3\}\}$ and $K_{F_1}=\{\{s_1\},\{s_2,s_3\}\}$, respectively.

**Table 1** Number of submission profiles per partition

| Partition | Number of submission profiles |
|---|---|
| $K_F=\{\{s_1\},\{s_2\},\{s_3\}\}$ | 27 |
| $K_{F_1}=\{\{s_1\},\{s_2,s_3\}\}$ | 9 |
| $K_{F_2}=\{\{s_2\},\{s_1,s_3\}\}$ | 9 |
| $K_{F_3}=\{\{s_3\},\{s_1,s_2\}\}$ | 9 |
| $K_C=\{s_1,s_2,s_3\}$ | 3 |

**Table 2** Submission profiles for the fine partition, $K_F=\{\{s_1\},\{s_2\},\{s_3\}\}$.

| | | |
|---|---|---|
| (L-I, L-I, L-I) | (M-I, L-I, L-I) | (H-I, L-I, L-I) |
| (L-I, L-I, M-I) | (M-I, L-I, M-I) | (H-I, L-I, M-I) |
| (L-I, L-I, H-I) | (M-I, L-I, H-I) | (H-I, L-I, H-I) |

| | | |
|---|---|---|
| (L-I, M-I, L-I) | (M-I, M-I, L-I) | (H-I, M-I, L-I) |
| (L-I, M-I, M-I) | (M-I, M-I, M-I) | (H-I, M-I, M-I) |
| (L-I, M-I, H-I) | (M-I, M-I, H-I) | (H-I, M-I, H-I) |
| (L-I, H-I, L-I) | (M-I, H-I, L-I) | (H-I, H-I, L-I) |
| (L-I, H-I, M-I) | (M-I, H-I, M-I) | (H-I, H-I, M-I) |
| (L-I, H-I, H-I) | (M-I, H-I, H-I) | (H-I, H-I, H-I) |

**Table 3** Submission profiles for $K_{F_1}=\{\{s_1\},\{s_2,s_3\}\}$, $K_{F_2}=\{\{s_2\},\{s_1,s_3\}\}$ and $K_{F_3}=\{\{s_3\},\{s_1,s_2\}\}$.

| | | |
|---|---|---|
| (L-I, L-I) | (M-I, L-I) | (H-I, L-I) |
| (L-I, M-I) | (M-I, M-I) | (H-I, M-I) |
| (L-I, H-I) | (M-I, H-I) | (H-I, H-I) |

This paper would also like to apply certain concepts inspired by the quasi-species model [4, 12].

For each author, we compute the probability of each partition as follows:

$$P(K_F)=P(s_3\vee i_3)\cdot P(i_3)+P(s_2\vee i_2)\cdot P(i_2)+P(s_1\vee i_1)\cdot P(i_1)$$

*Eq 2*

$$P(K_C)=\left(P(s_3\vee i_1)+P(s_2\vee i_1)\right)\cdot P(i_1)+\left(P(s_1\vee i_2)+P(s_3\vee i_2)\right)\cdot P(i_2)+\left(P(s_1\vee i_3)\right.$$

*Eq 3*

$$P(\{\{s_a\},\{s_b,s_c\}\}=P(s_a|i_a)\cdot P(i_a)+P(s_c|i_b)\cdot P(i_b)+P(s_b|i_c)\cdot P(i_c)$$

*Eq 4*

Where $a, b,$ and $c$ are elements of the set $[LOW, MEDIUM, HIGH]$, $s \in S$ and $i \in I$. Remember that $I$ is the set of categories for the actual impact of a manuscript, and $S$ is the set of categories for an author's perception the manuscript's quality.

Finally, the most probable partition is assigned to the author. Once the partition is established, we can compute each submission profile score for the selected partition by considering the frequency with which a set of authors produce manuscripts of each category. For example, for the $K_F$ partition, the best submission profile will always be (L-I, M-I, H-I), giving the best possible score for an author. For $K_{F_1}=\left(\left[s_1\right],\left[s_2,s_3\right]\right]$, it is necessary to decide between the submission profiles (L-I, M-I) or (L-I, H-I) according to the occurrence frequency for medium and high impact manuscripts.

Let $\pi_{(i)}(K)$ be the reward of submission profile (i) under partition K given as:

$$\pi_{(i)}(K)=\sum_{s\in S} f_s \pi_i(s) \quad \text{Eq 4}$$

With $S=\{s_1,s_2,s_3\}$ being the set of manuscript categories; $f_s$ being the frequency of a manuscript category; $\pi_i(s)$ being the reward function under submission profile (i) for manuscript category $s$, as defined in Eq 1. Then, we denote

$$\pi(K)=\left(\pi_{(1)}(K),\pi_{(2)}(K),\ldots,\pi_{(n)}(K)\right) \quad \textit{Eq 5}$$

as the reward vector of submission profiles under partition $K$. Among all the author's profiles, we define the best profile as the one with the highest score, seen as:

$$bestprofile(K)=argmax_{i\in P}\left\{\pi_{(i)}(K)\right\} \quad \textit{Eq 6}$$

with $P$ being the set of possible profiles for the partition $K$.

## 4 Methodology

### 4.1 Experimental setup

To apply our model, we deployed our website[1] with the aim of improving authors' skills when identifying the quality of manuscripts and letting them know their most probable partition as well as their recommended submission profile, according to the responses provided.

[1] https://blackcat.ugr.es/quasispecies/

The website was built using a combination of Typescript, HTML, and CSS using Angular framework. Our website is connected to a server written in Python by using a RESTful API. The webpage is responsive, so participants can use the webpage using either a computer or a smartphone. Figure 1 and Figure 2 show screenshots from the website. The supplemental material of this work contains screenshots of each section on the website, together with an explanation for each one.

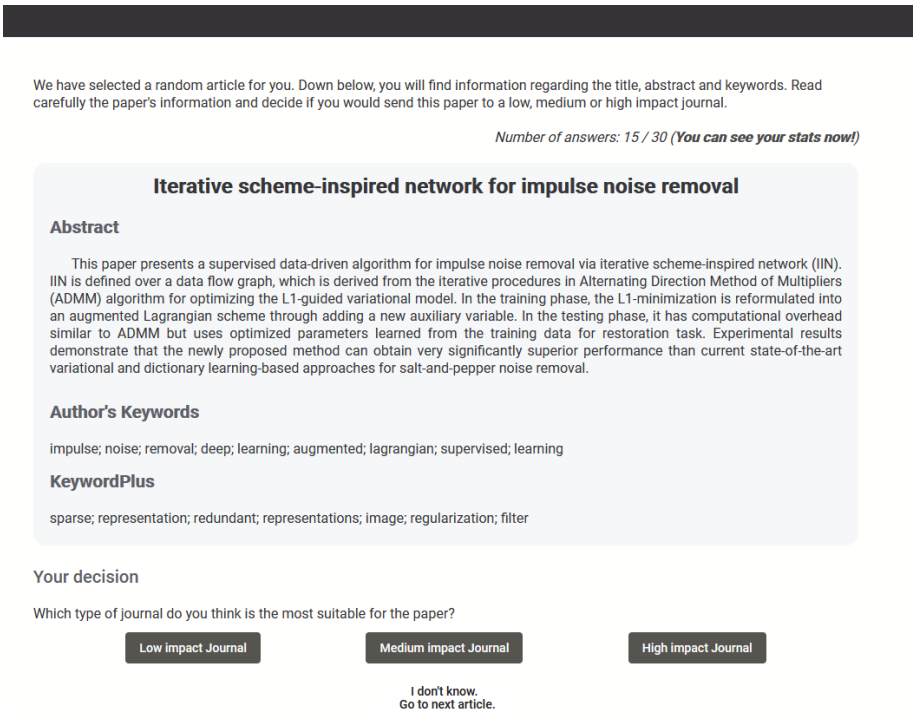
4.2 Participants

To test our proposed model, we asked 106 participants to register on our website and classify a minimum of 15 random articles.

We needed our participants to have experience in reading and working with scientific literature, so we asked them to have, at least, a bachelor's degree. In addition, our dataset consisted of computer science articles, so working in or having experience in an IT related area was another requirement. Individuals in our experiment came from two different sources:

- 86 participants worked in IT jobs.
- 20 participants were authors from Computer Sciences journals.

***Figure 1**: Quasi-species Peer review website. An example of a training session where the author has to decide which type of journal best fits with the displayed article.*

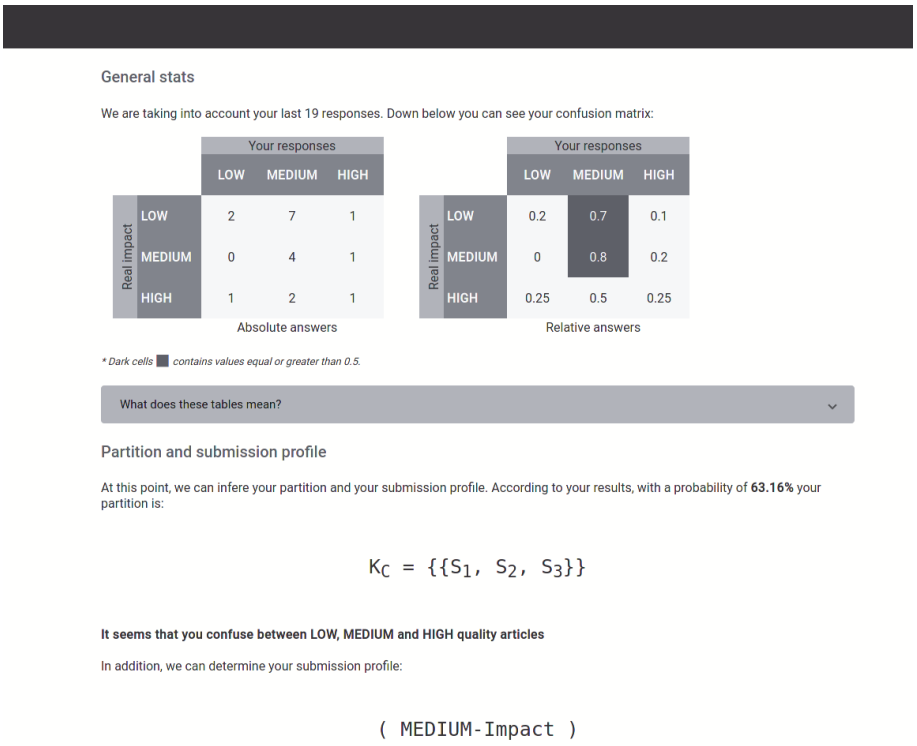



## 4.3 Materials

We created a dataset of articles published in JCR, specifically indexed journals from the *Computer Science* category, sub-area *Artificial Intelligence,* from 2019. The dataset contained 21,799 articles. For each article, the dataset included information regarding the title, keywords, abstract, and publishing journal. Concerning the journals, the dataset contained information about the journal's title, impact factor, and tertile level.

The dataset is published in Kaggle [4].

***Figure 2:*** *The quasi-species peer review website. An example of a stats session. The author can check their results in the form of a confusion matrix. Additionally, they can see their partition and the recommended submission profile.*




## 4.4 Design

In our experiment, we wanted to identify the partition and the submission profile of an author. For that purpose, we used the reward obtained by an author when sending a manuscript to a journal, $\pi_i(s_j)$. This reward is inferred from the user responses on our website. The user must decide by reading only limited information about the article (title, abstract, and keywords).

When users are in the training section, the articles they have to review are selected randomly. Users have to infer the article's quality and then decide which impact level journal they would submit the manuscript to. In order to establish a better correlation between qualities (low, medium, and high quality) and impact, we have assigned three different impacts to journals in our dataset (high, medium, and low impact). According to the JCR index, this impact is in line with the journal's impact factor during the year 2019. In this sense, journals in the first tertile are considered high impact journals, journals in the second tertile are considered medium-impact journals and journals in the third tertile are defined as low impact ones.

Concerning the quality of articles from the dataset, articles from high impact journals are considered high-quality manuscripts. Articles from a medium impact journal are considered medium-quality manuscripts, and, finally, articles from a low impact journal are considered low-quality manuscripts.

4.5 Procedure

The user experience on the website is as follows:

1. The user is signed up to the System.
2. After the signup process, users enter the *Training section*, where the papers are displayed, and a submitting decision must be made (see Figure 1). The user can skip the manuscript if they are unable to make a decision.

3. After submitting a minimum of 15 articles the user can access the *Stats section,* see *Figure 2*. In this section, they can see their partition type and recommended submission profile.
4. The user can go back to the Training section and keep training if they wish.

There is detailed information about the website's interface in the supplemental document.

From the user responses, we computed two confusion matrices for each user, $MA$ and $MR$. These matrices contain the same information about the users' responses but contain, respectively, absolute and relative results. $MA$ is only used to provide additional information to the user in the Stats section, while $MR$ is used to compute the partition type and the submission profile following the model described in the Model section.
. and Table 5 are examples of $MA$ and $MR$. Although both matrices contain the same information, we use $MR$ to determine the author partition. The first step is to calculate the probabilities of each type of partition:

$$P(K_{F_1})=0.421\ P(K_{F_2})=0.316\ P(K_{F_3})=0.263\ P(K_C)=0.632\ P(K_F)=0.368$$

With this information, the partition with a higher score is $K_C$, which determines the author's partition type. The second step is to compute the frequencies for each category

of manuscripts. $f_s$: $f_1$=0.158, $f_2$=0.684, $f_3$=0.158. These frequencies help us weigh the maximum possible score by considering the author's behavior. The third step is to compute the score for each possible submission profile. For $K_C$, available submission profiles are described in Table 3:

$(L-I)$=0.158 $(M-I)$=0.684 $(H-I)$=0.158

Finally, we selected the best profile, which is $(M-I)$. This profile is the one that the author should follow in order to increase their score. While the experiment was taking place, the webpage was measuring the response time for each article. Table 6 illustrates an example of a user with a Fine Partition, with their partition probabilities as follows: $P(K_{F_1})$=0.267 $P(K_{F_2})$=0.366 $P(K_{F_3})$=0.366 $P(K_C)$=0.105 $P(K_F)$=0.890.

To compute the submission profile, we used the same category frequencies as the example in Table 4, $f_s$: $f_1$=0.158, $f_2$=0.684, $f_3$=0.158. For $K_F$, we have 27 different submission profiles, with (L-I, M-I, H-I) being the most probable, with a score of 1.0.

**Table 4** Example of a $MA$ matrix. This matrix is used to give users additional information on the website.

| | | *Perceived quality* | | |
|---|---|---|---|---|
| | | **LOW** | **MEDIUM** | **HIGH** |
| *Real journal impact* | **LOW** | 2 | 7 | 1 |
| | **MEDIUM** | 0 | 4 | 1 |
| | **HIGH** | 1 | 2 | 1 |

**Table 5** Example of an $MR$ matrix. This matrix is used to compute the author's partition and their corresponding submission profiles. MR provides information about real journal impact and perceived quality in relative terms while MA contains absolute information.

| | | *Perceived quality* | | |
|---|---|---|---|---|
| | | **LOW** | **MEDIUM** | **HIGH** |
| *Real journal impact* | **LOW** | 0.105 | 0.368 | 0.053 |
| | **MEDIUM** | 0 | 0.211 | 0.053 |
| | **HIGH** | 0.053 | 0.105 | 0.053 |

**Table 6** Example of an $MR$ matrix of a user with a Fine Partition.

| | | *Perceived quality* | | |
|---|---|---|---|---|
| | | **LOW** | **MEDIUM** | **HIGH** |
| *Real journal impact* | **LOW** | 0.25 | 0 | 0.05 |
| | **MEDIUM** | 0.05 | 0.3 | 0.0025 |
| | **HIGH** | 0 | 0.0025 | 0.3 |

## 5 Results

As stated in the Participants section, 106 participants used our webpage and simulated the submission of at least 15 articles according to the articles' perceived quality.

Once all participants had finished their task we extracted the different partitions and submission profiles obtained from them. The results are shown in Table 7. Regarding

the submission profiles, participants received a recommended submission profile according to their partitions. The distribution of submissions profiles was as follows:

- {{$S_1$, $S_2$}. {$S_3$}}
  - (L-I, H-I): 11
  - (M-I, H-I): 12
  - (H-I, H-I): 2
- {{$S_1$, $S_3$}. {$S_2$}}
  - (L-I, M-I): 5
  - (H-I, M-I): 6
  - (H-I, L-I): 6
  - (H-I, L-I): 1
- {{$S_2$, $S_3$}. {$S_1$}}
  - (M-I, L-I): 16
  - (H-I, L-I): 6
  - (H-I, M-I): 2
- {{$S_1$}, {$S_2$}, {$S_3$}}
  - (L-I, M-I, H-I): 24
  - (L-I, L-I, H-I): 5
- {{$S_1$, $S_2$, $S_3$}}
  - (L-I): 3
  - (M-I): 5
  - (H-I): 2

Finally, concerning response time, the average time spent per article was 29.47 seconds, with a median time of 19.99 seconds, and a standard deviation of 66.82 seconds.

In order to check the significance of these results, we performed different analyses. Firstly, a One Way ANOVA was carried out. We grouped participant responses into partitions and extracted the average number of correct answers. ANOVA results are described in **Table 8**. From these results, we can determine that it is very probable that at least one of the groups is statistically significant.

**Table 7** Partitions obtained from participants' responses.

| Partition | Number of participants |
|---|---|
| $K_{F_3} = \{\{s_3\}, \{s_1, s_2\}\}$ | 25 |
| $K_{F_2} = \{\{s_2\}, \{s_1, s_3\}\}$ | 18 |
| $K_{F_1} = \{\{s_1\}, \{s_2, s_3\}\}$ | 24 |
| $K_C = \{s_1, s_2, s_3\}$ | 10 |
| $K_F = \{\{s_1\}, \{s_2\}, \{s_3\}\}$ | 29 |
| **Total** | **106** |

**Table 8** One Way ANOVA test. Participants' responses.

| **Source** | **Treatment** | **Error** | **Total** |
|---|---|---|---|
| **Sum of squares** | 318.2653 | 418.6120 | 736.8774 |
| **Degrees of freedom** | 4 | 101 | 105 |
| **Mean square** | 79.5663 | 4.1447 | |
| **F statistic** | 19.1972 | | |
| **p-value** | 9.0404e-12 | | |

The second step in our analysis was to evaluate the relationships between different groups by performing a Turkey HSD test in order to determine whether the means from each group were significantly different. **Table 9** illustrates the Turkey HSD p-values

obtained. Most of comparisons have a p-value lower than 0.01 and may be considered significant. We can see that $K_F$ and $K_C$ groups are different from the rest of groups. Significant differences were not found between $K_{F_1}$, $K_{F_2}$ and $K_{F_3}$.

**Table 9** Turkey HSD p-values. Participants' responses. Bold cells show comparisons whose p-value is lower than 0.01.

| | $K_C$ | $K_{F_1}$ | $K_{F_2}$ | $K_{F_3}$ | $K_F$ |
|---|---|---|---|---|---|
| $K_C$ | | **0.0010** | **0.0010** | **0.0010** | **0.0010** |
| $K_{F_1}$ | **0.0010** | | 0.9000 | 0.9000 | **0.0010** |
| $K_{F_2}$ | **0.0010** | 0.9000 | | 0.9000 | **0.0011** |
| $K_{F_3}$ | **0.0010** | 0.9000 | 0.9000 | | **0.0010** |
| $K_F$ | **0.0010** | **0.0010** | **0.0011** | **0.0010** | |

Similar to the analysis performed with answers from participants, we studied the significance differences between time per response. A One Way ANOVA test and a Turkey HSD test were performed. Results from these tests show that there were significant differences between individuals with a $K_C$ partition and the rest of groups. But no differences were found between the other groups. Results are described in **Tables 10** and **11**.

**Table 10** One Way ANOVA test. Participants' time per response.

| **Source** | **Treatment** | **Error** | **Total** |
|---|---|---|---|
| **Sum of squares** | 2,856.9484 | 14,522.4698 | 17,379.4182 |
| **Degrees of freedom** | 4 | 105 | 109 |
| **Mean square** | 714.2371 | 138.3092 | |
| **F statistic** | 5.1641 | | |
| **p-value** | 0.0008 | | |

**Table 11** Turkey HSD p-values. Participants' time per response. Bold cells show comparisons whose p-value is lower than 0.01. Italic cell indicates a p-value lower than 0.05.

| | $K_C$ | $K_{F_1}$ | $K_{F_2}$ | $K_{F_3}$ | $K_F$ |
|---|---|---|---|---|---|
| $K_C$ | | **0.0036** | **0.0018** | *0.012* | **0.0010** |
| $K_{F_1}$ | **0.0036** | | 0.9000 | 0.9000 | 0.7844 |
| $K_{F_2}$ | **0.0018** | 0.9000 | | 0.9000 | 0.9000 |
| $K_{F_3}$ | *0.012* | 0.9000 | 0.9000 | | 0.9000 |
| $K_F$ | **0.0010** | 0.7844 | 0.9000 | 0.9000 | |

## 6 Discussion

In this paper, we tried to answer three questions. With respect to the first one we can say that most of participants in our experiment (76.4%) had the ability to distinguish

between low, medium, and high-quality manuscripts (making minor mistakes) and only a minority of individuals were unable to distinguish the quality of a manuscript.

The Fine Partition, $K_F$, was the most common in our experiment (27.4%), followed by $K_{F_3}$(23.6%). $K_C$, the coarse partition, was assigned to only 9.4%. Having a fine partition means that the participants can distinguish between low, medium, and high-quality articles. $K_{F_1}$ was also a frequent partition (22.6%).

Sometimes, it can be difficult to distinguish between low and medium or high and medium articles. However, this type of error is less critical than confusing a high-quality article with a low-quality one. We can say that almost all participants (about 73.6%) had $K_F$, $K_{F_3}$ or $K_{F_1}$ partitions, which means that they were able to distinguish between different types or manuscripts according to their quality while, at the same time, authors in this partition could differentiate between low- and high-quality documents.

With respect to the second question posed in the paper, results from the experiment also mean that, for experienced users, the amount of information used in our research (Title, Abstract, and Keywords) was enough for them to achieve a quality perception that was quite close to the actual quality of an article.

Participants were required to have a bachelor's degree and working experience in IT while some of them were also authors for computer science journals. So, it is likely that

most of them had some experience in reading and understanding scientific documents. It would be worthwhile to research more users that have a variety of backgrounds to check their abilities as well. In addition, for future research, we would like to introduce additional datasets to our website in order to be more helpful to researchers from different fields of knowledge.

Regarding the third question raised in our paper, we can say that an experienced author spends about 29 seconds reviewing the Title, Abstract, and Keywords from an article and deciding the quality of a manuscript. The median time to do so is about 20 seconds. Nevertheless, a high standard deviation was observed. Identifying the quality and the potential impact of an article in less than one minute can save a significant amount of time and effort for authors, who do not always have access to the full document in order to decide whether the manuscript would fit their needs or not.

**7 Conclusions**

In our paper, we proposed a tool to give authors accurate information about their skills when recognizing the potential of a manuscript and their recommended submission profile. We also wanted to know whether a minimal amount of information, consisting of the only the Title, the Abstract, and the Keywords, would be enough for a researcher to

determine the article's quality and to know how much time would be required to score the article.

For the purpose of our research, we designed a website where researchers could test their abilities by evaluating article information and sending it to a journal from one of three impact types. After designing and launching our website, we ran an experiment where 106 experienced users classified at least 15 articles. The experiment results indicate that most of them were able to determine the quality of classified articles accurately. An article required an average time of 29 seconds to perform the evaluation. In the light of the results achieved, we can say that the Title, Abstract, and Keywords provide, in most cases (90.6% according to results from the experiment), enough information to identify, at least, one of the three quality categories defined in this work (low, medium, or high quality).

Future research must test the website with non-experienced users and compare their results with the experienced users' group. Furthermore, we would like to add articles from different fields of knowledge to analyze researchers' behavior according to their varying backgrounds. Finally, the minimum amount of information necessary to accurately score the impact of a manuscript is also an open issue requiring further investigation.

## 9 Declarations

**Conflict of interests:** The authors declare that there are no conflicts of interest regarding the publication of this paper.

## References

1. Bai, X., Liu, H., Zhang, F., Ning, Z., Kong, X., Lee, I., & Xia, F. (2017). An Overview on Evaluating and Predicting Scholarly Article Impact. *Information*, *8*(3), 73. https://doi.org/10.3390/info8030073

2. Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Journal of the American Medical Association*, *287*(21), 2847–2850. https://doi.org/10.1001/jama.287.21.2847

3. Campanario, J. M. (1998). Peer Review for Journals as it Stands Today—Part 1. *Science Communication*, *19*(3), 181–211. https://doi.org/10.1177/1075547098019003002

4. Chamorro-Padial, J., Rodriguez-Sánchez, R., Fdez-Valdivia, J., & Garcia, J. A. (2019). An evolutionary explanation of assassins and zealots in peer review. *Scientometrics*, *120*(3), 1373–1385. https://doi.org/10.1007/s11192-019-03171-3

5. Chamorro-Padial, J. (2021). *Computer Science Articles & Journals* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/1268595

6. García, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2015). The author–editor game. *Scientometrics*, *104*(1), 361–380. https://doi.org/10.1007/s11192-015-1566-x

7. García, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2019). Do the best papers have the highest probability of being cited? *Scientometrics*, *118*(3), 885–890. https://doi.org/10.1007/s11192-019-03008-z

8. Garcia, J.A., Rodriguez-Sánchez, R. & Fdez-Valdivia, J. (2021). The editor-manuscript game. *Scientometrics* 126**,** 4277–4295. https://doi.org/10.1007/s11192-021-03918-x

9. Johnson, R., Watkinson, A., & Mabe, M. (2018). The STM Report: An overview of scientific and scholarly publishing. *International Association of Scientific, Technical and Medical Publishers*, 1–214.

10. Kassirer, J. P., & Campion, E. W. (1994). Peer Review: Crude and Understudied, but Indispensable. *JAMA: The Journal of the American Medical Association*, *272*(2), 96–97. https://doi.org/10.1001/jama.1994.03520020022005

11. Mavrogenis, A. F., Quaile, A., & Scarlat, M. M. (2020). *The good, the bad and the rude peer-review* (No. 3; Vol. 44, pp. 413–415). Springer. https://doi.org/10.1007/s00264-020-04504-1

12. Mengel, F. (2012). On the evolution of coarse categories. *Journal of Theoretical Biology*, *307*, 117–124. https://doi.org/10.1016/j.jtbi.2012.05.016

13. Menon, V., Varadharajan, N., Praharaj, S. K., & Ameen, S. (2021). Quality of peer review reports submitted to a specialty psychiatry journal. *Asian Journal of Psychiatry*, *58*, 102599. https://doi.org/10.1016/j.ajp.2021.102599

14. Okike, K., Hug, K. T., Kocher, M. S., & Leopold, S. S. (2016). Single-blind vs. double-blind peer review in the setting of author prestige. 316(12), 1315-1316. *Journal of American Medical Association*. https://doi.org/10.1001/jama.2016.11014

15. Özçakar, L. & Franchignoni F. & Kara, M. & Lasa, S. (2012). Choosing a scholarly journal during manuscript submission: The way how it rings true for physiatrists. *European journal of physical and rehabilitation medicine*. 48(4). 643-647.

16. Rajput, A.S. (2022). Scientific writing: an analysis of Pune-based climate scientists' perceptions and training needs. *Weather*, 77: 99-103. https://doi.org/10.1002/wea.3967

17. Richard B. P. et al. (2019). Are scientific editors reliable gatekeepers of the publication process?. *Biological Conservation*. 238. 108232. https://doi.org/10.1016/j.biocon.2019.108232

18. Rodriguez-Sánchez, R., García, J. A., & Fdez-Valdivia, J. (2016). Evolutionary games between authors and their editors. *Applied Mathematics and Computation*, *273*, 645–655. https://doi.org/10.1016/j.amc.2015.10.034

19. Schoenwolf G. C. (2013). Getting published well requires fulfilling editors' and reviewers' needs and desires. *Development, growth & differentiation*, *55*(9), 735–743. https://doi.org/10.1111/dgd.12092

20. Schuster, P., & Swetina, J. (1988). Stationary mutant distributions and evolutionary optimization. *Bulletin of Mathematical Biology*, *50*(6), 635–660. https://doi.org/10.1007/BF02460094

21. Tennant, J. P., & Ross-Hellauer, T. (2020). The limitations to our understanding of peer review. *Research Integrity and Peer Review*, *5*(1), 6. https://doi.org/10.1186/s41073-020-00092-1

22. Wallach, J.D., Egilman, A.C., Gopal, A.D. *et al.* (2018). Biomedical journal speed and efficiency: a cross-sectional pilot survey of author experiences. *Research Integrity and Peer Review* **3,** 1. https://doi.org/10.1186/s41073-017-0045-8

23. Wong, V.S.S., Strowd, R.E., Aragón-García, R. (2017). *et al.* Mentored peer review of standardized manuscripts as a teaching tool for residents: a pilot randomized controlled multi-center study. *Research Integrity and Peer Review* **2,** 6. https://doi.org/10.1186/s41073-017-0032-0

# Capítulo 4

# Palabras Clave

## 4.1. Text categorization through dimensionality reduction using Wavelet Transform

### 4.1.1. Datos generales

1. **Autores**: Jorge Chamorro-Padial, Rosa Rodríguez-Sánchez.
2. **Revista**: Journal of Information & Knowledge Management.
3. **Datos sobre la publicación**:
   - **Referencia**: Chamorro-Padial y Rodríguez-Sánchez (2020d).
   - **Volumen**: 19.
   - **Número**: 4.
   - **Páginas**: 2050039.
   - **Año**: 2020.
   - **Editorial**: World Scientific Publishing.
   - **DOI**: https://doi.org/10.1142/S0219649220500392.
4. **Estado**: Publicado.
5. **Métricas**:
   - **Factor de impacto**: 1,27 (SCImago, 2021).
   - **Ranking**:
     - *Emerging Sources Citation Index (ESCI)*:
       - *Information Science & Library Science*: Q3 - 88/164 (año 2021).
       - *Information Science & Library Science*: Q3 - 91/164 (año 2020).

- *SJR*:
  - *Library and Information Sciences*: Q3 - (año 2020).
  - *Computer Science Applications*: Q4 - (año 2020).
  - *Computer Networks and Communications*: Q4 - (año 2020).

### 4.1.2. Contribuciones principales

1. Se propone un nuevo método para resolver el problema de la reducción de dimensionalidad, en el ámbito de la Clasificación de Texto.

2. Mediante la aplicación de la transformada wavelet (Ver Subsección 2.5.2), se pueden obtener los términos más relevantes en un documento.

### 4.1.3. Resumen

El objetivo de la *clasificación de texto* es asignar una o varias categorías a un documento, para que sea identificable y distinguible en un *corpus*. Las estrategias utilizadas en este ámbito para representar la información, a menudo, implican que el número de dimensiones equivalga al tamaño del vocabulario de un corpus, siendo esta una cifra muy elevada y que nos lleva al problema de la *Maldición de la dimensión* (Trunk, 1979). Para abordar este problema, se utilizan técnicas de reducción de la dimensionalidad basadas en la selección de características (escoger las características más relevantes del espacio original) o en la extracción de características (proyectar un nuevo espacio con menor dimensionalidad) (Li et al., 2018).

En nuestro artículo, proponemos el uso de la *Transformada Discreta Wavelet* (que explicamos más en detalle en la sección 2.5.2), (TDW), como técnica de extracción de características. TDW tiene aplicaciones en el campo de la compresión de señales, donde comprime la energía de la señal de entrada en una serie de coeficientes, denominados *coeficientes wavelet* (Pearlman y Said, 2007). TDW también se ha utilizado en el campo de la Minería de Texto y, concretamente, nuestro trabajo continúa la propuesta de Xexéo et al. (2008) donde se propone un nuevo modelo de clasificación de documentos utilizando los coeficientes wavelet, donde cada documento es representado por coeficientes wavelet con una cantidad significativa de energía.

En nuestro método, en primer lugar se construye una matriz documento-término donde se relaciona la frecuencia de aparición de cada término en cada uno de los documentos que conforman el corpus de estudio. Esta matriz es la señal de entrada de la Transformada Wavelet y, a partir de aquí, se seleccionan los coeficientes con mayor energía y que sean relevantes en diferentes bandas realizando un **análisis multi-escala**, es decir, señalando aquellos términos que sean relevantes en diferentes escalas de la señal. De acuerdo con (Xexéo et al., 2008), se tratar de agrupar los términos con una mayor correlación, lo que reducirá las transiciones en la banda de detalles

hasta que se incremente la escala. Por nuestra parte, nosotros optamos por una estrategia diferente: agrupamos los términos por orden alfabético, lo que genera transiciones más acentuadas entre escalas.

Una vez que TDW ha realizado una descomposición multi-escala de la señal de entrada, es el momento de filtrar los coeficientes wavelet. Para ello, establecemos a cero el valor de cualquier coeficiente wavelet que quede por debajo de un umbral determinado. Si el valor de los coeficientes supera o iguala el umbral, entonces no realizamos ningún cambio en el mismo. El objetivo final de esta fase es conseguir seleccionar términos que cumplan las siguientes propiedades [1]:

1. El término debe ser relevante en comparación con otros otros términos del mismo documento, observados en la misma escala.

2. El término debe ser relevante en un documento, e irrelevante en otros documentos del corpus.

3. El término debe ser relevante cuando se hace una comparación de términos entre documentos.

Estas condiciones están relacionadas, a su vez, con las cinco propiedades que cumplen los términos relacionadas con la descomposición wavelet:

1. **Propiedad 1: Revelancia de un término entre documentos**. Un término debe ser altamente frecuente en un documento y altamente infrecuente en el resto del corpus, lo que implica que este término tiene una magnitud muy fuerte en los coeficientes wavelet de la banda HL.

2. **Propiedad 2: Irrelevancia de un término entre documentos**. Un término que tiene una frecuencia similar entre documentos implica que en la banda HL tendrá una magnitud débil.

3. **Propiedad 3: Relevancia de un término en un documento**: Un término que es altamente frecuente en un documento frente al resto de términos en el mismo documento implica que este término esta asociado con coeficientes wavelet de alta energia en la banda LH.

4. **Propiedad 4: Irrelevancia de un término en un documento**: Si un término presenta un frecuencia baja a lo largo de un documento los coeficientes wavelets asociado en la banda LH tendrán una energía baja.

5. **Propiedad 5: Relevancia de un término en el corpus**: Si un término es altamente frecuente a nivel de corpus, en comparación con

[1] Estas tres condiciones, en nuestro trabajo vienen representadas por la condición que denominamos *C*.

otros términos del corpus, sus coeficientes mostrarán una alta cantidad de energía en la banda HH.

Finalmente, tras aplicar un proceso de umbralización a los coeficientes wavelets, y así quedándonos con los coeficientes de mayor energía y aplicar sobre estos la Tansformada Wavelet inversa, obtenemos una aproximación a la matriz de frecuencias. Aquellos términos en la matriz de frecuencias aproximada sea mayor o igual que 1 se mantienen para representar los documentos. Ya solamente nos queda detectar qué términos han sido eliminados y producimos un vocabulario de salida que se corresponde con los términos que han superado el filtro de relevancia.

Siendo esta nuestra propuesta inicial, en la parte final del artículo, también estudiamos el efecto de aplicar (Latent Diritchlet Allocation, LDA) al conjunto de vocabulario reducido. Encontramos que esta combinación de métodos es capaz de realizar una mayor reducción de dimensionalidad que aplicando cada método por separado, lo que nos lleva a obtener muy buenos resultados en comparación con otras técnicas de reducción de dimensionalidad.

# Text categorization through dimensionality reduction using Wavelet Transform

Jorge Chamorro-Padial[a,1] , Rosa Rodríguez-Sánchez[a,2]

[a]Departamento de Ciencias de la Computación e Inteligencia Artificial., CITIC-UGR, Universidad de Granada, 18071 Granada, Spain.

Tel +34 958244019 Fax +34 958 243317

**Keywords**: *text classification; dimensional reduction; wavelet; transform; coefficient wavelet properties; term-document*

[1] Corresponding author. Email address: jorgechp@correo.ugr.es (Jorge Chamorro-Padial).
[2] Email address: rosa@decsai.ugr.es (Rosa Rodríguez-Sánchez).

# Text categorization through dimensionality reduction using Wavelet Transform

## Abstract

This paper proposes a new method of dimensionality reduction when performing text classification, by applying the Discrete Wavelet Transform to the document-term frequencies matrix.

We analyze the features provided by the wavelet coefficients from the different orientations: 1) The high energy coefficients in the horizontal orientation correspond to relevant terms in a single document. 2) The high energy coefficients in the vertical orientation correspond to relevant terms for a single document, but not for the others. 3) The high energy coefficients in the diagonal orientation correspond to relevant terms in a document in comparison to other terms.

If we filter using the wavelet coefficients and fulfill these three conditions simultaneously, we can obtain a reduced vocabulary of the corpus, with less dimensions than in the original one. To test the success of the reduced vocabulary, we recoded the corpus with the new reduced vocabulary and we obtained a statistically relevant level of accuracy for document classification.

## 1. Introduction

The goal of Text Classification (TC) problems is to assign one or more categories to a document inside a corpus. In Text Classification, choosing features to represent a document allows us to reduce the dimensionality of the corpus. When working with text data in fields like Text Classification or Text Mining, it's typical to deal with the curse of dimensionality: the number of dimensions is, in

most cases, the size of the vocabulary of a corpus. The computational effort required to perform a classification task with a very high number of dimensions tends to grow exponentially. In this context, the compression is very important if we want to improve the computational efficiency and to achieve better classification results especially when dealing with documents that have noise (Beck, Gonon, & Jentzen, 2020) .

If a vocabulary has noise and redundant information to represent a set of documents, this generates a *tf-idf* matrix with many more rows (more terms) and columns (more documents) than it would actually take for a document classification procedure to work, with the same or greater precision. With this premise in mind, we try to find a way to reduce the vocabulary such that the *tf-idf* matrix has a smaller dimension. This reduction will speed up the classification processes. Besides the representation with this reduced vocabulary is a subset of the original vocabulary. This aspect is important in order to have an interpretability of the new representation with respect to the original vocabulary.

In pursuit of this goal, we have analyzed the properties when the DWT is applied to the term-document matrix, in which each cell we have the frequency of a term in a document. In this article we postulate that the terms that must remain in the reduced vocabulary must correspond to a set of wavelet coefficients that have a high energy in the 3 orientation bands of the DWT. This is equivalent to searching for terms that comply simultaneously: 1) The term is relevant in a document; 2) The term is relevant in some documents and not in other documents; 3) The term is more relevant than other terms in other documents.

Therefore, this paper focuses on reducing the number of terms required to represent a set of documents that comprises a corpus in order to characterize them by using *tf-idf*. In this search for a reduced vocabulary, the DWT is applied into the document-term frequency matrix emphasizing that, during the process of computing the wavelet coefficients along the different scales and orientations we can identify properties inside the *tf-idf* characterization. Furthermore, the representation with

the reduced vocabulary is able to be interpreted along with the representation given by the original vocabulary. The method has been compared with the wavelet approximation (Xexéo et al., 2008) ,Latent Diritchlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), Chi-Squared and Mutual Information methods. Unlike the methods proposed in (Xexéo et al., 2008), Chi-Squared and Mutual Information, our proposal does not need an input parameter that indicates the length of the output vocabulary.

The rest of the paper is organized as follows:

- *Section 2*: A review of the methods to reduce dimensionality.
- *Section 3*: Describes the proposed methods and establishes the relationship between a relevant term in a document with the features presented with the wavelet coefficients.
- *Section 4*: Presents the results obtained by our methodology.
- Finally, in *Section 5* the main conclusions are presented.

## 2. Literature review

Typically, a Text Classification system is composed of (Kowsari et al., 2019): feature extraction/selection, dimensionality reduction, document characterization, classifier and evaluation. During the feature extraction, the training set is selected. The feature selection is increasingly more necessary when the number of samples is small and the observations have a high dimensionality (Gang Kou, 2020). The dimensionality reduction is not always included in every classification system but is relevant when the corpus is large or when there is noise in the documents. How the corpus will be represented is decided during the characterization phase and there are different possible characterizations: frequencies, *tf-idf*, etc. Next, the classifier is selected along with an error metric to compute the distance between documents. During the evaluation phase, we evaluate our model by testing it on a test set. Sometimes, it's possible to use a validation set.

When dealing with a significant amount of data distributed on a high number of dimensions, we can use dimensionality reduction techniques. These techniques project data into a, typically, linear combination of original features, but reduces the number of dimensions. We can divide dimensionality reduction techniques into feature selection and feature extraction. The goal of feature extraction is to project a new feature space with a low number of dimensions. Feature selection picks relevant features from the original space (Li et al., 2017).

One important step in the classification process is to select those features that can be relevant and provide useful information, and to remove those features that can generate noise or give poor information. *tf-idf* is one of the most used weighing strategies to perform dimensionality reduction. Moreover, *tf-idf* is not exempt from limitations when dealing with a large corpus which tend to generate a large number of dimensions, as well as worsen the results obtained by selection and classification algorithms. It's common to use some strategies to reduce dimensionality when dealing with a large corpus. For example, the use of Latent Semantic Analysis (LSA) and Linear Discriminant Analysis (LDA), sometimes in combination with *td-idf* ( Dzisevic & Sesok, 2019).

Two of the most common features selection methods are chi-squared (Fengxi S., 2005) and mutual information (Manning C.D, 1999). In terms of feature selection, (Párraga-Valle, García-Bermúdez, Rojas, Torres-Morán, & Simón-Cuevas, 2020) compares the use of chi-squared and mutual information as metrics to perform feature selection. According to the test performed in their work, chi-squared achieved better results than mutual information in all the benchmarks that were applied. In addition, the authors applied a standardized vocabulary but they didn't observe any relevant improvement with respect to applying a specific vocabulary. Chi-square is also used as a feature selection technique by (Bahassine, Madani, Al-Sarem, & Kissi, 2020). In their work, the authors applied feature selection with the goal of performing text classification using Arabic text documents and trying to overcome the limitations that result from chi-square.

Nevertheless, other authors use Mutual Information metrics to perform feature selection. In (Das et al., 2020) the authors propose three steps to perform feature selection: 1) identify irrelevant features, 2) identify potential redundant features and 3) identify redundant features. These steps, in combination with the use of classification techniques like Decision Tree, Naive Bayes and SVM, achieved better performance in contrast with other methods.

The Discrete Wavelet Transform (DWT) is usually used in signal compression. In particular, DWT is very effective in compressing most of the energy of the input signal into a few wavelet coefficients (Said & Pearlman, 2002). In Text Mining problems, DWT has been widely used (Al-Mofareji, Kamel, & Dahab, 2017; Park, Ramamohanarao, & Palaniswami, 2005; Xexéo, Souza, Castro, & Pinheiro, 2008; Hussin, El-Rube & Kamel,2008).

In (Xexéo et al., 2008) a new document classification method was presented using the wavelet coefficients to characterize each document. Previously, this method established term ordering as a way to build a document-term matrix. Each document is represented by a determined number of wavelet coefficients which contain high energy.

In (Hussin et al., 2008) a term in a document is represented by a weight. This weight is generated by a fusion process of a set of coefficients wavelets. They analyzed three fusion process: maximum, minimum and average for the set of coefficients wavelets.

In (Park et al., 2005) a new focus to classify text documents is proposed by using the DWT to distinguish term patterns in documents through different resolutions.

For his part, (Al-Mofareji et al., 2017) proposes a new method to group web documents by using multi-resolution analysis of the DWT. In this article, the authors use the wavelet transform to represent the document, but for dimensional reduction they use three methods: document frequency, mean *tf-idf* and term frequency variation (TFV).

Feature selection is also a task that can be carried out using wavelet, in (Mahajan, Sharmistha, & Roy, 2015) the authors start by representing the documents as vectors to transform the corpus in terms of wavelet coefficients. Then, highly informative coefficients are selected. The authors test their method to analyze content from Twitter as well as to detect spam, achieving good classification results.

More recently, DWT has been used to detect malicious bots from social network like Twitter ( Ameena, & Surendran, 2019). Each tweet was labeled as human, malicious or non-malicious content and a wavelet decomposition was applied in order to perform classification, Daubechies wavelet was used.

(Wolter, Lin, & Yao, 2020) applies wavelet to compress recurrent neural networks (RNN). They defined a method for learning local basis functions and let linear layers be compressed by using their wavelet-based representation.

## 3. Dimensionality reduction in the representation of documents

In this section we want to review the most relevant methods to perform dimensionality reduction in text document representation.

### LSA

Latent Semantic Analysis (LSA) (Furnas et al., 1988) was introduced as a technique to improve information retrieval (IR). The goal of LSA was to reduce the dimensions on IR problems. LSA is applied to Natural Language Processing to find relationships between documents. Thus, LSA relates documents with terms by generating a set of terms (topics or concepts) so that the words that appear in a similar text fragments have a similar meaning. LSA applies the following steps (Dumais, 2005):

1. **Document-term Matrix**: To represent the documents a document-term matrix is generated. The columns are the words of the corpus and each row represents a document in the

corpus. A cell *(i,j)* in the matrix refers to the number of times that the term *i* appears in the document *j*. This representation does not take into consideration the order of the terms. This way of representing a document is called *Bag of Words*.

2. **Frequency Matrix transformation**: The previously mentioned Document-Term Matrix is converted into a *tf-idf* Document-Term matrix. $tf-idf(i,j)=tf_{i,j}\times\log\left(\frac{N}{df_j}\right)$ where $tf_{i,j}$is the frequency of the term *j* in the document *i*, $df_j$ is the number of documents where the term *j* appears, and *N* is the total number of documents.
3. **Dimensionality reduction**: The matrix is reduced by applying a Singular Value Decomposition (SVD), where the *k* singular values are kept, and the rest are set to 0. In this way, we can get a k-dimensional approximation of the original matrix.

The document-term matrix is typically sparse and noisy so SVD is applied to reduce the number of dimensions. This dimension reduction process allows us to spend less resources during the classification phase. The problem with this method is that estimating the optimal number of dimensions is difficult and there is a lack of interpretability among the terms that form a certain topic. In addition, LSA needs a large set of documents and vocabulary to achieve positive results.

### LDA

Latent Diritchlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is a Bayesian method. Given a corpus represented by a set of terms, LDA obtains the set of topics related with the corpus. In this model, each document can be generated from a different topic distribution. In contrast with probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999), the topic distributions follows a sparse Dirichlet distribution as prior distribution and this assumes that documents are associated with a small set of topics and that every topic is formed by only a few words. The main idea behind LDA is that every document can be described as a topic distribution and, at the same time, every topic can be described as a word distribution.

LDA, in contrast with LSA, associates topics to words from the original vocabulary. Moreover, LDA requires, as an input parameter, the number of topics to be generated.

Nowadays, LDA is still being studied in different contexts (Edisn H. et. al, 2020; Zhou, 2020; Jedrzejowicz J., 2020). In (Edisn H. et. al, 2020) LDA is applied to analyse the transcripts of the U.S. Federal Open Market Committee. In (Zhou, 2020) news text is used as the research object and they propose a LDA text topic clustering algorithm based on the Spark big data platform. In (Jedrzejowicz J., 2020) a hybrid approach has been proposed to using the well-examined Latent Dirichlet Allocation algorithm expanded by the knowledge acquired via word embeddings representing words.

## Wavelets and Text Classification

The Discrete Wavelet Transform (DWT) produces a description of a signal in terms of frequency-time or scale-orientation. DWT works by applying the wavelet function to a fragment of the signal, resulting in a coefficient called wavelet coefficient. If the fragment is similar to the function wavelet, then the wavelet coefficients will have more energy. This comparison is repeated until every frame that shares the signal is completed. Then, the signal is split into new larger fragments and the process starts again. In that way, we can obtain a multiscale or multiresolution decomposition of the signal.

The wavelet functions are designed to identify where the transitions happen. On one hand, these transitions are coded into detail coefficients. On the other hand, approximation coefficients are smoothed out versions of the input signal. The smoothing becomes larger when the scale is increased. In (Daubechies & Bates, 1993) more information can be found about the Wavelet Transform.

By using the wavelet transform, we can expect that the dimensionality reduction can sustain a vocabulary compression process in text classification. In this way, many different articles (Al-Mofareji et al., 2017; Mahajan, Sharmistha, & Roy, 2015; Park et al., 2005; Xexéo et al., 2008; Hussin et al., 2008) have used the Wavelet Transform to classify a text. For example, (Mahajan et al., 2015)

focuses on classifying short documents but in dimensional reduction step they don't use the wavelet transform. They use the wavelet transform for representation the document. (Xexéo et al., 2008) reorganizes the terms of the corpus in order to represent them in the wavelet domain by setting the number of wavelet coefficients to normalize the corpus. Both papers apply the Wavelet Transform document by document, independently.

In our paper, we see the set of documents as a whole, represented by the document-term frequency matrix, which is the input to the Wavelet Transform. Then, we search for the coefficients that have more energy. The coefficient must also be relevant in the three different orientations. This condition is supported by the properties of *tf-idf* and expound them. Furthermore, we generate a reduced vocabulary that is a subset of the original vocabulary.

## 4. Methodology

The steps of the proposed method are as follows:

- **Corpus**: Every document in the corpus has been represented as a term sequence.
- **Cleaning**: The documents are pre-processed to remove words that could increase the entropy. Thus, special characters and punctuations are filtered out during this step. We used the *StopWords* list provided by *The NLTK Project*[3] to further filter the document. In addition, every word was transformed into its singular form by applying a lemmatizer. By doing that, we can prevent the possibility of a single concept being expressed by two or more different verbal forms.
- **Frequency Matrix:** A document $D_j$ is represented as a term vector $t_j=(t_{j1},t_{j2},\ldots,t_{jN})$, where N is the size of the vocabulary in the corpus. This representation lets us compute the frequency of every term in each document in the corpus and build a matrix *F* where *F(i,j)* represents the number of times that the term *j* appears in the document *i*.

[3] https://www.nltk.org/_modules/nltk/stem/wordnet.html

Ordering the terms is a key point to building the frequency matrix. The terms of the frequency matrix are placed in the columns. If we use a particular order for the terms the ability of the DWT to detect transitions will be affected.

If we define the order of the terms by choosing the pairs of terms with the maximum correlation, the wavelet transform will not show relevant information (transitions) in the detail bands until we increase the number of scales. This ordering system is proposed in (Xexéo et al., 2008).

In Figure 2, we use the Reuters dataset[4] where we can see (left side) the correlation between two consecutive terms when sorting is done looking for the pairs of terms with a maximum correlation (Xexéo et al., 2008). On the right side we can see the correlation between two terms when alphabetic sorting is applied.

In our analysis, the best results were achieved when using the alphabetic sorting. This ordering generates strong transitions since the lower scales (with more detail) make it possible for HL, LH and HH bands to have strong magnitudes.

Furthermore, the results of the classification of the documents do not show significant differences regarding the document ordering used in the frequency matrix as long as the number of scales is large enough.

The proposed method has been run on the same database several times to check if the ordering of the documents to build the frequency matrix affects the accuracy of the method. For this purpose, the results of the proposed method are shown in Figure 3 when executed 30 times on the Reuters database and in each execution the documents are randomly

[4] https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection.The dataset can be downloaded from http://www.daviddlewis.com/resources/testcollections/reuters21578/

arranged. As can be seen in the graphs, the dispersion of the results is within a range of one thousandth.

- **Multiscale Decomposition**: In this step, a multiscale decomposition method is applied to the frequency matrix using the Wavelet Transform. In general, a multiscale decomposition allows us to highlight details of a signal in different scales and orientations. In an analogous way, representing documents as a document-term frequency matrix allows us to highlight the most relevant terms in the corpus. At each scale, the multiscale decomposition method determines the detail (the details correspond with the bands *HL*, *LH*, *HH* in the diagram) and the approximation (*LL*) of the signal. The detail can be computed by comparing a value and its neighbourhood with a high-pass signal called wavelet function. The approximation is the smoothed input signal, with less detail than the original one. After getting the detail and the approximation we have to apply a new decomposition process over the approximation signal in order to obtain a new scale. This is equivalent to obtaining the detail value over a larger neighbourhood. Thus, increasing the scale makes the neighbourhood larger.

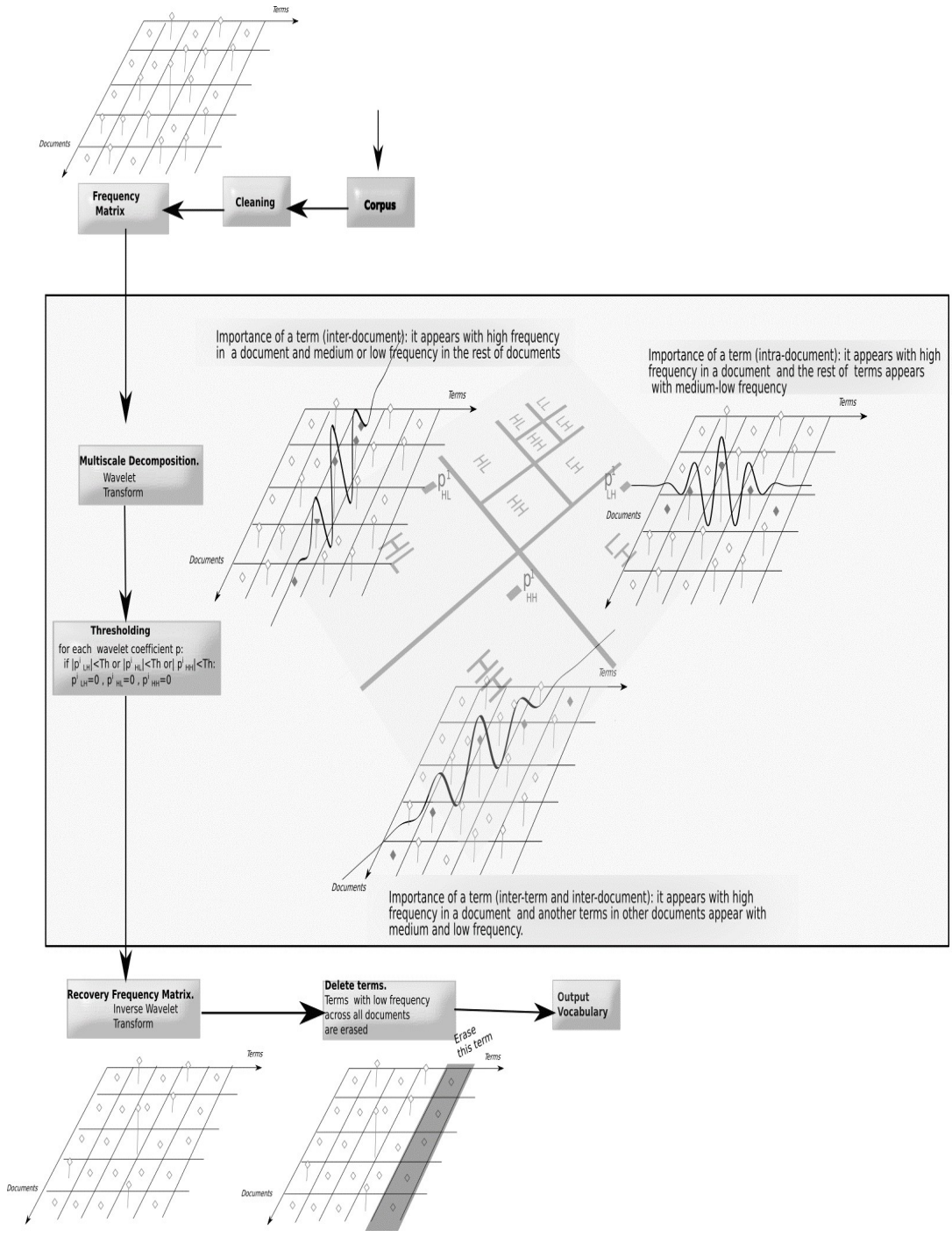



*Figure 1: Scheme of the proposed methodology.*

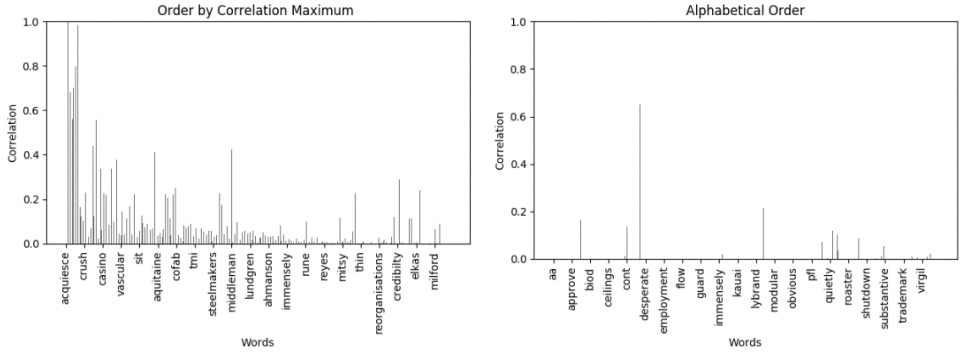



*Figure 2 Left: Sequence of terms in which two consecutive terms in the sequence have the màximum correlation. Right: correlation betweeb two terms whe tne sequence of terms is given in alphabetical order.*

The orientation bands allow us to analyse details in different orientations. In our case, when using DWT we obtain details in the horizontal, vertical and diagonal orientations. The information of the details reveals important properties that will be analysed in the next section. The details from the DWT can give us information about the relevance of a term (or set of terms) in the corpus when we do the analysis: 1) In the same scale, when compared with other terms in the same document. 2) How relevant a term of a document is when compared with the same term in other documents or even with other terms in other documents.

- **Thresholding**: In this step we consider the detail bands ($LH^i$,$HL^i$,$HH^i$ ), with respect to the approximation signal their coefficients are set to 0.

  Let the wavelet coefficients $p^i_{LH}$,$p^i_{HL}$, $p^i_{HH}$ represent the detail information of a set of terms in the horizontal, vertical and diagonal orientations at scale *i*. When the energy of some of these coefficients is lower than the threshold, these coefficients are set to 0 (that means that coefficients $p^i_{LH}$,$p^i_{HL}$, $p^i_{HH}$ are set to 0), otherwise the coefficients values are kept the same. This multiple condition, called *C* in this paper, refers to: 1) the set of terms associated with the coefficient are relevant when compared with terms in the same document at scale *i* ($p^i_{LH}$

must have high energy); 2) The set of terms associated with the wavelet coefficients are relevant in one document but not in the others ($p^j_{HL}$ must have high energy); 3) and finally, the term is relevant when compared with other terms in other documents ($p^j_{HH}$ must have high energy).

Fulfilling these three conditions means that *C* is met, and this generates restrictions to the terms that must be represented in the corpus. These terms will allow us to perform a better description of the corpus from the point of view of operations like classification, topic search, etc.

- **Recovery Frequency Matrix**: After the thresholding step, we apply the inverse wavelet transform to recover the frequency matrix. Let $\widetilde{F}(i,j)$ be the recovery frequency matrix.
- **Delete terms:** In this stage we detect non relevant terms to represent the corpus. Thus, we remove columns where all their values are lower than 1. These columns correspond to terms that will be removed.
- **Output Vocabulary:** We present the corpus after removing the irrelevant terms in our reduced vocabulary set.

Finally, we compute the *tf-idf* values for the set of documents with their reduced vocabulary. Next, a classification process is applied. We classify the test set to check the validity of the classification and the quality of the reduced vocabulary.

Figure 1 shows the diagram of the proposed methodology. In this figure, the first step, Corpus, is to represent each document of the dataset by a sequence of terms. The clean-up step then removes the noise and redundant information from documents, followed by building a term-document matrix. DWT is applied in the Multiscale decomposition step. On the right of the figure there is a schematic of the DWT that superimposes which transitions each orientation band should indicate in the frequency matrix. Wavelet coefficients with high magnitude represent strong transitions. In the Thresholding stage, the wavelet coefficients with high

magnitude in all bands (HL, LH and HH) are maintained and the rest are modified to zero. Then, applying the inverse of the DWT, a recovery frequency matrix is obtained. From this matrix, in the next step, the terms with low frequency across all documents are erased. The result of this process is a reduced vocabulary to represent the input corpus.

### Properties related with the terms in the corpus and the wavelet decomposition

In the following section we explain the different properties found in the document-term representation and, thus, in the terms of the corpus and the related wavelet decomposition:

- **Property 1: Inter document relevance of a term**. A term that is highly frequent in one document and infrequent in the rest of the corpus implies that, along the *HL* bands, the term has a strong magnitude in its related wavelet coefficients.
- **Property 2: Inter document irrelevance of a term.** A term having a similar frequency in all documents in the corpus implies that, along the bands *HL*, the term has a weak magnitude in its related wavelet coefficients (because there are no possible transitions).

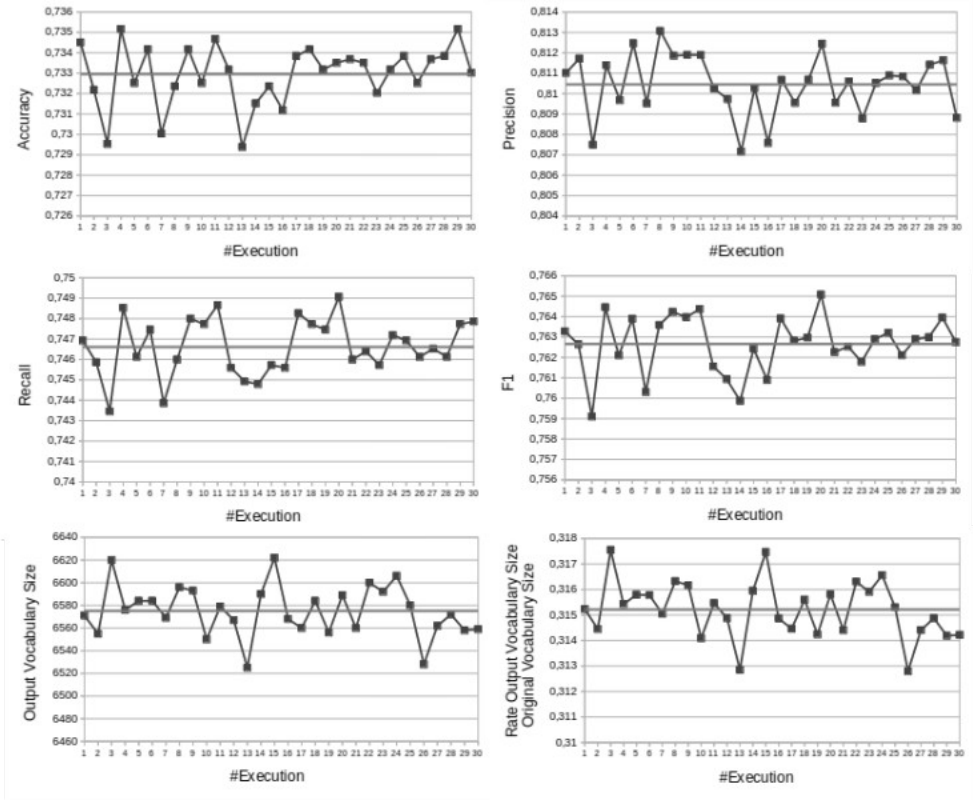



*Figure 3 Scores for Reuters dataset obtained by the proposed methodology. The method has been executed 30 times. In each plot the red line is the mean value of the score across the 30 executions.*

- **Property 3: Intra document relevance of a term.** A term that is highly frequent in one document in contrast with the rest of the terms of the same document implies that, along the *LH* bands, the term has a strong magnitude in its related wavelet coefficients.
- **Property 4: Intra document irrelevance of a term.** If two terms of a document have similar frequencies, then their related wavelet coefficients in the LH band will have low energy.
- **Property 5: Relevance of a term in the corpus.** If a term in a document is highly frequent in contrast with other terms of the corpus, then their related wavelet coefficients in the *HH* band will have high energy.

It is possible to make an analogy with the characterization of $tf-idf(i,j)=tf_{i,j} x log\left(\frac{N}{df_j}\right)$ where $tf_{i,j}$ is the frequency of the term *j* in the document *i*, $df_j$ is the number of documents where the term *j* appears and *N* is the size of the corpus. We can establish that properties 1 and 3 search the terms with a high *tf-idf* score: terms with high frequency in a single document and very infrequent in the rest of the documents. Property 5 is not considered by *tf-idf*.

## Toy Example

For the toy example, we can use the information from Table 1 to test this theory.The vocabulary of the corpus and the term frequencies of a document are shown in Table 2.

From Table 2 we can compute the frequencies matrix *F*:

| Code | Document |
|---|---|
| d1 | Shipment of gold damaged in a fire |
| d2 | Delivery of silver arrived in a silver truck |
| d3 | Shipment of gold arrived in a truck |
| d4 | In a silver truckt arrived gold |

*Table 1: Documents in the Toy Example.*

| Word | d1 | d2 | d3 | d4 |
|---|---|---|---|---|
| A | 1 | 1 | 1 | 1 |
| Arrived | 0 | 1 | 1 | 1 |
| Damaged | 1 | 0 | 0 | 0 |
| Delivery | 0 | 1 | 0 | 0 |
| Fire | 1 | 0 | 0 | 0 |
| Gold | 1 | 0 | 1 | 1 |
| Of | 1 | 1 | 1 | 0 |
| Shipment | 1 | 0 | 1 | 0 |
| Silver | 0 | 2 | 0 | 1 |
| Truck | 0 | 1 | 1 | 1 |

*Table 2: Frequencies of words in the documents shown in Table 1.*

$$F = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 2 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Then, we apply a DWT over the matrix. We use the Daubechies family of wavelets, specifically, we use *db1 Daubechies*[5] also called *Haar* Wavelet Transform[6]. Then, we set all the detail coefficients that do not fulfill the multiple condition *C* to 0. Coefficients related with the low pass band are completely removed. Then we apply the inverse wavelet transform getting the following results:

$$\hat{F} = \begin{pmatrix} 0.25 & 0 & 0.5 & 0 & 0.5 & 0.5 & 0 & 0 & 0.25 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 1.25 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0.25 & 0.25 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0.25 & 0 & 0 & 0.5 & 0 \end{pmatrix}$$

Then, we apply a ceiling to the next integer $\widetilde{F}$:

Every column of $\widetilde{F}$ represents the recovered frequency of a term along the whole corpus. All the terms whose related columns in $\widetilde{F}$ are equal to zero get removed.

In our example, we remove the words *{'a', 'arrived', 'in', 'of'}*. So that the reduced vocabulary is *{'damaged', 'delivery', 'fire', 'gold', 'shipment', 'silver', 'truck'}*.

$$\hat{F} = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

[5] Daubechies Wavelet. http://wavelets.pybytes.com/wavelet/

[6]

If we did not apply a ceiling and had kept only the frequencies that were greater than or equal to 1, we would have had the following results:

$$\hat{F} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

And the reduced vocabulary would be *{'silver'}*.

In Figure 4 we can see the steps taken to obtain the output vocabulary *{'silver'}*.

## 5. Experimental results

In this section we will test the proposed methods to reduce dimensionality in a corpus and then classify the documents. Before applying the classification, the documents, which have been represented with the new reduced vocabulary, are characterized by *tf-idf*.

The classification algorithms we used are k-Nearest Neighbors *k-NN* (Altman, 1992)
and Support Vector Machines (also Support Vector Classifier *SVC*) (Cortes & Vapnik, 1995).

For *k-NN* we used *k=3*. For *SVC* we applied a linear kernel.

We have compared the following methods:

1. **Original**: The original vocabulary of the corpus.
2. **LDA**: A reduced vocabulary according to the most relevant topics found by Latent Diritchlet Analysis (LDA).
3. **Xexéo:** This method was proposed in (Xexéo et al., 2008).
4. **Chi-Squared:** The $\chi 2$ test is used to select the most relevant terms.
5. **Mutual Information:** With this information measure, the most relevant terms are selected.
6. **Wavelet**: A reduced vocabulary computed by using the proposed method.

7. **F-Wavelet-LDA:** A reduced vocabulary applying *Wavelet* and, then, *LDA*.

## 5.1 Database

### *BBC Dataset*

The BBC database consists of 22,225 news articles from the BBC website. These news articles are from 2004-2005. There are five topical areas:

- Business
- Entertainment
- Politics
- Sport
- Tech

The dataset can be downloaded from Insight Project Resources site of the University College Dublin (http://mlg.ucd.ie/datasets/bbc.html). The corpus was split, 80% for training and 20% for testing.

### *The Reuters Dataset*

The Reuters-21578 database is a database with 21,578 news articles from the Reuters News Agency. These news articles were collected in 1987. In this paper we have used The Reuters-21578 Distribution 1.0, "ApteMod" version that contains 10,788 news articles divided into a training group and a test group. The training group consists of 7,769 documents and the 3,019 tests. There are 90 categories and each document has at least one category. The dataset was downloaded from http://www.daviddlewis.com/resources/testcollections/reuters21578/

### *The 20 newsgroup Dataset*

The 20 newsgroup dataset is composed of 18,000 publications extracted from newsgroups and categorized into 20 different topics. Each entry is assigned to a single topic. The number of documents in the training set is 11,314 and, for the test, 7,532. The dataset was downloaded from http://qwone.com/~jason/20Newsgroups/

## 5.2 Configuration

For our methods *Wavelet* and *F-Wavelet-LDA* it is necessary to define a value for the threshold determining the minimum number of wavelet coefficient to keep. In our case the threshold is the absolute value of the wavelet coefficient associated to the median. The number of scales is determined by the minimum dimension of the frequency matrices, minus 2.

In all the experiments, as mentioned before, we used a wavelet of the Daubechies family, specifically, "*db4*" Daubechies[7]. For all the methods tested we proceeded in the same manner: Firstly, obtaining the output vocabulary and, secondly, representing each document with this reduced vocabulary.

Thus, a word in a document is omitted if it is not present in the output vocabulary.

In the case of LDA, the input parameter was the number of topics, which we have defined as the number of categories in the database. From each topic a percentage of the most significant words in the topic were chosen.

For the Xexéo (Xexéo et al., 2008), Chi-Squared and Mutual Information methods, it is necessary to enter the length of the output vocabulary. For this, we have used two lengths of the output vocabulary.

Once all the documents with the new vocabulary are represented, each document was characterized with *tf-idf*.

Next, for the three databases, we classified the documents using k-NN and SVC.

[7] Daubechies Wavelet: http://wavelets.pybytes.com/

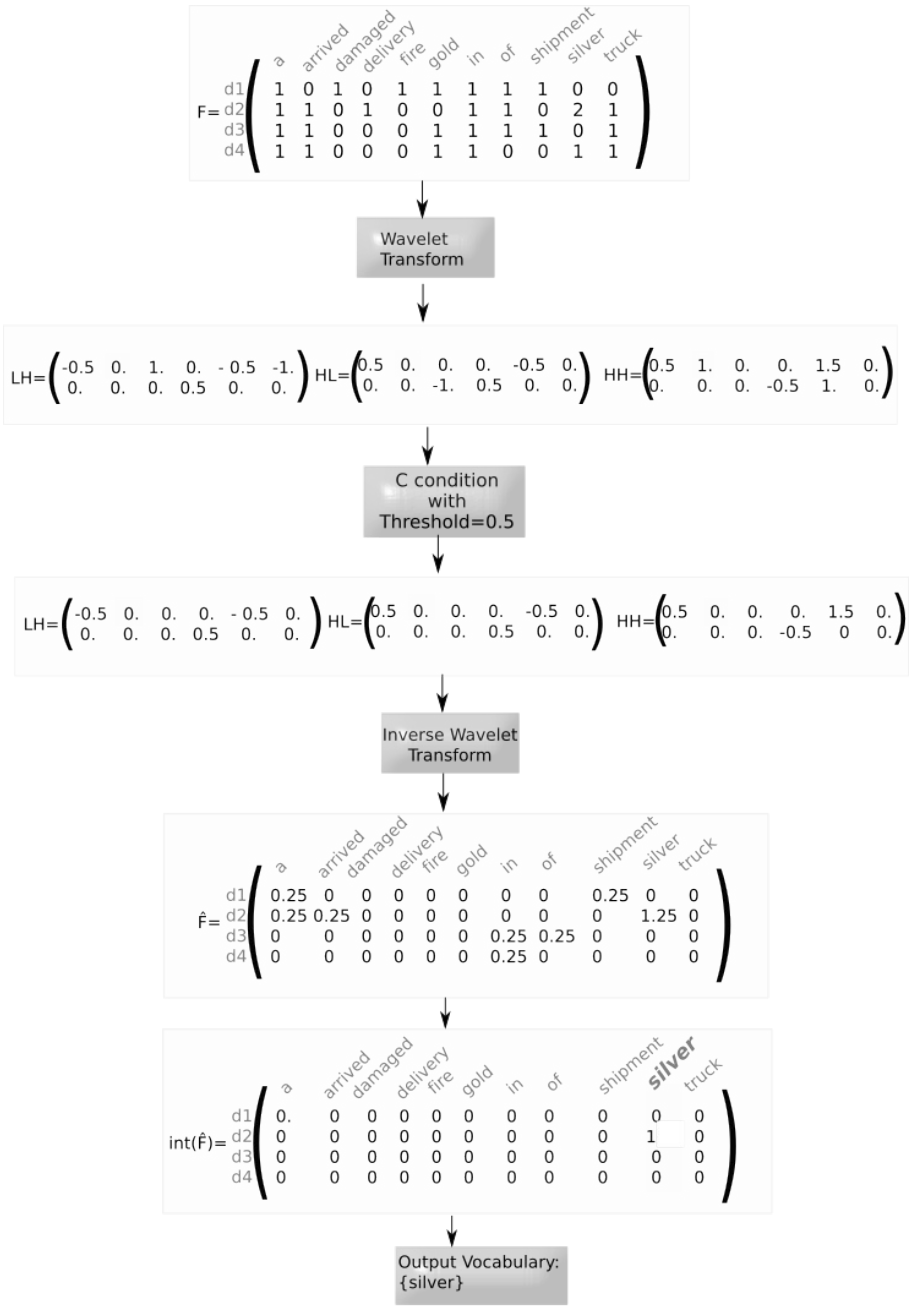



*Figure 4: Steps of the method for the toy example. In this example the wavelet used is the Haar Wavelet. The steps are:1.- Obtain the frequency matrix. 2.-Obtain the wavelet transform. 3.- Keep only the coefficient in LH, HL and HH that are bigger than or equal to the threshold (simultaneously).4.-Apply the inverse wavelet transform in order to obtain $\widetilde{F}$. Each coeficient is rounded down to the lower or equal integer (int($\widetilde{F}$)). 5.- The terms corresponding to columns with all values zeros are removed from the output vocabulary.*

## 5.3 Results

In Table 3 we can see the results obtained with the Reuters dataset.

Once the documents are cleaned, the vocabulary contains 20,845 terms. Applying LDA reduces the vocabulary to 14,763 terms, which represents a ratio of 0.71. The Xexéo method was run with two output vocabulary sizes: 4,096 and 8,192. In these cases, the ratio is 0.19 and 0.39, respectively.

*Wavelet* gets a vocabulary of 6,578 terms that represents a ratio of 0.315. And finally, *F-Wavelet-LDA* gets a vocabulary of 4,741 terms that represents a ratio of 0.227.

When we apply the k-NN classifier we get the following indicators:

- **F-Wavelet-LDA** achieves the best accuracy.
- The best precision is obtained with the original vocabulary.
- **Xexéo (**with a length of 4095**)** gets the best recall and F1 value. Next the best behaviour it is achieved by **F-Wavelet-LDA.**

When we apply the SVC classifier we get the following indicators:

- **F-Wavelet-LDA** achieves the best accuracy, recall and F1 value.
- **LDA** gets the best precision value.

*F-Wavelet-LDA* gets the best classification performance with both classifiers (k-NN and SVC).

In Table 4 we can see the results for the BBC dataset. The Xexéo method was run with two output vocabulary sizes: 4,096 (the most compressed vocabulary) and 8,192. In these cases, the ratio is 0.21 and 0.42, respectively.

The next most compressed vocabulary is obtained by *F-Wavelet-LDA* with a ratio of 0.22, followed by *Wavelet* and *LDA* with ratios of 0.31 and 0.89, respectively.

| k-Nearest Neighbors | | | | | | |
|---|---|---|---|---|---|---|
| **Method** | **Original** | **Xexéo** | | **LDA** | **Wavelet** | **F-Wavelet-LDA** |
| **Vocabulary Size** | 20845 | **4096** | 8192 | 14763 | 6578 | 4741 |
| **Accuracy** | 0.644 | 0.728 | 0.660 | 0.721 | 0.734 | **0.744** |
| **Precision** | **0.932** | 0.882 | 0.763 | 0.800 | 0.813 | 0.819 |
| **Recall** | 0.655 | **0.818** | 0.659 | 0.736 | 0.747 | 0.756 |
| **F1** | 0.764 | **0.821** | 0.673 | 0.753 | 0.768 | 0.773 |
| **Support Vector Classifier** | | | | | | |
| **Vocabulary Size** | 20845 | **4096** | 8192 | 14763 | 6578 | 4741 |
| **Accuracy** | 0.810 | 0.808 | 0.808 | 0.812 | 0.814 | **0.815** |
| **Precision** | 0.908 | 0.908 | 0.911 | **0.913** | 0.911 | 0.911 |
| **Recall** | 0.794 | 0.792 | 0.793 | 0.799 | 0.805 | **0.808** |
| **F1** | 0.836 | 0.834 | 0.835 | 0.840 | 0.843 | **0.846** |

*Table 3: Results for Reuters data set Distribution 1.0.*

When the classification with k-NN is performed, the highest accuracy is obtained with the vocabulary from *LDA*.

Regarding F1 values, the best results are obtained both with the original vocabulary and the reduced vocabulary generated by LDA, followed by *Wavelet* and *F-Wavelet-LDA*.

Note that *F-Wavelet-LDA* obtains precision, recall and F1 values that differs with the best results in the order $1.0e^{-3}$. Nevertheless, the differences in the reduced vocabulary sizes are remarkable.

When we apply the classification with SVC the best result is *Wavelet* for accuracy, recall and F1 values.

In *Table 5* we can see the results obtained for the 20 newsgroup dataset. In this case the original vocabulary has 80,783 terms. LDA reduces vocabulary to a ratio of 0.71, *Wavelet* reduces to a ratio of 0.38 and *F-Wavelet-LDA* gets the maximum reduction to 0.25. The Xexéo method enters an output vocabulary of size 16,384 and 32,768 with the vocabulary reduced to a ratio of 0.20 and 0.41, respectively.

For k-*NN* the best results for accuracy and F1 values are obtained for the original vocabulary. The best result for precision is obtained with Xexéo. And the best result for recall is achieved with

*Wavelet*. Alternatively, when we apply *SVC* the best results for recall, accuracy (equal to *F-Wavelet-LDA*) and F1 values, are obtained by *Wavelet*. The best value for precision with the original vocabulary and the Xexéo method.

| k-Nearest Neighbors | | | | | | |
|---|---|---|---|---|---|---|
| **Method** | **Original** | **Xexéo** | | **LDA** | **Wavelet** | **F-Wavelet-LDA** |
| **Vocabulary Size** | 19182 | **4096** | 8192 | 17054 | 5943 | 4288 |
| **Accuracy** | 0.923 | 0.728 | 0.851 | **0.924** | 0.922 | 0.921 |
| **Precision** | **0.955** | 0.882 | 0.915 | **0.955** | 0.953 | 0.953 |
| **Recall** | **0.951** | 0.818 | 0.908 | 0.949 | 0.948 | 0.947 |
| **F1** | **0.952** | 0.821 | 0.909 | **0.952** | 0.950 | 0.949 |
| **Support Vector Classifier** | | | | | | |
| **Vocabulary Size** | 19182 | **4096** | 8192 | 17054 | 5943 | 4288 |
| **Accuracy** | **0.939** | 0.926 | 0.928 | 0.936 | **0.939** | 0.938 |
| **Precision** | **0.983** | **0.983** | 0.976 | **0.983** | 0.982 | 0.981 |
| **Recall** | 0.976 | 0.970 | 0.967 | 0.975 | **0.977** | 0.975 |
| **F1** | **0.979** | 0.977 | 0.972 | **0.979** | **0.979** | 0.978 |

*Table 4: Results for BBC data set.*

| k-Nearest Neighbors | | | | | | |
|---|---|---|---|---|---|---|
| **Method** | **Original** | **Xexéo** | | **LDA** | **Wavelet** | **F-Wavelet-LDA** |
| **Vocabulary Size** | 80783 | **16384** | 32768 | 57465 | 30444 | 19931 |
| **Accuracy** | **0.639** | 0.290 | 0.483 | 0.636 | 0.638 | 0.624 |
| **Precision** | 0.768 | **0.822** | 0.813 | 0.765 | 0.765 | 0.764 |
| **Recall** | 0.583 | 0.290 | 0.483 | 0.636 | **0.637** | 0.624 |
| **F1** | **0.694** | 0.421 | 0.597 | 0.692 | 0.692 | 0.685 |
| **Support Vector Classifier** | | | | | | |
| **Vocabulary Size** | 80793 | **16384** | 32768 | 57465 | 30444 | **19931** |
| **Accuracy** | 0.645 | 0.626 | 0.641 | 0.648 | **0.653** | 0.651 |
| **Precision** | **0.932** | 0.929 | **0.932** | 0.930 | 0.927 | 0.925 |
| **Recall** | 0.656 | 0.638 | 0.652 | 0.661 | **0.665** | **0.665** |
| **F1** | 0.764 | 0.749 | 0.761 | 0.766 | **0.768** | 0.767 |

*Table 5: Results for 20 newsgroup.*

It should be noted that the differences between the values of *F-Wavelet-LDA* and the best values are of the order of $1.0e^{-3}$. This is a remarkable result seeing the difference in vocabulary reduction that *F-Wavelet-LDA* gets.

In Tabla 6 we have obtained the results for Chi-Squared and Mutual Information methods. For these methods, we have set the length of the output vocabulary to the length of the output vocabulary obtained for the Wavelet and F-Wavelet-LDA methods.

| k-Nearest Neighbors | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **DataSet** | **Reuters** | | | | **BBC** | | | | **20newsgroup** | | | |
| **Method** | **$\chi^2$** | | **Mutual Information** | | **$\chi^2$** | | **Mutual Information** | | **$\chi^2$** | | **Mutual Information** | |
| **Vocabulary Size** | 4741 | 6578 | 4741 | 6578 | 4288 | 5943 | 4288 | 5943 | 19931 | 30444 | 19931 | 30444 |
| **Accuracy** | 0.699 | 0.686 | 0.669 | 0.619 | 0.676 | 0.703 | 0.811 | 0.782 | 0.389 | 0.463 | 0.272 | 0.198 |
| **Precision** | 0.781 | 0.786 | 0.801 | 0.804 | 0.892 | 0.877 | 0.893 | 0.887 | 0.806 | 0.824 | 0.766 | 0.798 |
| **Recall** | 0.638 | 0.622 | 0.634 | 0.585 | 0.764 | 0.783 | 0.897 | 0.853 | 0.389 | 0.462 | 0.271 | 0.198 |
| **F1** | 0.642 | 0.626 | 0.687 | 0.638 | 0.788 | 0.798 | 0.885 | 0.856 | 0.482 | 0.581 | 0.382 | 0.253 |
| **Support Vector Classifier** | | | | | | | | | | | | |
| **DataSet** | **Reuters** | | | | **BBC** | | | | **20newsgroup** | | | |
| **Method** | **$\chi^2$** | | **Mutual Information** | | **$\chi^2$** | | **Mutual Information** | | **$\chi^2$** | | **Mutual Information** | |
| **Vocabulary Size** | 4741 | 6578 | 4741 | 6578 | 4288 | 5943 | 4288 | 5943 | 19931 | 30444 | 19931 | 30444 |
| **Accuracy** | 0.804 | 0.807 | 0.790 | 0.790 | 0.939 | 0.925 | 0.896 | 0.885 | 0.634 | 0.640 | 0.549 | 0.580 |
| **Precision** | 0.904 | 0.904 | 0.887 | 0.890 | 0.981 | 0.978 | 0.984 | 0.979 | 0.930 | 0.931 | 0.915 | 0.921 |
| **Recall** | 0.787 | 0.790 | 0.769 | 0.773 | 0.976 | 0.969 | 0.964 | 0.958 | 0.644 | 0.652 | 0.560 | 0.591 |
| **F1** | 0.830 | 0.832 | 0.812 | 0.817 | 0.978 | 0.974 | 0.973 | 0.968 | 0.755 | 0.759 | 0.684 | 0.709 |

*Tabla 6 Results obtained by Chi-Squared and Mutual Information for Reuters, BCC and 20newsgroup datasets.*

When comparing Tables Table 3,Table 4 and Table 5 with Table Tabla 6, , we can see that the results obtained by Wavelet and F-Wavelet-LDA are better than the results obtained by Chi-Squared and Mutual Information for all datasets.

## 8. Conclusions

In this paper a new *Wavelet* proposal has been given to reduce the dimensionality of a set of documents. For this, a multiscale decomposition of the document-term frequency matrix has been applied, using a wavelet transform.

Unlike the Latent Semantic Analysis (LSA), the proposed method does not impact the interpretability of the terms.

We have introduced properties that a term must have in order to remain in the reduced vocabulary. These properties are connected to the detail wavelet coefficients. Therefore, the terms that survive in the output vocabulary must be: important intra documents, important inter documents and important inter documents & inter terms. These properties have been correlated to high energy wavelet coefficients in the orientation bands (*LH*, *HL*, *HH*), respectively. If the wavelet coefficients on a scale associated with a term or set of terms have a high magnitude in the three orientations, then the associated terms meet the three properties and will remain in the output vocabulary.

Then, the reduced vocabulary corpus was classified to validate the method. It was tested with three different databases: The Reuters Dataset, BBC and The 20 newsgroup Dataset. The *Wavelet* method was compared with Latent Diritchlet Analysis *LDA*, (Xexéo et al., 2008), Chi-Squared and Mutual Information methods. *Wavelet* produced better results than *LDA* in terms of vocabulary compression, obtaining even better ranking values in the proposed databases.

An alternative method is to apply the *Wavelet* process and then the *LDA* method giving rise to the method called *F-Wavelet-LDA*.

At the same compression ratio of the output vocabulary, both Wavelet and F-Wavelet-LDA perform better than Chi-squared and Mutual Information methods.

Unlike the methods in (Xexéo et.al,2008), Chi-Squared and Mutual Information, *Wavelet* and *F-Wavelet-LDA* do not need to enter the size of the output vocabulary. Analyzing the results obtained with k-NN and SVC, the behavior of *Wavelet* and *F-Wavelet-LDA* is more robust than the Xexéo obtaining better results. When we analyzed the compression ratio of the output vocabulary, the *F-Wavelet-LDA* had a very competitive behavior in relations to the other methods. To summarize,

*Wavelet* and *F-Wavelet-LDA* are recommended in order to reduce dimensionality in the text classification process.

## 9. Acknowledgements

This research was sponsored by the Spanish Board for Science, Technology and Innovation under grant TIN-2017-85542-P, and co-financed with European FEDER funds.

## 10. Bibliography

Al-Mofareji, H., Kamel, M., & Dahab, M. Y. (2017). WeDoCWT: A New Method for Web Document Clustering Using Discrete Wavelet Transforms. *Journal of Information & Knowledge Management*, *16*(01), 1750004. https://doi.org/10.1142/S0219649217500046

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, *46*(3), 175–185. Retrieved from http://www.jstor.org/stable/2685209

Ameena N.S. and Surendran S. (2019) Identification of Malicious Bots in Twitter using Wavelets (August 3, 2019). Proceedings of International Conference on Recent Trends in Computing, Communication & Networking Technologies (ICRTCCNT) 2019. Available at SSRN: https://ssrn.com/abstract=3431587 or http://dx.doi.org/10.2139/ssrn.3431587

Beck, C., Gonon, L., & Jentzen, A. (2020). *Overcoming the curse of dimensionality in the numerical approximation of high-dimensional semilinear elliptic partial differential equations*. Retrieved from http://arxiv.org/abs/2003.00596

Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences*, *32*(2), 225–231. https://doi.org/10.1016/j.jksuci.2018.05.010

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Das, A. K., Kumar, S., Jain, S., Goswami, S., Chakrabarti, A., & Chakraborty, B. (2020). An information-theoretic graph-based approach for feature selection. *Sadhana - Academy Proceedings in Engineering Sciences*, *45*(1), 1–9. https://doi.org/10.1007/s12046-019-1238-2

Daubechies, I., & Bates, B. J. (1993). Ten Lectures on Wavelets. *The Journal of the Acoustical Society of America*, *93*(3), 1671–1671. https://doi.org/10.1121/1.406784

Dzisevic, R., & Sesok, D. (2019). Text Classification using Different Feature Extraction Approaches. *2019 Open Conference of Electrical, Electronic and Information Sciences, EStream 2019 - Proceedings*. https://doi.org/10.1109/eStream.2019.8732167

Dumais, S. T. (2005). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*(1), 188–230. https://doi.org/10.1002/aris.1440380105

Edison H & Carcel H (2020). Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. Journal Applied Economics Letters. https://doi.org/10.1080/13504851.2020.1730748

Fengxi S., Liu S., and Yang J. (2005). A comparative study on text representation schemes in text categorization.Pattern analysis and applications, 8(1-2):199–209

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. a., Streeter, L. a., & Lochbaum, K. E. (1988). Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure. *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 465–480. https://doi.org/10.1145/62437.62487

Gang Kou, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, Fawaz E.(2020) Alsaadi,Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods,Applied Soft Computing, 86, https://doi.org/10.1016/j.asoc.2019.105836.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *SIGIR 1999*, 50–57. https://doi.org/10.1145/312624.312649

Jedrzejowicz J., Zakrzewska M. (2020) Text Classification Using LDA-W2V Hybrid Algorithm. In: Czarnowski I., Howlett R., Jain L. (eds) Intelligent Decision Technologies 2019. Smart Innovation, Systems and Technologies, vol 142. Springer, Singapore

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information (Switzerland)*, Vol. 10. https://doi.org/10.3390/info10040150

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys*, *50*(6). https://doi.org/10.1145/3136625

Mahajan, A., Sharmistha, & Roy, S. (2015). Feature selection for short text classification using wavelet packet transform. *CoNLL 2015 - 19th Conference on Computational Natural Language Learning, Proceedings*, 321–326. https://doi.org/10.18653/v1/k15-1034

Manning C and Schütze H. (1999). Foundations of Statistical Natural Language Processing. The MIT Press ISBN-13: 978-0262133609

Hussin, MF, I El-Rube and MS Kamel (2008). Enhanced document clustering using fusion of multiscale wavelet decomposition. In IEEE/ACS International Conference on ComputerSystems and Applications (AICCSA 2008). https://doi.org/10.1109/AICCSA.2008.4493632

Párraga-Valle, J., García-Bermúdez, R., Rojas, F., Torres-Morán, C., & Simón-Cuevas, A. (2020). *Evaluating Mutual Information and Chi-Square Metrics in Text Features Selection Process: A Study Case Applied to the Text Classification in PubMed*. 636–646. https://doi.org/10.1007/978-3-030-45385-5_57

Park, L. A. F., Ramamohanarao, K., & Palaniswami, M. (2005). A novel document retrieval method

using the discrete wavelet transform. *ACM Transactions on Information Systems*, *23*(3), 267–298. https://doi.org/10.1145/1080343.1080345

Said, A., & Pearlman, W. A. (2002). A New Fast/Efficient Image Codec Based on Set Partitioning in Hierarchical Trees. In *Wavelet Image and Video Compression* (pp. 157–170). https://doi.org/10.1007/0-306-47043-8_9

Wolter, M., Lin, S., & Yao, A. (2020). *Towards deep neural network compression via learnable wavelet transforms*. Retrieved from http://arxiv.org/abs/2004.09569

Xexéo, G., Souza, J. de, Castro, P. F., & Pinheiro, W. A. (2008). Using Wavelets to Classify Documents. *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 272–278. https://doi.org/10.1109/WIIAT.2008.221

Zhou, Z., Qin, J., Xiang, X., Tan, Y., Liu, Q. et al. (2020). News Text Topic Clustering Optimized Method Based on TF-IDF Algorithm on Spark *CMC-Computers, Materials & Continua*, *62(1)*, 217–231.

## 4.2. Finding answers to COVID-19 specific questions: An Information Retrieval System based on latent keywords and adapted TF-IDF

### 4.2.1. Datos generales

1. **Autores**: Jorge Chamorro-Padial, Francisco-Javier Rodrigo-Ginés y Rosa Rodríguez-Sánchez.

2. **Revista**: Journal of Information Science.

3. **Datos sobre la publicación**:

   - **Referencia**: Chamorro-Padial et al. (2022).
   - **Año**: 2022.
   - **Editorial**: SAGE Publishing.
   - **DOI**: https://doi.org/10.1177/01655515221110995.

4. **Estado**: Publicado.

5. **Métricas**:

   - **Factor de impacto**: 2,462 (JCR, 2021).
   - **Ranking** [2]:
     - *Social Sciences Citation Index (SSCI)*:
       - *Computer Science, Information Systems*: Q3 - 112/164 (año 2021).
     - *Science Citation Index Expanded (SCIE)*:
       - *Information Science & Library Science*: Q3 - 43/84 (año 2021).

### 4.2.2. Contribuciones principales

En este trabajo, utilizamos minería de texto para encontrar información latente en documentos científicos, en combinación con palabras clave escogidas por un usuario, que nos permitan encontrar respuestas a preguntas complejas. Proponemos un modelo que ha sido validado por humanos (con estudios superiores y experiencia laboral o científica en el ámbito de las Ciencias de la Salud) por medio de diferentes cuestionarios.

[2] A fecha de depósito de esta tesis, aún no se disponen de datos del año 2022.

### 4.2.3. Resumen

En el año 2020, durante el desarrollo de esta Tesis Doctoral, asistimos a la emergencia sanitaria causada por la Pandemia de COVID-19. Este suceso desencadenó una muy elevada producción de literatura académica con el fin de dar explicación a la pandemia desde diferentes perspectivas y ámbitos de conocimiento. Gestionar esta gran cantidad de información supuso un reto en toda regla que, además, tuvo que resolverse con la mayor premura posible.

En marzo del 2020, *Allen Institute for AI* [3], en coordinación con la Casa Blanca [4] publicaron un dataset [5] que contenía, en el momento de la redacción de nuestro artículo, unos 50.000 artículos científicos relacionados con la pandemia o con el *SARS-CoV-2*, así como otras áreas de estudio relacionadas. Junto a la publicación del mencionado dataset, se pidió la contribución de la comunidad para aportar ideas con respecto a la búsqueda de información concreta por medio de técnicas de minería de texto y procesado de lenguaje natural.

En nuestro caso, realizamos una aportación a esta iniciativa explorando la efectividad de un método que combinase palabras clave seleccionadas por un usuario que busca respuestas a una determinada pregunta, con información latente presente en un artículo científico [6].

Para la extracción de la información latente, primero generamos pseudo-documentos Zuo et al. (2021) que, en un siguiente paso, son pre-procesados. Una vez que tenemos un pseudo-corpus pre-procesado, entonces aplicamos *Latent Dirichlet Allocation* para extraer los tópicos de cada pseudo-documento. Los términos más relevantes de cada tópico son incluídos en un conjunto de palabras clave *latentes* la que denominamos *LDA terms* en contraposición con las palabras clave introducidas por el usuarios, *input terms*.

A partir de aquí, generamos una matriz de co-ocurrencia entre *LDA terms* e *input terms* con el fin de comprobar el número de veces que una palabra de un conjunto co-ocurre con una palabra del otro conjunto. A partir de esta matriz, calculamos la puntuación de cada términos según *Topic Inverse Document Frequency* (TIDF). TIDF es una adaptación de TF-IDF propuesta en nuestro artículo para puntuar los diferentes términos. Aquellos términos que consigan la mayor puntuación y un determinado umbral, son incorporadoes a *input terms*. A partir de aquí, se repite todo el proceso durante tantos ciclos como se configure el algoritmo de nuestro modelo.

Tras la propuesta del modelo, se procedió a realizar una validación del

[3] https://allenai.org/ (Accedida el 2 de mayo del 2023).

[4] White House Office of Science and Technology Policy. https://www.whitehouse.gov/ostp/ (Accedida el 2 de mayo del 2023).

[5] https://allenai.org/data/cord-19 (Accedida el 2 de mayo del 2023).

[6] nótese que esta información es estática, y no varía en función de la información que estemos buscando, en contraposición a la información introducida por el usuario, que varía en función del objetivo de la búsqueda.

mismo mediante dos evaluaciones diferentes: una evaluación *a priori*, realizada por participantes en dos cuestionarios diferentes donde tuvieron que evaluar el rendimiento del modelo. Y una evaluación *a posteriori* donde se realizó una descarga y clasificación manual de 150 artículos seleccionados por nuestro método, y donde comprobamos si realmente eran artículos válidos. Los resultados de la validación fueron positivos en ambos casos.

### 4.2.4. Logros

Este artículo fue incluído en la *SAGE Public Health Emergency Collection* [7], iniciativa mediante la cual el editor distribuye voluntariamente el artículo bajo la modalidad de Acceso Abierto [8].

[7] https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/ (Accedida el 2 de mayo del 2023).

[8] El artículo se puede encontrar en esta modalidad en el repositorio de la National Library of Medicine (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9379592/) así como en Europe PMC https://europepmc.org/article/PMC/PMC9379592, ambos parte de PubMed.

# Finding answers to COVID-19 specific questions: An Information Retrieval System based on latent keywords and adapted TF-IDF

**Jorge Chamorro-Padial**
CITIC-UGR, Universidad de Granada, 18071 Granada, Spain.
Corresponding author.

**Francisco-Javier Rodrigo-Ginés**
NLP & IR Group, UNED, 28040 Madrid, Spain.

**Rosa Rodríguez-Sánchez**
Departamento de Ciencias de la Computación e I.A, CITIC-UGR, Universidad de Granada, 18081 Granada, Spain.

**Abstract**

The scientific community has reacted to the COVID-19 outbreak by producing a high number of literary works that are helping us to understand a variety of topics related to the pandemic from different perspectives. Dealing with this large amount of information can be challenging, especially when researchers need to find answers to complex questions about specific topics.

We present an Information Retrieval System that uses latent information to select relevant works related to specific concepts. By applying LDA models to documents, we can identify key concepts related to a specific query and a corpus. Our method is iterative in that, from an initial input query defined by the user, the original query is expanded for each subsequent iteration. In addition, our method is able to work with a limited amount of information per article.

We have tested the performance of our proposal using human validation and two evaluation strategies, achieving good results in both of them. Concerning the first strategy, we performed two surveys to determine the performance of our model. For all the categories that were studied, precision was always greater than 0.6, while accuracy was always greater than 0.8. The second strategy also showed good results, achieving a precision of 1.0 for one category and scoring over 0.7 points overall.

**Keywords**

Information Retrieval; Latent Dirichlet Allocation; TF-IDF; ATF.IDF; Document Filtering; COVID-19

# 1. Introduction

Coronavirus disease (also called COVID-19), was first detected in Wuhan, in the Hubei province of China and was reported to the WHO by the Chinese government on December 31, 2019 [1]. The illness has been spreading since the beginning of 2020 and officially became a pandemic in March when the virus had already infected more than 150,000 people worldwide [2,3].

Since the beginning of 2020 and due to the social, economic, and political repercussions generated by the evolution of the virus [4,5], the media and the scientific community have published an incredible amount of information on the aforementioned disease [6]. In just three months from the initial notification of the disease by the Chinese authorities, a total of 1,596 publications had been generated about COVID-19, 66% of them from China, while in April the number of publications rose to more than 6,500 [7,8]. At a social level, the pandemic has also attracted the attention of social network users [5]. In these circumstances, the excess of information and the lack of knowledge about the pandemic has led to the spread of inaccurate or false information (also known as *fake news*).

In the current environment, it is necessary and relevant to establish strategies to analyze the large amount of information that is continuously being generated in the scientific world in order to make the task of responding to the different needs that arise from different fields easier. Such fields include, health, social, economics, ethics, educational, and political, to name a few.

In this article, we propose a method to extract scientific literature based on a variety of topics. Our method facilitates the work of identifying which works written about COVID-19 could help respond to certain questions. Furthermore, our method also makes it possible to differentiate between articles that do and do not address purely health-related issues. This paper aims to make it easier for researchers to find and retrieve scientific literature related to complex or abstract topics, thus making it easier to find answers to complicated questions and to provide them with a complementary method for using information retrieval systems based on standalone keywords.

To do this, we worked with a dataset of articles on COVID-19 as well as other related areas. We extracted the different topics for each article, analyzing its title and abstract. Then, we filtered the documents by performing a coincidence analysis on the terms of the topics with the terms of the query.

We propose a method to find answers to complex and abstract questions by exploring latent concepts hidden in the titles and abstracts of scientific works. Our main goal was to respond to the call to action from the White House and other various research groups. They had prepared a dataset of COVID-19 related articles so as to solve urgent and relevant problems related to the pandemic[9]. While plenty of proposals have been offered by different authors (some of them are mentioned in the following section), we wanted to design a simpler model that would be able to find answers on COVID-19 adequately. More importantly, humans have validated our method, while most authors use automatic systems to measure or benchmark their proposals.

While our method has been tested specifically with COVID-19 data, it can be generalized and used in different areas.

Our work is structured as follows:

- The “State of the Art” section reviews the latest published studies, mainly in the field of bibliometrics and Natural Language Processing, as related to the COVID-19 disease.
- The “Methodology” section describes the dataset used and the proposed method.
- The “Results” section shows the results obtained by our method.
- Finally, the “Conclusions” section states the main findings of our work.

## 2. State of the Art

Since the outbreak of the pandemic, there have been many efforts to analyze the behavior of the scientific community through Bibliometry. Other authors [10,11] analyze the relationships between authors and highly cited articles about both the COVID-19 disease and the SARS-CoV-2 virus, as well as the number of papers produced by country, providing empirical data that proves the growing interest that the pandemic has had among not only authors, but also journals of very high impact. In both papers, scientific publications are also broken down by different subject areas, with epidemiology and virology being the fields that have received the greatest number of publications. Although [3] also performs a bibliometric analysis where they reached similar conclusions as [10,11], the authors also provided an interpretation for the significant differences in publications observed by country, based solely on the Gross Domestic Product (GDP) and the number of inhabitants. A high level of saturation in the healthcare system can have a negative impact on the number of publications in a country. This is the case, for example, in Italy.

This massive amount of information about the pandemic requires useful strategies that can help facilitate the scientific community in finding the desired information, while weeding out those works that may not be relevant for them. The large number of pre-existing works in the highly specific fields in the world of Medicine (Chahrour et al., 2020; Lou et al., 2020; Nasab & Rahim, 2020) also makes it more difficult to find information on certain less well-studied areas, such as education or ethics.

In addition, the evolution of the pandemic has unleashed plenty of social repercussions that should be studied and taken into consideration. For example, the vaccine opposition and COVID-19 denial movements which have had a considerable impact on social networks [12,13]. Work-from-home has also become a trend during the outbreak and is paving the way for an important transformation in terms of labour relations [14]. The pandemic has also had repercusions on cities, where marginalized populations are receiving a disproportional impact on their health and well-being. Urban planning is also learning and growing from this situation as well [15].

[10] highlights the limitations of their bibliometric analysis under the current conditions, where a large number of articles on the subject of COVID-19 are being published, in different languages. Being such a current event, it is impossible to draw firm conclusions in a field so dependent on current events.

In order to facilitate the task of obtaining relevant information for researchers, it is possible to apply techniques from the field of Natural Language Processing (NLP). IBM, for example, offers a service to extract relevant content from a corpus of articles about COVID-19, allowing researchers to ask specific questions about COVID-19 and analyze the related existing information [1].

Natural Language Processing is also being used to analyze the social interactions caused by the pandemic.[16] extracts topics from a corpus of text from different social networks and performs a discourse analysis to infer key concepts that have been used by the users of social media and the patterns of information dissemination. [17] analyzes Twitter with text mining techniques in order to conduct a multi-language analysis of the speech. In [18] an analysis of comments on Twitter about the pandemic was carried out but, this time, by analyzing the dissemination of truthful information and misinformation through the social network. In [19] the authors analyzed the phenomenon of sinophobia on 4chan and Twitter in relation to the pandemic through word embeddings, linking different news about Donald Trump, the World Health Organization (WHO), and the Chinese Government.

With the aim of helping researchers apply Natural Language Processing in the search for information concerning the pandemic, [20] have designed a toolbox that includes a set of English dictionaries with relevant concepts related to COVID-19. [21] lists the different areas of study where the use of artificial intelligence and machine learning could be relevant, as well as a compilation of datasets, resources, and initiatives carried out to improve the current knowledge about the disease.

Keyword extraction models have been widely used to classify different domains of knowledge in scientific papers [22]. We can define keywords from two different perspectives: sociocultural and statistical. Traditionally, any word that includes culturally and socially relevant concepts has been intuitively considered as a keyword [23]. From the point of view of corpus linguistics, keywords are extracted by using statistical processes, commonly comparing their frequency in the text to be analyzed with their frequency in a reference corpus. Using this type of technique, three types of keywords are usually obtained: proper names, concepts that explain

[1] https://www.research.ibm.com/covid19/deep-search/ Retrieved December 28, 2021.

the content of the text, and frequent words such as pronouns and prepositions that can be used as style and not content indicators [24].

Some information retrieval systems have been created to help researchers find relevant information. Named Entity Recognition (NER) can be useful in collecting COVID-19 information from statements, which can be considered an alternative to document retrieval systems. [25] Developed a web-based system to find textual evidence from COVID-19 document corpus.

CO-Search is an information retrieval system that is able to extract search queries from natural language questions and to retrieve scientific literature about COVID-19 [26]. CO-Search uses a SBERT model to create a latent space with queries and documents. CO-Search results are evaluated using TREC-COVID, an evaluation system that helps researchers find searching algorithms and information discovering methods to manage the existing literature around COVID-19 [27].

For researchers, collecting scientific literature is crucial in order to be up to date with the newest and most relevant knowledge for their research areas and to provide solid background knowledge that will allow them to effectively contribute to their field. The explosive growth of new scientific literature makes it difficult to identify suitable papers and is becoming an increasingly complex task[28].

When doing a literature review, researchers need to maximize the “relevance” of the collected literature, but “relevance” is not directly measurable, and some level of uncertainty is inevitable [29]. While information retrieval systems are nowadays an cornerstone for researchers, question answering systems are becoming a powerful tool that may help to find more relevant knowledge [26] [30]

# 3. Methodology

## Dataset description

In March 2020, due to the COVID-19 global pandemic, the *Allen Institute for AI* coordinated by the *White House Office of Science and Technology Policy* and in association with several initiatives, published CORD-19[2], an open dataset of over 50,000 papers on COVID-19, SARS-CoV-2, coronavirus, and other related study areas.

The idea behind the publication of this dataset was that after the increase in the academic literature on COVID-19 [8], the computer science community could apply text mining and natural language processing methods in order to digest and retrieve significant information and provide it to the medical research community.

The CORD-19 dataset contains multilingual information, but we have only dealt with English papers in this study. Thus, a new stage removing non-English information was applied.

[2] https://www.semanticscholar.org/cord19

After deleting non-English or duplicated papers and discarding papers without abstracts, titles, or references, we got a reduced dataset of 25,004 instances. Each instance of the dataset has the following information: title, authors, abstract, body text, references, and publication date.

## Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative statistical model used to extract the latent topic structure of text documents. LDA is a machine learning technique used in different areas such as retrieval field, document classification, and topic modeling [31].

In this paper, we've implemented the Latent Dirichlet Allocation provided by the Scikit-Learn project: a machine learning library for the Python Programming Language [3].

## Method

In this section, we will present the main contribution of this study: an information retrieval system that allows users to extract relevant papers when given certain search terms. Our approach is based on keyword extraction using the LDA topic modeling technique on pseudo-documents generated from the papers in the dataset.

The idea of this work is to relate both concepts of keywords. By adding latent keywords in the dataset papers, we can broaden the search spectrum and obtain more relevant results. As in other Information Retrieval Systems, the user introduces input information to perform a query and obtain an output. The input information is a set of keywords containing the most important ideas that the user wishes to examine. In our method and thanks to LDA, we extracted new latent keywords that complemented the input provided by the user.

From now on, we will refer to the input keywords provided by the query as *input_u*. Those latent keywords extracted by LDA will be referred to as *topic_terms*.

To be noted, *topic_terms set* provides information about the latent structure of the corpus and is independent of the query. *input_u set* gives information about the query so that their main goal is to facilitate which terms and concepts from the corpus are desired by the user.

In order to obtain the *topic_terms set*, we have processed each instance of the dataset in three phases: Pseudo-document generation, Text preprocessing, and Topic modeling. Figure 1 shows a schematic representation of our model.

**Pseudo-document generation**. In this initial step, for every paper in the dataset, we aggregate it into a pseudo-document using the title of the paper, the text of its abstract, and

[3] https://scikit-learn.org/stable/index.html Retrieved June 3, 2020.

the titles of its references. The generation of pseudo-documents is a commonly used strategy to combat data sparsity [32].

The use of the title and the abstract when extracting key information from a paper is quite common since it contains many keywords in a very limited space. The text of the paper is not usually taken into account since its content are usually quite heterogeneous, and the results vary greatly depending on the section being analyzed [33]. That is why we have aggregated the title and the abstract and not the full text into the pseudo-document.

References are also a useful source of information in articles. Bibliographical coupling and co-citation were used in the very first approach to study relationships between articles [34]. In addition, references have been used to extract key concepts of a document in order to generate better keywords [35]. In the context of this work, where keywords are absent in the dataset, it is important to extract key concepts in order to help us to analyze the latent information of each article.

**Text preprocessing**: In this phase, we conducted a series of tasks that aim to transform the text into a more digestible form so that the irrelevant information is reduced and the LDA model can perform better.

To preprocess a pseudo-document, we performed the following tasks:

- The text is converted to lowercase and the punctuation is deleted.

- We applied a *stop words* filter: Stop words are words that do not contribute any meaning to the document, such as articles, pronouns, and prepositions. Removing stop words avoids generating style indicator keywords that are not useful in the desired context. We have used the list of clinical stop words provided by [36].

- Each word is lemmatized: The lemmatization process is a linguistic process by which the root of a word is determined. We have used the spaCy lemmatizer algorithm [4].

- For reasons of efficiency, only unigrams are used in our model, but some n-grams have been kept for relevant reasons, those n-grams are: *World Health Organization*, *Public Health*, *Social Media*, *Fake News*, and *Social Sciences*.

- Finally, papers written in languages other than English are eliminated in order to reduce the computing time in the keyword topic modeling step.

**Topic modeling:** LDA analysis was applied for each article of the dataset in order to extract the latent topics hidden in the documents. It is important to note that we have not applied the same model for the whole corpus. A new LDA model was generated for each article. Our goal here was to detect minority topics in the dataset that often can be hidden in a single article. Excess noise from popular topics could be introduced if the same LDA model was applied to the entire corpus.

[4] https://www.nltk.org/_modules/nltk/stem/wordnet.html Accessed on 31 January 2022.

After completing the above steps, we extracted the latent topics from the corpus. These latent topics do not depend on the user input but rather the documents inside the dataset. In any case, our goal here was not to use the topics extracted by LDA, but to work with the terms inside these topics. So we combined the most relevant terms from each topic into a set of latent keywords. This set is called *LDA terms*.

The following steps combined the latent structure of the corpus with the concepts provided by the query.

**Co-occurrence Matrix**: After modeling *LDA terms* for each article, we computed a co-occurrence matrix between all the input terms (*term_u)* and all the LDA terms (*term_topic)* for the whole corpus. So that we could count the number of times that a *term_u* co-occurs with a *term_topic*, in the corpus.

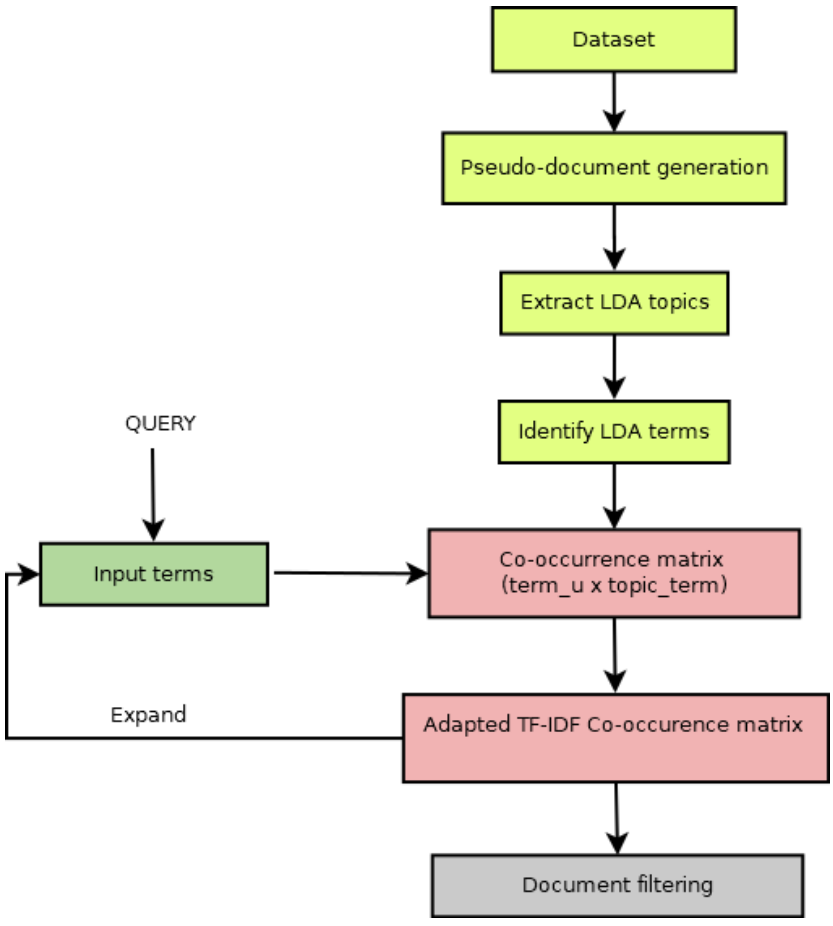



*Figure 1: Schematic representation of the method. Yellow squares represent steps required to obtain latent keywords in the dataset. The gray square represents the step that combines the latent information from the corpus with the user input terms.*

**Adapted TF-IDF co-occurrence Matrix**. TF-IDF combines two metrics used in text information processing techniques: *term frequency* (TF) and *Inverse Document Frequency* (IDF). TF is a metric used to represent the number of occurrences of a term in a document, while IDF indicates the number of times a term appears in a corpus. TF-IDF represents the importance of a term in the document. Considering important a term which appears frequently in a document but is rare in the corpus. This term can be used to represent the key information of the document.

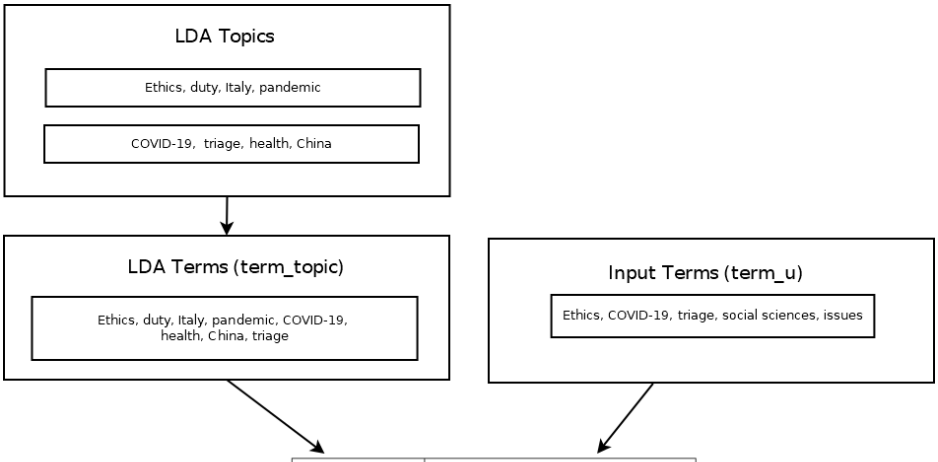
In our study, we translated the idea behind TF-IDF and used it to analyze the relation between *term_topic* and *term_u*. TF-IDF's goal is to link the information provided by topic terms and input terms. The TF-IDF score obtained by each of the topic terms would depend on the query and, therefore, on the input terms.

Firstly, we computed the topic frequency TF as the number of occurrences of an LDA term in all the LDA topics for each paper:

$$TF(term) = \sum_{i=1}^{p} S_i(term)$$

*Equation 1*

Where $S_i$(term) = 1 if the term is in LDA terms $\{term\_topic\}_i$ for the document i, and p is the total number of documents in the corpus.




| Co-occurrence Matrix (term_u x term_topic) | | term_topic | | | | |
|---|---|---|---|---|---|---|
| | | *Ethics* | *COVID-19* | *Triage* | *Social sciences* | *issues* |
| term_u | *Duty* | 1 | 5 | 6 | 1 | 0 |
| | *Italy* | 2 | 2 | 1 | 0 | 4 |
| | *pandemic* | 2 | 1 | 0 | 4 | 5 |
| | *health* | 4 | 2 | 1 | 0 | 1 |
| | *China* | 8 | 4 | 5 | 0 | 4 |
| | *Triage* | 2 | 12 | 5 | 1 | 4 |

*Figure 2: Co-occurrence matrix generation*

Then, we computed the Topic Inverse Document Frequency (TIDF) as the inverse frequency of the number of times that a term occurs in the list of term topics for each article:

$$TIDF(term) = log\left(1 + \frac{p}{TF(term)^c}\right)$$

*Equation 2*

Where c is a constant number. A greater value of c would indicate a lower TIDF score in relation to frequent terms, while smaller c values would indicate a better TIDF score in relation to frequent terms. For our dataset, we have used c=2 as the value that allowed us to extract latent terms with similar popularity to the user inputs. Readers can refer to the supplementary material section to see further analysis of the behavior of the c value. A deeper justification about the use of the adapted version of TF-IDF that we have proposed in our paper will be further explained in the section *Energy and c parameter behavior*.

We defined LDA-Term Co-occurrence (LTC) as the number of times that a *term_u* co-occurs with a *term_topic*. This value is provided by the co-occurrence matrix defined above:

$$LTC(term_{topic}, term_u) = \mathrm{C}(term_{topic}, term_u)$$

*Equation 3*

Where C is the co-occurrence matrix.

Our adapted TF-IDF metric is expressed as follow:

$$ATF.IDF(term_{topic}, term_u) = LTC(term_{topic}, term_u) \cdot TIDF(term_{topic})$$

*Equation 4*

In our method, LTC fulfills the same role as Term Frequency in TF-IDF [37]. We worked with term topics instead of documents, so the aim of LTC is to weigh terms inside of term topics. The more frequent a term is present inside a term topic, the more important this term becomes for that term topic. LTC is a simple way to measure the importance of a term within a term topic. In the same way, if a term is very frequent in each term topic, then this term is not relevant or may give us irrelevant information. Therefore, the objective of TIDF is to present a high score to less frequent terms. Finally, we combined LTC with TIDF results to get a high score from those very frequent terms in only a few terms topics which are, at the same time, very infrequently seen throughout the rest of the term topics.

Finally, we built an adapted TF-IDF co-occurrence Matrix (ATF.IDF):

$$ACO(term_{topic}, term_u) = ATF.IDF(term_u, term_{topic})\ \forall term_u \in U, \forall\ term_{topic} \in T$$

*Equation 5*

Where $T$ is a set with all LDA term topics found in the corpus and $U$ are the input terms.

**Keywords expansion**: In this step, we expanded the input terms with new topics extracted from LDA terms. From the ATF.IDF co-occurrence matrix built in the previous step; we computed the sum of the ATF.IDF scores for all the LDA terms topic as:

$$total_{score}(ACO) = \sum_{term_u \in U} \sum_{term_{topic} \in T} ACO(term_{topic}, term_u)$$

*Equation 6*

Then, we applied an energy threshold to the score. This energy threshold is a classical mechanism to preserve information until reaching a specific threshold score [38]:

$$threshold(ACO, energy) = total_{score}(ACO) \cdot energy$$

*Equation 7*

Term topics are ranked according to their ATF-IDF score so that the term topics with greater scores appears first.

$$score(term_{topic}) = \sum_{term_u \in U} ACO(term_{topic}, term_u)$$

*Equation 8*

The next step computed the cumulative sum of all term topics, according to their score, and selected only the LDA terms whose cumulative ATF-IDF score were under the threshold. Finally, these selected LDA terms were added to the input terms set.

$$selectedTerms(ACO, energy) = \{t_1, t_2, \ldots, t_n\}\{t_j \in LDA \wedge cumsum(\mathrm{score(t_j)}) < threshold(ACO, energy)$$

*Equation 9*

We can thus repeat all the steps again using the new input terms in order to find new latent keywords.

**Document filtering**. For this final step, documents were filtered according to the expanded input terms set ($\{term_u\}$). So the filters were built using the initial user input terms, and the expanded input terms derived from the latent structure. Next, we extracted all the documents from the corpus that contained the keywords. The minimum number of keywords that had to be in a document can be set as an additional threshold. In our case, we have imposed a minimum of 2 keywords.

## Time complexity

The time complexity was mainly affected by the execution of the LDA method and is $O(p \cdot (mnt + t^3))$ where *p* is the number of LDA documents in the corpus, *m* is the number of topics, *n* is the number of terms, and $t = \min(m, n)$ [39]. While this topic is not directly related to our work, there are some proposals to reduce the time complexity of LDA algorithms [39,40] that can be useful for readers.

### Experimental set-up

In this paper, we have executed the proposed method in three iterations, in order to extract three levels of keywords. After performing a hyperparameter optimization search using the Grid Search method provided by the Scikit-learn library[5], the best energy threshold for our dataset was 0.025.

The c parameter of TIDF was 2 (Please, refer to Section 4 check the supplemental material of this paper for more details about ATF.IDF).

In respect to LDA, the number of topics was set to 30, which gave us good results in terms of coherence. We used the UMASS-Coherence to determine the number of topics [41].

## 4. Results

In this section, we will proceed to show the results obtained using our proposed method. The goal of our Information Retrieval System was to find an answer to the question: "*What has been published about ethical and social science considerations?*". And, specifically, we wanted information from seven different thematic areas[6]:

| Number | Thematic area |
|---|---|
| 1 | Efforts to articulate and translate existing ethical principles and standards to salient issues in COVID-19. |
| 2 | Efforts to embed ethics across all thematic areas, engage with novel ethical issues that arise, and coordinate to minimize duplication of oversight. |
| 3 | Efforts to support sustained education, access, and capacity building in the area of ethics. |
| 4 | Efforts to establish a team at World Health Organization that will be integrated within multidisciplinary research and operational platforms and that will connect with existing and expanded global networks of social sciences. |
| 5 | Efforts to develop qualitative assessment frameworks to systematically collect information related to local barriers and enablers for the uptake and adherence to public health measures for prevention and control. This includes the rapid identification of the secondary impacts of these measures. (e.g. use of surgical masks, modification of health seeking behaviors for SRH, school closures). |
| 6 | Efforts to identify how the burden of responding to the outbreak and implementing public health measures affects the physical and psychological health of those providing care for Covid-19 patients and identify the immediate needs that must be addressed. |
| 7 | Efforts to identify the underlying drivers of fear, anxiety and stigma that fuel misinformation and rumor, particularly through social media. |

*Table 1: Thematic areas.*

[5] https://scikit-learn.org/stable/ Accessed on 31 January 2022.
[6] The questions and thematic areas referred in our work were extracted from the Kaggle Challenge

As explained in previous sections, the dataset provided by Kaggle does not contain information about the keywords of the articles, so we could only work with the title, abstract, and references of each article. The absence of keywords is a challenge when identifying the category in which an article is framed, since the title and abstract do not always contain enough information on all the topics covered by a scientific work.

In this context, evaluating the performance of our method can be a challenging problem. Fortunately, the provided dataset contains enough information to identify the articles that compose the corpus; thus we were able to perform a manual evaluation in two steps:

1. **A priori evaluation**: By checking the titles and abstracts of the selected article, it was verified whether the articles selected by our method could answer the question posed. This evaluation was performed exclusively using the data provided by the dataset.
2. **A posteriori evaluation**: By manually obtaining each one of the selected articles the articles were checked to see if they would be able to answer the question posed. This evaluation was performed by using external data to the provided corpus by the dataset.

In addition, keywords of the selected articles were extracted and compared with the topics generated with our method in order to check the level of co-occurrence.

For each thematic area, a list of keywords was chosen, as shown in Table 2. These keywords are the initial input terms that compound our queries and were chosen manually.

| Thematic area | Initial input terms |
|---|---|
| 1 | ethic, moral, fair, justice, immoral, standard |
| 2 | ethic, oversight, justice, care, sociology, education, anthropology, bibliometric, social, morale, dilemma, concerns |
| 3 | education, access, ethics, fellowship, teaching, principles, philosophy, students, training |
| 4 | WorldHealthOrganization, research, global, multidisciplinary, social, science, university, collaboration |
| 5 | local, barrier, public, measures, society, pandemic, enablers, publicHealth, prevention, control, impact, closures, quarantine |
| 6 | outbreak, publicHealth, public, measures, psychology, care, covid-19, needs, urgently, response, resilicency, pandemic, nurse, medic, employee, professional, worker |
| 7 | stigma, misinformation, rumor, socialMedia, media, news, papers, networks, fake, fakeNews, facebook, twitter |

*Table 2: Thematic areas and initial input terms.*

## A priori evaluation

During "a priori" evaluation, the top 10 most relevant papers for each thematic area were selected. We performed two surveys with the goal of checking whether the filtered papers were able to answer the question posed in the corresponding thematic area. To this end, in the first survey, the participants had to evaluate whether the title and the abstract of a set of articles were related to a thematic area. Thus, participants were only allowed to see the same information used by the method to filter and retrieve the selected papers. Participants were randomly assigned to two thematic areas and had to read the title and abstract of ten articles retrieved by our method for the assigned thematic areas.

| **Expansion** (#iteration) | **Input terms** | **Documents** |
|---|---|---|
| 1 | Ethic, Moral, Fair, Justice, Immoral, Standard | 260 |
| 2 | Ethic, Moral, Fair, Justice, Immoral, Standard, **Consent, Bioethic, Duti, Principi** | 306 |
| 3 | Ethic, Moral, Fair, Justice, Immoral, Standard, Consent, Bioethic, Duti, Principi, **Oblig, Alloc** | 319 |

*Table 3: Example of filtering documents and input terms expansion after three expansions. Bold terms are introduced from LDA terms during the previous expansion. Documents column indicates the number of filtered documents. Note that terms are stemmed.*

There were three possible answers:

1. The article fits the thematic area.
2. The article does not fit the thematic area.
3. With the provided information, we cannot know whether the article fits the thematic area or not.

Answers 1 and 2 helped to evaluate an article as a True Positive and a False Positive, respectively.

Table 4 shows the results obtained in the survey per thematic area. The range of different responses given by the participants can be seen in Figure 3.

The survey was answered by 57 participants, all of them use Amazon Mechanical Turk[7], a web platform where users get an economic reward for doing tasks that require human intelligence. All participants were native English speakers. The time to complete the survey and response patterns were analyzed in order to prevent random responses.

[7] https://www.mturk.com/ Retrieved June 3, 2020.

Table 4 shows the responses obtained in the survey. For all the thematic areas, the first answer (The article fits the thematic area) was the most common, being selected for the majority of articles. According to our results, the second thematic area ("Efforts to embed ethics across all thematic areas, engage with novel ethical issues that arise, and coordinate to minimize duplication of oversight.") obtained the best results, with about 81% of users answering with the first answer while the last thematic area ("Efforts to identify the underlying drivers of fear, anxiety, and stigma that fuel misinformation and rumors, particularly through social media.") achieved the worst results, with a score of only 67.144% for the first answer.

| Thematic area | Response 1 (%) | | Response 2 (%) | | Response 3 (%) | |
|---|---|---|---|---|---|---|
| | Mean | STD.D | Mean | STD.D | Mean | STD.D |
| 1 | 79 | 15.23 | 16 | 11.73 | 4 | 5.16 |
| 2 | 81.03 | 11.79 | 13.42 | 10.29 | 5.56 | 5.85 |
| 3 | 73.59 | 11.72 | 19.42 | 12.24 | 6.15 | 4.86 |
| 4 | 73.031 | 10.96 | 21.113 | 9.20 | 5.832 | 5.62 |
| 5 | 76.67 | 9.72 | 15.55 | 10.73 | 7.77 | 5.36 |
| 6 | 77.141 | 13.08 | 15.717 | 12.51 | 7.14 | 12.14 |
| 7 | 67.144 | 9.03 | 21.428 | 9.52 | 11.429 | 4.99 |

*Table 4: Responses per thematic area (See Table 1 for information about each thematic area). Response 1 = The article fit the thematic area. 2 = The article does not fit the thematic area. 3 = With the provided information, it is not possible to determine whether the article fits the thematic area or not.*

Figure 3 shows the answer interval for every thematic. That means that, for example, in the first thematic area there was an article that was classified as a True positive by 50% of the participants. At the same time, there was an article that was classified as a True positive by 100% of the participants assigned to that thematic area. Additionally, in the supplemental material section, we have included a Figure where readers can get additional information about the precision obtained in the survey results.

In the second survey, we randomly divided the same papers from the first survey into five groups, ensuring that each group had two papers per subject area (14 papers per group). Then, we arbitrarily assigned each participant to a single group.

Participants then had to read the abstract and the title of each paper in their assigned group and assign a thematic area. They also had the option of not assigning any thematic area if they felt unable to classify the paper.

For the second survey, we had 65 participants from Amazon Mechanical Turk and all participants were native English speakers. As in the previous survey, we analyzed the time and response patterns to prevent random responses. Results from the second survey are presented in the confusion matrix shown in Table 5. From this confusion matrix, some metrics have been extracted and can be found in Table 6. As we can see, precision is, again, obtaining high scores. At this point, it is important to state that the second survey required participants to perform a harder task than in the first survey, so lower scores were expected. The recall also had high scores except for the third thematic area, where results were < 0.5. Also, accuracy and F-scores were over 0.5 points for each thematic area.

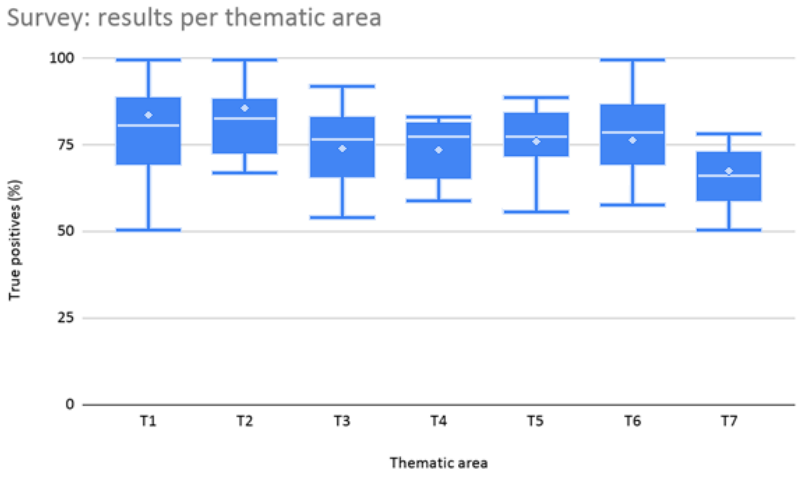



*Figure 3: Results per thematic area. We can see the range of true positive values obtained by our method according to the different responses given by the participants in the survey.*

Precision, recall, accuracy, True/Negative rates, and F-scores were calculated by using definitions from [42].

| Thematic area | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| **1** | 8 | 1 | 0 | 1 | 0 | 1 | 0 |
| **2** | 2 | 8 | 1 | 0 | 2 | 0 | 0 |
| **3** | 0 | 1 | 7 | 1 | 1 | 0 | 0 |
| **4** | 0 | 0 | 1 | 8 | 2 | 1 | 1 |
| **5** | 0 | 0 | 0 | 0 | 5 | 0 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 5 | 3 |
| **8** | 0 | 0 | 1 | 0 | 0 | 3 | 6 |

*Table 5: Confusion matrix representing results from the second survey. Rows represent the participants' responses, while columns represent the estimated classification performed by our method.*

| TA | TP | FP | FN | Precision | Recall | TN | TNR | Accuracy | F-score |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 8 | 2 | 3 | 0,73 | 0,8 | 39 | 0,95 | 0,90 | 0,76 |
| **2** | 8 | 2 | 5 | 0,62 | 0,8 | 39 | 0,90 | 0,87 | 0,70 |
| **3** | 7 | 3 | 3 | 0,70 | 0,7 | 40 | 0,93 | 0,89 | 0,70 |
| **4** | 8 | 2 | 5 | 0,62 | 0,8 | 39 | 0,95 | 0,87 | 0,70 |
| **5** | 5 | 5 | 0 | 1,00 | 0,5 | 42 | 0,89 | 0,90 | 0,67 |
| **6** | 5 | 5 | 3 | 0,63 | 0,5 | 42 | 0,89 | 0,85 | 0,56 |
| **7** | 6 | 4 | 4 | 0,60 | 0,6 | 41 | 0,91 | 0,85 | 0,60 |

*Table 6: Metrics extracted from the second survey. TA= Thematic Area. TP = True Positives. FP = False Positives. FN = False Negatives. TN = True Negatives. TNR = True/Negative Rate.*

Despite the results obtained, “a priori” evaluation has several limitations. Firstly, it is still difficult, even for humans, to decide what the content of an article is by only checking its title and abstract. Secondly, while there are documents with very long abstracts, other articles have short or even non-existent abstracts. Additionally, the dataset did not exclusively contain scientific papers, and it was possible to encounter editorial articles and other documents whose summaries were not homologated with abstracts.

In any case, “a priori” analysis can be very useful in evaluating results obtained when working with the same conditions as used in our method.

## A posteriori evaluation

To perform the “a posteriori” evaluation, we downloaded and manually classified 150 papers by using the following guidelines:

1. Check if the title, abstract, and references contained terms or topics related to the thematic area.
2. Check if keywords matched topics from the thematic area.
3. Check if conclusions contained keywords or topics related to the thematic area.
4. Read the full article if the previous steps did not provide enough evidence to make a decision.

After this manual classification, from the subset of 150 previously classified papers, we randomly extracted ten papers per thematic area and tested the performance of our model in terms of precision, accuracy, recall, and F-scores. This type of evaluation allowed us to more precisely know whether an article was appropriately selected because we could access information from the whole document (title, authors, keywords, abstract, references, and text). "A posteriori" evaluation was then performed by the authors. “A posteriori” results are described in Table 7 and Table 8.

If we compare both evaluation strategies, “a priori” (second survey) and “a posteriori”, as noted, the “a posteriori” evaluation gets better scores for nearly all thematic areas. In T5, both evaluations scored a 100% in terms of precision.

The lack of information during the “a priori” evaluation (first survey) was clearly reflected during the second step of the evaluation process. For example, for the first thematic area, one of the articles retrieved by the method was titled: “Understanding perceptions of global healthcare experiences on provider values and practices in the USA: a qualitative study among global health physicians and program directors” [43].

| Thematic area | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| **1** | 7 | 1 | 3 | 1 | 2 | 0 | 0 |
| **2** | 2 | 8 | 1 | 0 | 0 | 0 | 0 |
| **3** | 0 | 1 | 6 | 1 | 0 | 1 | 0 |
| **4** | 1 | 0 | 0 | 8 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 8 | 0 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 7 | 2 |
| **8** | 0 | 0 | 0 | 0 | 0 | 2 | 8 |

*Table 7: Confusion matrix presenting results from “a posteriori” results. Rows represent the real classification, while columns represent the estimated classification performed by our method.*

| TA | TP | FP | FN | Precision | Recall | TN | TNR | Accuracy | F-score |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 7 | 3 | 7 | 0,50 | 0,7 | 45 | 0,94 | 0,84 | 0,58 |
| **2** | 8 | 2 | 3 | 0,73 | 0,8 | 44 | 0,96 | 0,91 | 0,76 |
| **3** | 6 | 4 | 3 | 0,67 | 0,6 | 46 | 0,92 | 0,88 | 0,63 |
| **4** | 8 | 2 | 1 | 0,89 | 0,8 | 44 | 0,96 | 0,94 | 0,84 |
| **5** | 8 | 2 | 0 | 1,00 | 0,8 | 44 | 0,96 | 0,96 | 0,89 |
| **6** | 7 | 3 | 2 | 0,78 | 0,7 | 45 | 0,94 | 0,91 | 0,74 |
| **7** | 8 | 2 | 2 | 0,80 | 0,8 | 44 | 0,96 | 0,93 | 0,80 |

*Table 8: Metrics extracted from the “a posteriori” evaluation. TA= Thematic Area. TP = True Positives. FP = False Positives. FN = False Negatives. TN = True Negatives. TNR = True/Negative Rate.*

The title and the abstract of the article did not contain words explicitly related to the thematic area like “ethics”, “principles”, and “standards” so that it was more difficult to evaluate whether the article fit the topic or not. In addition, the article was not about COVID-19 disease or any other pandemic. Nevertheless, by carefully checking the article, it was clear that the article was a True Positive. This article achieved the worst results in its thematic area, where only 50% of the participants believed that the article fit the topic, while 40% of participants believed that the article did not fit the topic. Meanwhile, our method retrieved this article as one of the top 10 for the thematic area.

From these results, we can say that it seems that using certain methods to extract the latent structure of a document can help us establish relationships between words that are not necessarily evident to the human mind. Though this is not the scope of our work, further analysis should be done.

On the other hand, there are two thematic areas where “a priori” evaluation overcame “a posteriori” results. For example, in T4, our method retrieved an article titled “*The Highest Cited Papers on Brucellosis: Identification Using Two Databases and Review of the Papers' Major Findings*” [44]. This article is a bibliometric analysis but is not connected with key concepts to this thematic area, like “Web Health Organization” or global network of social sciences, and yet most participants answered that the article fit the topic.

### Energy and c parameter behavior

As mentioned before, energy and c were parameters used in Equation 7, Equation 9 (energy), and Equation 2 (c). We performed an “a posteriori” evaluation strategy to study the behavior of energy and the c parameter. According to our tests, the best performance was achieved when c was 2 and energy was 0.02. Nevertheless, is important to note that every thematic area had a different score. For example, for thematic areas 2 and 3, c=3 gave a better performance, while energy = 0.25 was the best option for thematic area 2. Figure 4 and Figure 5 illustrate the behavior of c and energy parameters for each thematic area, while Figure 6 shows an average F-score across all thematic areas.

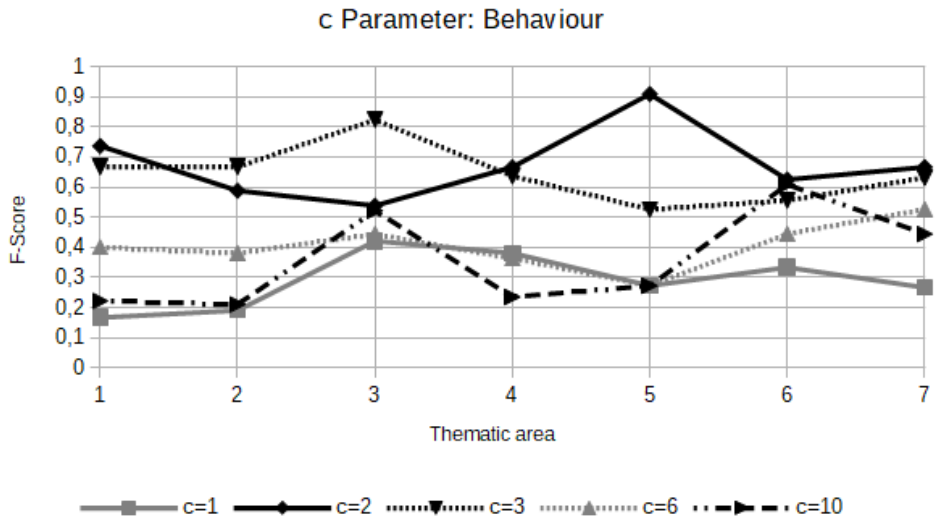



*Figure 4: Behavior of the c parameter. Energy is fixed at 0.025.*

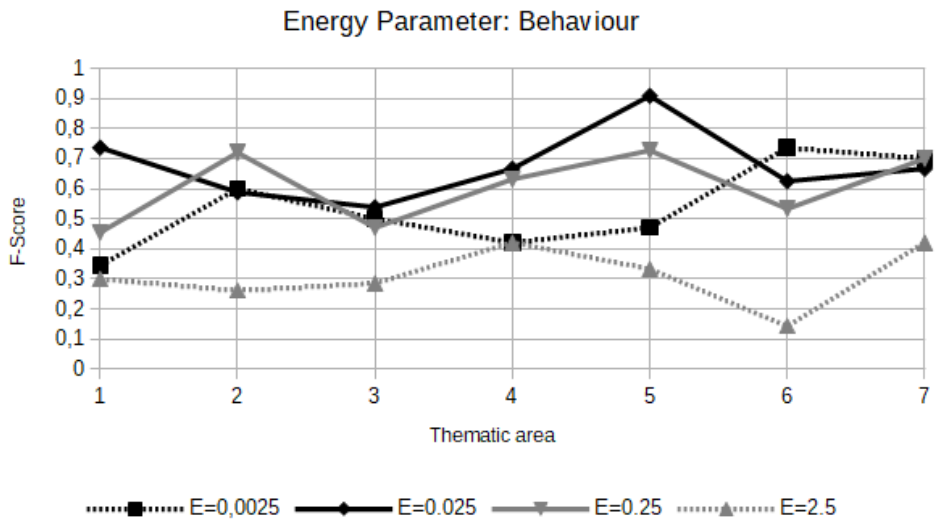



*Figure 5: Behavior of the energy parameter. c is fixed at 2.*

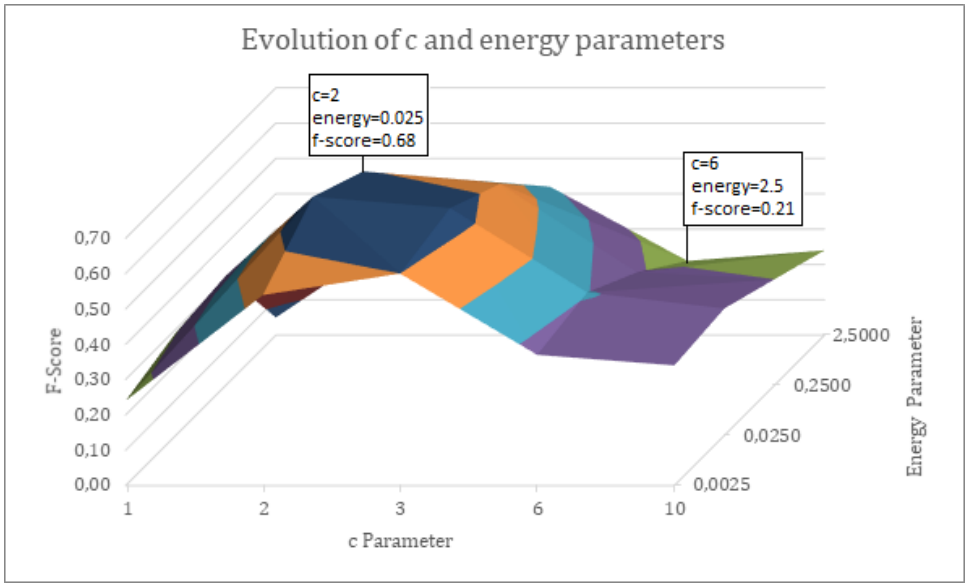



*Figure 6: A surface plot representing F-score evolution according to c and energy parameters. The F-score is measured as the mean F-score for whole thematic areas.*

## The role of ATF.IDF

In Section 3, we introduced ATF.IDF, and we adapted the TF-IDF proposal. Figure 7 shows a comparison of the keyword expansion stage according to the value of c in terms of *popularity.* While Figure 7 groups terms by their stage in the method, Figure 8 shows the same information

but clusters terms by popularity. The popularity of a term is the number of articles in the corpus where the term appears. For example, if term A appears in 2,500 articles and term B appears in 12 documents, the popularity of term A (popularity = 2,500) is greater than the popularity of term B (popularity = 12). At the same time, we can say that term B is a more specific term than term A.

In both figures, we analyzed the input terms of thematic area number 5 (See Table 2) If we use c = 1, the input terms are expanded with the LDA topics described below:

- First iteration: “health” and “diseas”, “influenza”, “outbreak”.
- Second iteration: “viru”, “infect”, “respiratori”, “epiderm”.

Finally, in Figure 8, elements represented with a grey mark are terms that had relations with the query but were not selected.

When using c = 2, expanded terms were represented by a diamond in Figure 7 for terms extracted during the first iteration and with a star for the second iteration.

In the corpus, “health” and “outbreak” were examples of very popular terms which appeared throughout the corpus. For that reason, their frequency score was very high. We can observe how the second iteration takes very generalist terms into account. At the same time, expanded keywords are very far from the initial input terms.

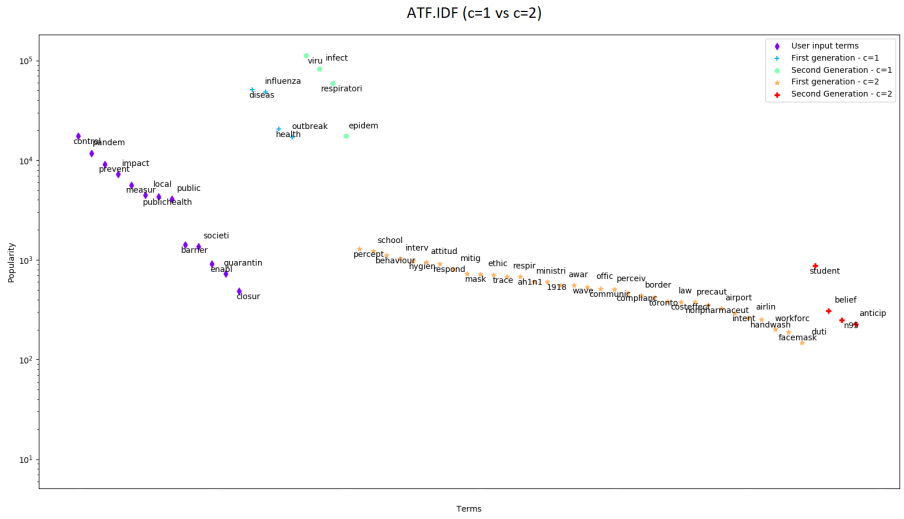



*Figure 7: Terms by popularity. Thematic area #5. They are sorted by stage and popularity.*

For that reason, it is relevant to give a higher preponderance to specific terms by setting a proper value to c. This value will depend on the corpus structure and their word distributions in the corpus where a small number of concepts tend to be present in almost every document, greater values of c will be needed. Nevertheless, setting and a high value of c can lead to ATF.IDF values that are very close to zero, resulting in a random selection.

In our corpus, c=2 was providing a good preponderance for specific terms, and during the first iteration, the expanded terms were very close to the initial query. In addition, the second iteration expanded the query to two terms that were even more specific than the user input terms. This behavior was the opposite when c=1, where the terms in the second generation were less specific than the first generation.

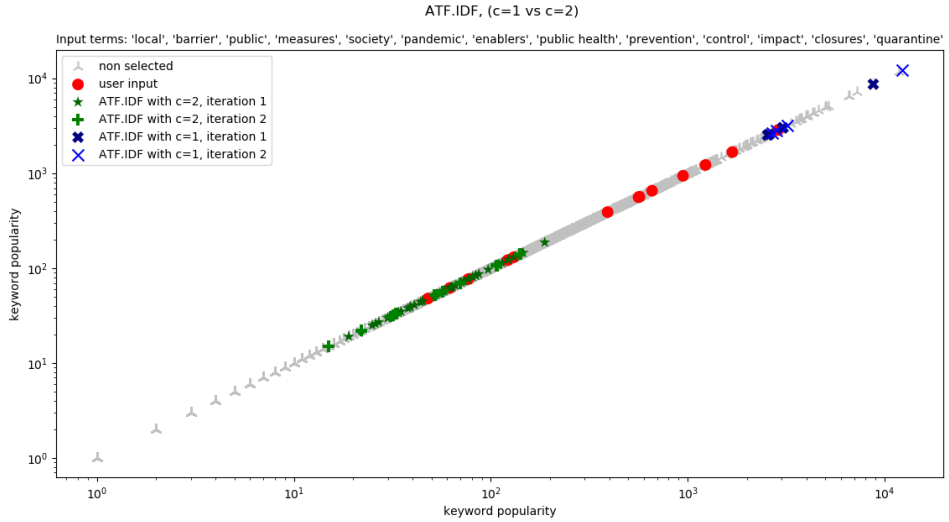



*Figure 8: Comparison of ATF,IDF selected items according to c value. c=1 vs c=2. Thematic area #5. Sorted by popularity.*

As we can observe in Figure 7, when c=2, more terms were selected during the first iteration, and they were very similar in terms of popularity. When c=1, fewer terms were selected during the first iteration, and popularity dispersion increased.

Initial query terms had very different popularity, but only the two most popular terms appeared in more than $10^3$ documents. At the same time, all the terms selected for c=1 were very popular, with all of them appearing in more than $10^4$ documents. This situation contrasted with terms selected by ATF.IDF when c=2, where the terms popularity was in the interval between $10^2$ and $10^3$ documents.

To see results for other thematic areas, please refer to the supplemental material of this article.

## 5. **Conclusions**

In the current context of the COVID-19 pandemic, a massive influx of scientific knowledge is being produced. This amount of information makes it difficult for researchers to search for useful information.

Keywords are a very important source when identifying articles that fall within a specific category. In our method, we used LDA to generate keywords that could identify latent concepts existing in a scientific article. The co-occurrence analysis between input terms and LDA terms allowed us to retrieve articles that were close to a series of keywords related to the subject of study. ATF-IDF played a key role in the co-occurrence analysis, acting as a selection function and extracting query-related specific concepts while filtering generic ones.

Two different validation strategies were carried out: “a priori” and “a posteriori” evaluations. Both of them were performed by humans, and all in all, results were satisfactory. Therefore, our method can be a useful tool for identifying articles on very specific topics that might occupy less popular places within a corpus.

“A posteriori” evaluation, specifically, improved the performance scores obtained by “a priori” evaluation by identifying relationships between concepts that were not as evident for humans as for latent models.

Our method can be a valuable tool when used as an Information Retrieval System, with the capacity to focus on retrieving specific information about complex topics in a dataset where the thematic area has a secondary role. Nevertheless, we have performed our analysis working with the corpus provide by the White House in collaboration with the Allen Institute, where only abstracts and titles were presented in the corpus so that the full semantic information from papers have not been analyzed. This limitation should be taken into consideration.

For future works, it is necessary to analyze the behavior of this method in different thematic areas as well as to test our work with a full semantic information corpus.

**Declaring of Conflict of Interest**
The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

**ORCID iDs**

- Jorge Chamorro-Padial: 0000-0002-6334-3786
- Francisco-Javier Rodrigo-Ginés: 0000-0001-6235-6860
- Rosa Rodríguez-Sánchez: 0000-0001-7886-9329

**References**

1. Hui DS, I Azhar E, Madani TA, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases* 2020; 91: 264–266.

2. World Health Organization. WHO Director-General’s opening remarks at the media

briefing on COVID-19 - 6 March 2020, https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---6-march-2020 (2020, accessed 26 May 2020).

3. Chahrour M, Assi S, Bejjani M, et al. A Bibliometric Analysis of COVID-19 Research Activity: A Call for Increased Output. *Cureus*; 12. Epub ahead of print 22 March 2020. DOI: 10.7759/cureus.7357.

4. Atkson AG. *What Will be the Economic Impact of COVID-19 in the US? Rough Estimates of Disease Scenarios*. Los Angeles. Epub ahead of print 2020. DOI: doi.org/10.21034/sr.595.

5. Li S, Wang Y, Xue J, et al. The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users. *International Journal of Environmental Research and Public Health* 2020; 17: 2032.

6. Liu Q, Zheng Z, Zheng J, et al. Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: A Digital Topic Modeling Approach (Preprint). *Journal of Medical Internet Research*; 22. Epub ahead of print 4 April 2020. DOI: 10.2196/19118.

7. Huynh TLD. The COVID-19 risk perception: A survey on socioeconomics and media attention. *Economics Bulletin* 2020; 40: 758–764.

8. Torres-Salinas D. Ritmo de crecimiento diario de la producción científica sobre Covid-19. Análisis en bases de datos y repositorios en acceso abierto. *El Profesional de la Información*; 29. Epub ahead of print 14 April 2020. DOI: 10.3145/epi.2020.mar.15.

9. COVID-19 Open Research Dataset Challenge (CORD-19) | Kaggle, https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge (accessed 30 January 2022).

10. Lou J, Tian SJ, Niu SM, et al. Coronavirus disease 2019: A bibliometric analysis and review. *European Review for Medical and Pharmacological Sciences* 2020; 24: 3411–3421.

11. Nasab FR, rahim F. Bibliometric Analysis of Global Scientific Research on SARSCoV-2 (COVID-19). *medRxiv* 2020; 2020.03.19.20038752.

12. Bonnevie E, Gallegos-Jeffrey A, Goldbarg J, et al. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *Journal of Communication in Healthcare* 2021; 14: 12–19.

13. Baraybar-Fernández A, Arrufat-Martín S, Rubira-García R. Public Information, Traditional Media and Social Networks during the COVID-19 Crisis in Spain. *Sustainability* 2021; 13: 6534.

14. Feng Y, Zhou W. Is Working From Home The New Norm? An Observational Study Based on a Large Geo-tagged COVID-19 Twitter Dataset, http://arxiv.org/abs/2006.08581 (2020, accessed 28 December 2021).

15. Sharifi A, Khavarian-Garmsir AR. The COVID-19 pandemic: Impacts on cities and major lessons for urban planning, design, and management. *Science of the Total Environment* 2020; 749: 142391.

16. Cinelli M, Quattrociocchi W, Galeazzi A, et al. The COVID-19 Social Media Infodemic, http://arxiv.org/abs/2003.05004 (2020, accessed 26 May 2020).

17. Lopez CE, Vasu M, Gallemore C. Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset, http://arxiv.org/abs/2003.10359 (2020, accessed 26 May 2020).

18. Singh L, Bansal S, Bode L, et al. A first look at COVID-19 information and misinformation sharing on Twitter, http://arxiv.org/abs/2003.13907 (2020, accessed 26 May 2020).

19. Schild L, Ling C, Blackburn J, et al. 'Go eat a bat, Chang!': An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19, http://arxiv.org/abs/2004.04046 (2020, accessed 26 May 2020).

20. Riloff E, Schafer C, Yarowsky D. Inducing information extraction systems for new languages via cross-language projection. 2002; 1–7.

21. Latif S, Usman M, Manzoor S, et al. Leveraging Data Science To Combat COVID-19: A Comprehensive Review. Epub ahead of print 30 April 2020. DOI: 10.36227/TECHRXIV.12212516.V1.

22. Shah PK, Perez-Iratxeta C, Bork P, et al. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics* 2003; 4: 20.

23. Williams R. *Keywords: A Vocabulary of Culture and Society*. 2nd ed. New York, NY, USA: Oxford University Press, 1985.

24. Baker P. Querying Keywords. *Journal of English Linguistics* 2004; 32: 346–359.

25. Wang X, Liu W, Chauhan A, et al. Automatic Textual Evidence Mining in COVID-19 Literature, http://arxiv.org/abs/2004.12563 (2020, accessed 28 December 2021).

26. Esteva A, Kale A, Paulus R, et al. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *npj Digital Medicine* 2021; 4: 1–9.

27. Voorhees E, Alam T, Bedrick S, et al. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection, https://lhncbc.nlm.nih.gov/LHC-publications/pubs/TRECCOVIDConstructingaPandemicInformationRetrievalTestCollection.html (2020, accessed 28 December 2021).

28. Best P, Taylor B, Manktelow R, et al. Systematically retrieving research in the digital age: Case study on the topic of social networking sites and young people's mental health. *Journal of Information Science* 2014; 40: 346–356.

29. Karlsson A, Hammarfelt B, Steinhauer HJ, et al. Modeling uncertainty in bibliometrics and information retrieval: an information fusion approach. *Scientometrics* 2015; 102: 2255–2274.

30. Dimitrakis E, Sgontzos K, Tzitzikas Y. A survey on question answering systems over linked data and documents. *Journal of Intelligent Information Systems* 2020; 55: 233–259.

31. Mohammed SH, Al-Augby S. LSA & LDA topic modeling classification: Comparison

study on E-books. *Indonesian Journal of Electrical Engineering and Computer Science* 2020; 19: 353–362.

32. Zuo Y, Wu J, Zhang H, et al. Topic modeling of short texts: A pseudo-document view. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, pp. 2105–2114.

33. Chen G, Xiao L. Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. *Journal of Informetrics* 2016; 10: 212–223.

34. Weinberg BH. Bibliographic Coupling: A Review. *Information Storage and Retrieval*, https://eric.ed.gov/?id=EJ101204 (1974, accessed 25 March 2019).

35. Garfield E. KeyWords Plus - ISI's Breakthrough Retrieval Method. 1. Expanding Your Searching Power on Current-Contents on Diskette. *Current Contents* 1990; 1: 5–9.

36. Ganesan K, Lloyd S, Sarkar V. Discovering Related Clinical Concepts Using Large Amounts of Clinical Notes. *Biomedical Engineering and Computational Biology* 2016; 7s2: BECB.S36155.

37. Roelleke T, Wang J. TF-IDF uncovered: A study of theories and probabilities. In: *ACM SIGIR 2008 - 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings*. New York, New York, USA: ACM Press, pp. 435–442.

38. Singla A, Patra S. A fast automatic optimal threshold selection technique for image segmentation. *Signal, Image and Video Processing* 2017; 11: 243–250.

39. Cai D, He X, Han J. Training linear discriminant analysis in linear time. In: *Proceedings - International Conference on Data Engineering*. 2008, pp. 209–217.

40. Sontag D, Roy DM. *Complexity of Inference in Latent Dirichlet Allocation*. 2011.

41. Mimno D, Wallach H, Talley E, et al. Optimizing Semantic Coherence in Topic Models. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, pp. 262–272.

42. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, http://arxiv.org/abs/2010.16061 (2020, accessed 30 January 2022).

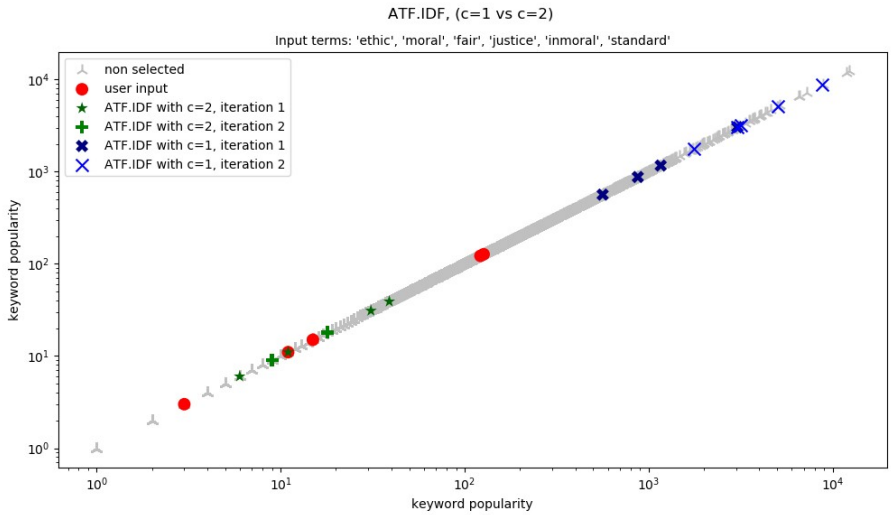
43. Matthews-Trigg N, Citrin D, Halliday S, et al. Understanding perceptions of global healthcare experiences on provider values and practices in the USA: A qualitative study among global health physicians and program directors. *BMJ Open* 2019; 9: e026020.

44. Bakri FG, Alqadiri HM, Adwan MH. The Highest Cited Papers in Brucellosis: Identification Using Two Databases and Review of the Papers' Major Findings. *BioMed Research International*; 2018. Epub ahead of print 2018. DOI: 10.1155/2018/9291326.

# Finding answers to COVID-19 specific questions: An Information Retrieval System based on latent keywords and adapted TF-IDF

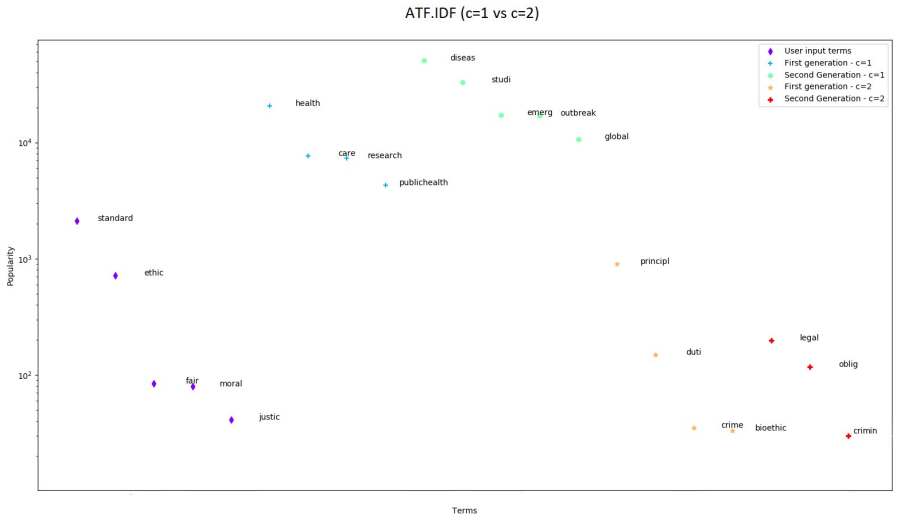
Jorge Chamorro-Padial[1]
Francisco-Javier Rodrigo-Ginés[23]
Rosa Rodríguez-Sánchez[4]

## 1. Supplemental material

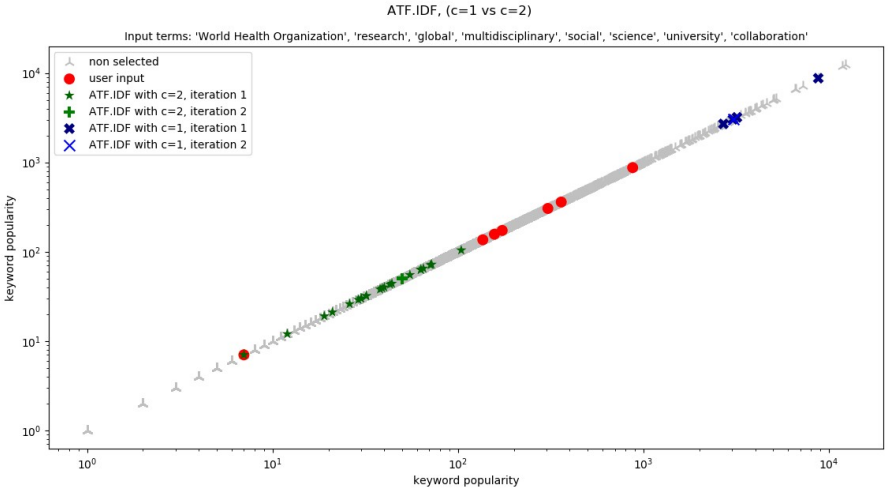



- **Title**: Comparison of ATF.IDF selected items according to c value. c=1 vs c=2. Thematic area #1. Sorted by popularity.
- **Description**: All selected terms when c=1 are more popular than the input terms. When c=2, we can observe how all the selected terms are in consonance with the input terms.

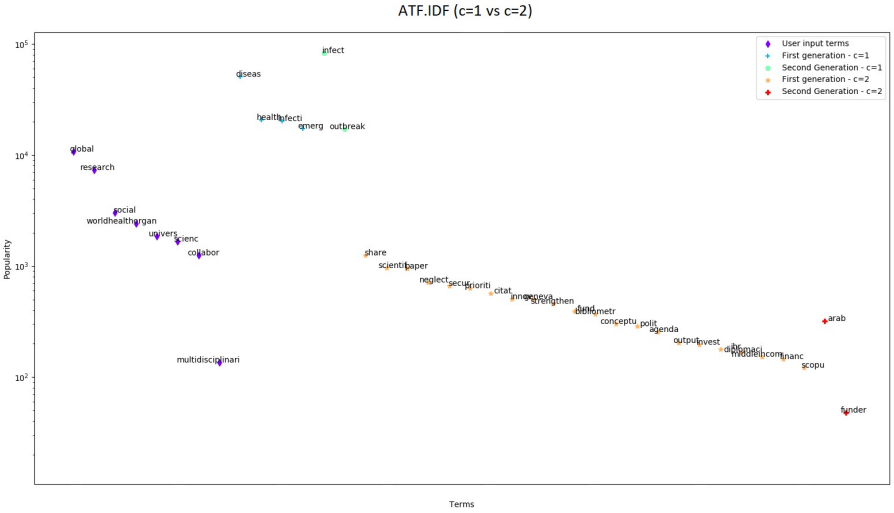
[1] CITIC-UGR, Universidad de Granada, 18071 Granada, Spain.
[2] Universidad Nacional de Educación a Distancia, Spain.
[3] Corresponding autor.
[4] Departamento de Ciencias de la Computación e I.A, CITIC-UGR, Universidad de Granada, 18081 Granada, Spain



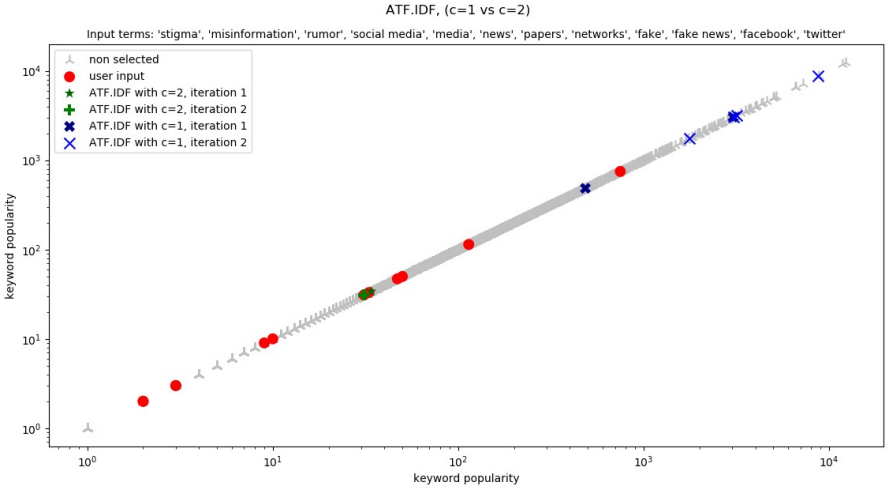

- **Title**: Comparison of ATF.IDF selected items according to c value. c=1 vs c=2. Thematic area #1. Sorted by stage and popularity.
- **Description**: All selected terms when c=1 are more popular than the input terms. When c=2, we can observe how all the selected terms are in consonance with the input terms.



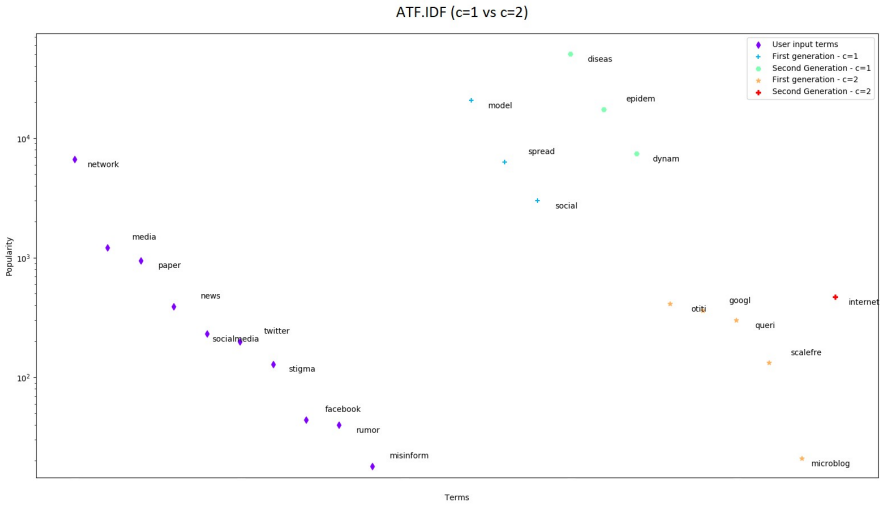

- **Title**: Comparison of ATF.IDF selected items according to c value. c=1 vs c=2. Thematic area #4. Sorted by popularity.
- **Description**: In this thematic area, we can see clearly how selected terms are very popular when c=1 and very specific in for c=2. Distance between selected terms and input terms are closer when c=2.



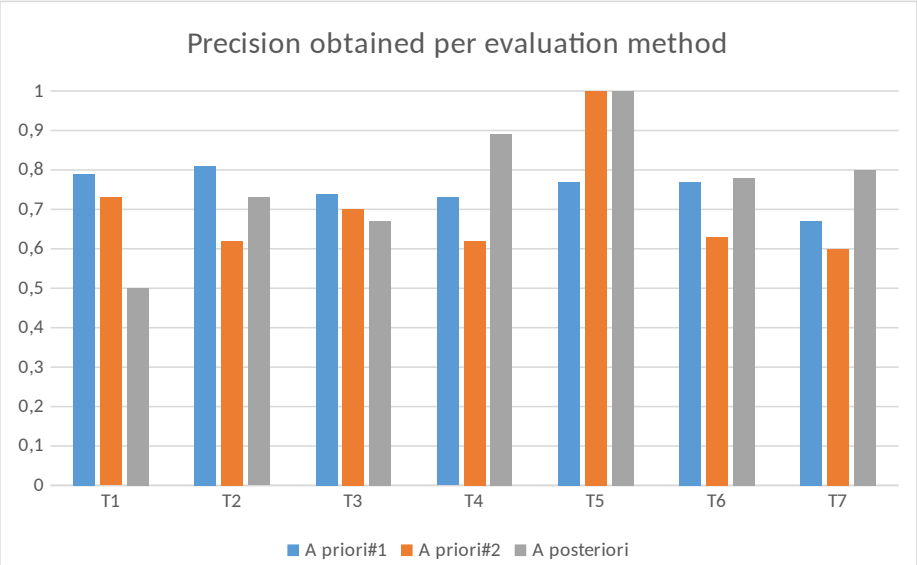

- **Title**: Comparison of ATF.IDF selected items according to c value. c=1 vs c=2. Thematic area #4. Sorted by stage and popularity.
- **Description**: In this thematic area, we can see clearly how selected terms are very popular when c=1 and very specific in for c=2. Distance between selected terms and input terms are closer when c=2.




- **Title**: Comparison of ATF.IDF selected items according to c value. c=1 vs c=2. Thematic area #7. Sorted by popularity.
- **Description**: As we can see in the chart, the popularity of query terms is unequal, having very popular terms as well as very specific terms. Nevertheless, for c=1 selected terms are more popular. The distance between the query and the input terms are closer when c=1.




- **Title**: Comparison of ATF.IDF selected items according to c value. c=1 vs c=2. Thematic area #7. Sorted by popularity. Sorted by stage and popularity.
- **Description**: As we can see in the chart, the popularity of query terms is unequal, having very popular terms as well as very specific terms. Nevertheless, for c=1 selected terms are more popular. The distance between the query and the input terms are closer when c=1.




- **Title**: Precision scores comparison. A priori#1 makes reference to the first survey. A priori#2 makes reference to the second survey. Results from A priori#1 are the average responses received in the first survey. T1, T2… T7 is the number of thematic area.

- **Description**: The three evaluation methods performs with scores > 0.5 on every thematic area. “A posteriori” evaluation gets better scores for almost all thematic areas, except the fifth and the sixth ones. In T5, “A posteriori” evaluation scored a 100% on precision.

## 4.3. Attention –Survival Score: A Metric to Choose Better Keywords and Improve Visibility of Information

### 4.3.1. Datos generales

1. **Autores**: Jorge Chamorro-Padial, Rosa Rodríguez-Sánchez.
2. **Revista**: Algorithms.
3. **Datos sobre la publicación**:
   - **Referencia**: Chamorro-Padial y Rodríguez-Sánchez (2023a)
   - **Año**: 2023.
   - **Editorial**: MDPI.
   - **DOI**: https://doi.org/10.3390/a16040196
4. **Estado**: Publicado.
5. **Métricas**:
   - **Ranking**:
     - *Emerging Sources Citation Index (ESCI)* [9]:
       - *Computer Science, Artificial Intelligence*: Q3 - 121/190 (año 2021).
       - *Computer Science, Theroy & Methods*: Q3 - 81/143 (año 2021).

### 4.3.2. Contribuciones principales

En este artículo, analizamos el rol que la Comunidad Científica otorga a las palabras clave en base a dos factores diferentes: Atención y Supervivencia. Nuestro enfoque permite valorar la importancia de una palabra clave basándonos en criterios tanto sociales como de comportamiento actual de las personas sobre un concepto. De esta manera, proponemos una métrica que permite asistir a los autores a la hora de seleccionar las palabras más adecuadas para etiquetar sus trabajos que, además, no tienen por qué ser documentos en lenguaje escrito, como es el caso de otras métricas populares tales como TF-IDF.

[9] A fecha de depósito de esta tesis, aún no se disponen de datos del año 2022.

### 4.3.3. Resumen

En nuestro trabajo, proponemos una métrica para evaluar la idoneidad de una palabra clave en base a dos criterios:

1. Atención: La atención hace referencia a la popularidad de una palabra clave en un motor de búsqueda. A mayor nivel de atención, más usuarios están interesados en conseguir información relacionada con ese término. Traslado al campo de la literatura científica: a mayor atención, más lectores potenciales puede tener un artículo.

2. Supervivencia: La supervivencia, por su parte, hace mención a la cantidad de documentos que están etiquetados con un determinado término y, por tanto, compiten entre sí por la atención de un lector. A mayor cantidad de artículos en un término, menos probabilidad de que un artículo *sobreviva* al proceso de búsqueda de información.

Nuestra propuesta es usar estos dos conceptos, atención y supervivencia, con el fin de obtener una métrica denominada *Attention-Survival score* que se expresa de la siguiente manera:

$$AS_{\cup}(K) = \alpha \cdot S_{\cup}(K) + (1 - \alpha) \cdot A_{\cup}(K)$$

Donde:

- K es el conjunto de palabras clave que se desea evaluar.
- $\alpha$ es un valor de ponderación entre atención y supervivencia. $\alpha \in [0, 1], \alpha \in \mathbb{R}$.
- $S_{\cup}$ Es el valor de supervivencia de K.
- $A_{\cup}$ Es el valor de atención de K.

En primer lugar, realizamos un análisis teórico sobre el comportamiento de las palabras clave y proponemos un algoritmo para refinar términos. Esto es, dado un término, y una ontología, sugerir al usuario una palabra clave cercana a la inicial pero que mejore su puntuación en base a *Attention-Survival*. A continuación, realizamos diferentes pruebas experimentales donde estudiamos la estructura de dos ontologías diferentes: *WordNet* (Princeton University, 2010) y *The Computer Science Ontology* (CSO) (Salatino et al., 2018). También realizamos una validación de nuestro algoritmo de refinamiento mediante una encuesta dirigida a personas con conocimientos en el ámbito de las ciencias de la computación. En esta validación de nuestro algoritmo, además de las dos ontologías mencionadas anteriormente, utilizamos DBpedia (https://www.dbpedia.org/ Accedida el 2 de mayo del 2023)).

El conocimiento ontológico juega un papel clave en nuestro algoritmo de refinamiento, que impacta directamente sobre la calidad de los resultados ofrecidos. Una ontología demasiado genérica normalmente va a provocar que el algoritmo sugiera términos muy alejados del original, por lo que resulta conveniente utilizar ontologías específicas relacionadas con el dominio en el que se está trabajando.

# Attention–Survival score: A metric to choose better keywords and improve visibility of information

**Jorge Chamorro-Padial[1],[2] and Rosa Rodríguez-Sánchez[3]**

[1]CITIC-UGR. Universidad de Granada. 18071 Granada, Spain. ORCID: 0000-0002-6334-3786. jorgechp@correo.ugr.es

[2]Corresponding author.

[3]Departamento de Ciencias de la Computación e IA. CITIC-UGR. Universidad de Granada. 18071 Granada, Spain. ORCID: 0000-0001-7886-9329. rosa@decsai.ugr.es

*Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials)

**Abstract:** In this paper, we propose a method to aid authors in choosing alternative keywords that help their papers gain visibility. These alternative keywords must have a certain level of popularity in the scientific community and, simultaneously, be keywords with fewer competitors. The competitors are derived from other papers containing the same keywords. Having fewer competitors would allow an author's paper to have a higher consult frequency. In order to recommend keywords, we must first determine an attention–survival score. The attention score is obtained using the popularity of a keyword. The survival score is derived from the number of manuscripts using the same keyword. With these two scores, we created a new algorithm that finds alternative keywords with a high attention–survival score. We used ontologies to ensure that alternative keywords proposed by our method are semantically related to the original authors' keywords that they wish to refine. The hierarchical structure in an ontology supports the relationship between the alternative and input keywords. To test the sensibility of the ontology, we used two sources: WordNet and The Computer Science Ontology (CSO). Finally, we launched a survey for the human validation of our algorithm using keywords from Web of Science papers and three ontologies: WordNet, CSO and DBpedia. We obtained good results from all our tests.

**Keywords:** ontology; attention; survival; bibliometrics; keywords; papers

[1] CITIC-UGR. Universidad de Granada. 18071 Granada, Spain. ORCID: 0000-0002-6334-3786. jorgechp@correo.ugr.es

[2] Corresponding author.

[3] Departamento de Ciencias de la Computación e IA. CITIC-UGR. Universidad de Granada. 18071 Granada, Spain. ORCID: 0000-0001-7886-9329. rosa@decsai.ugr.es

**Introduction**

Authors use keywords to emphasize the most important topics of their papers. Keywords can play an important role when other researchers use recommendation systems to discover works related to a specified set of input terms. In addition, journals provide keywords with the title and abstract as part of the preliminary public information about a paper, so this information can be crucial for a researcher to determine whether they will read the full paper. Nevertheless, keyword selection is a process that is not perfect and often has many problems, such as selecting very specific or very generic keywords, misprints, author bias, and inexperience. These biases are aggravated by choosing keywords without following a keyword selection methodology.

When choosing generic or trendy terms, authors face the risk of sharing these terms with many papers that will compete with each other to claim researchers' attention. However, in choosing rare or specific terms, there is the risk of using terms that do not attract researchers.

In this paper, we studied how popularity relates to higher competition among keywords when choosing them for a manuscript. We analyzed ontologies and, by doing so, we were able to intuitively understand whether generic terms tend to be more popular, as well as more crowded, than specific ones. We aimed to formalize properties that would help us study this phenomenon, and we continued by analyzing the structure of keywords using an ontology.

While in the field of information retrieval we can find different measures to extract topics and keywords, such as TF-IDF, in this paper, we focused on the effects of social behaviors on the keywords and how different tendencies can affect the visibility of the information.

Finally, after the analytic stage, we propose a method to help authors refine their keyword selection processes. This method is based on measuring the popularity and crowding of desired terms while using the knowledge provided by ontologies to enhance the keywords proposed by an author.

This paper is structured as follows:

- Section 1 provides a literature review, which describes the current state-of-the-art theory that underpins the paper;
- Section 2 details the attention–survival model, which presents our theoretical proposal;
- Section 3 provides information on our experimental design, including details about the dataset used, an analysis of the WordNet and CSO ontologies, and examples illustrating our refinement algorithm;
- Section 4 summarizes our conclusions based on the results obtained from our experimental design;

Finally, this work contains a supplementary information section with more information regarding the algorithm results as well as the first and the second experiments.

**Literature review**

The International Organization for Standardization defines keywords as a word or group of words, possibly in a lexicographically standardized form, taken out of a title or the text of a document characterizing its content and enabling its retrieval (ISO 5963 1985). Apart from texts, keywords are often used to describe the content of a work by using words that contain the essential topics or themes that are represented in the work. For example, papers in the scientific community frequently come with a set of keywords, typically six. When authors decide what keywords they want to use for their manuscripts, we call them author keywords. The author can choose these keywords freely or from a prespecified list of terms (Lu et al. 2020). Sometimes, keywords are extracted using automatic procedures. This is the case for KeyWords Plus, where keywords are selected from the titles of articles cited in the references section (Zhang et al. 2016).

Keywords can have multiple applications, with one of the most used being information retrieval, where the scientific community uses keywords to search for information on certain topics (Grant 2010; Hartley and Kostoff 2003; Sesagiri Raamkumar, Foo, and Pang 2017). Keywords are also used to easily identify the most relevant content of an article, study the behavior of authors (Gil-Leiva and Alonso-Arroyo 2007; González et al. 2018), map the structure of the science (Lozano et al. 2019), or build a taxonomy, among others (X. Liu et al. 2012). In order to detect research trends (Wei Lu et al. 2021), author-defined keywords are used to represent topics in an ex ante approach, called author-defined keyword frequency prediction (AKFP), to detect research trends. Another example is searching by keywords using a traditional web service (Purohit, L., et al.2016).

Nowadays, plenty of search engines enable researchers to discover papers by typing in a set of keywords. Search engine users prefer this approach because users do not worry about grammatical rules; hence, they require less time and effort to formulate a query (Hasany, N., et al. 2010). Then, the search engine presents a list of recommended papers related to keywords, and the researcher can choose from them. This process was analyzed by H. Liu et al. (2020), who proposed a method to refine this recommendation process by choosing popular articles and very correlated keywords.

Aside from articles, keywords also have different degrees of popularity (Fernandes, Vinagre, and Cortez 2015). Keyword popularity is a topic that has gained attention in the field of marketing research. Jerath, Ma, and Park (2014) studied the different behaviors among customers who searched for popular terms and customers who used less popular keywords, concluding that the second group of customers spent more effort on their search process and were more likely to buy something in the end.

It is also relevant to note that author keywords can prevent interpretation biases by other authors (González et al. 2018). Nevertheless, author keywords are not free from other types of biases, as there are differences between experienced and nonexperienced authors (Sesagiri Raamkumar, Foo, and Pang 2017).

Authors' behavior when choosing keywords was studied by Hartley and Kostoff 2003. This paper performed a study on the habits of authors and editors regarding keywords, finding that, in the case of authors, it was very common to simply select as many keywords as desired, while editors tend to let authors choose the keywords for a manuscript. Hartley and Kostoff (2003) also focused on the problems generated by the inability of some authors to choose good

keywords and the inefficiency of search systems. Some of the problems mentioned are the use of ambiguous keywords and the overuse of keywords without justification.

For authors, it is crucial to select the correct keywords so that their papers are more easily visible to others. At the same time, from an editor's point of view, having a manuscript with proper keywords can help them improve their journal's impact (Pearce, Hicks, and Pierson 2018).

Some strategies have been proposed to mitigate the effects of selecting poor keywords. For example, Zhang et al. (2016) grouped terms with a similar meaning into single primary terms, while Lozano et al. (2019), in addition to removing excessively specific words, divided generic terms into specific ones. Intuitively, it seems that using overly generic or overly specific keywords is bad practice.

In the computer science field, ontologies are an explicit specification of a conceptualization (Gruber 1993). Ontologies represent concepts and their relations in terms of generalization and specificity and can have different applications in fields such as artificial intelligence, web semantics, or linguistics (Dong et al. 2021; Guarino, Oberle, and Staab 2009).

Ontologies have multiple applications in information retrieval and keywords. For example, Khan et al. (2004) used ontologies to process natural language. The input words were expanded into related concepts that help to create a keywords domain. Jose, V. et al. (2020) built a dynamic ontology based on The Computer Science Ontology. With the help of Word2Vec, the ontology was expanded in order to identify new academic research areas. Haribabu, S. et al. (2019) proposed a paradigm to provide webpage recommendations using semantic information and semantic ontologies. Their paradigm uses the similarity between words and a keyword expansion mechanism to identify new terms that may be interesting for the user who performs a search.

In addition, ontologies have been used to obtain the topics of documents. For example, Kong, H. et al. (2006) and Huang, M. et al. (2007) used an ontology structure to determine the topic of a web document. In Liu, M. et al.'s (2017) study, the semantic similarity among academic documents was determined. For this, the authors calculated the similarity between topic events based on the domain ontology to acquire the semantic similarity between articles.

**The Attention–Survival model**

**Basic model**

Our theoretical model is based on the premise that there is an information retrieval system in which the user introduces a set of keywords. The system randomly returns a list of papers that contain all these keywords but in a random order, and from there the user chooses one of the returned articles. We consider that the paper the user selects is the only one that "survives" the process.

Our basic model does not consider different biases that would typically exist in an information retrieval system, such as ordering by citation, relevance, impact factor, and publication date, as well as the attention's bias generated by the user when choosing an article.

An example of a more complex model where the information retrieval system is biased with respect to the publication date is presented in the appendix.

**The Attention–Survival score**

Let $K_j = \{k_1, k_2, \ldots, k_n\}$ be the set of initial candidate keywords, as defined by the user for manuscript $j$.

When authors search for manuscripts by entering specific inputs, we presume that only one manuscript will be extracted from each search process. Therefore, we denote the article selected as the survivor manuscript.

Let $C(k)$ be the community of keyword $k$. We define community as the set of manuscripts that contains a given keyword. $C(k)$ defines the set of articles containing the keyword $k$; thus, they are the ones that will compete with our manuscripts to survive.

Let $S(k), k \in K$ be the survival score of a keyword. Given an article that contains the keyword $k$, $S(k)$ is the propensity of the keyword to survive, according to our basic model. We assume that our recommendation system is neutral so that the retrieval process is unbiased. In order to help readers understand our model, we did not consider values by relevance, length, cites, impact, or publication date (i.e., basic model). If we apply a biased retrieval process, then we have to redefine $S(k), k \in K$, according to the bias applied. This situation is explained in the Appendix. Under our basic model supposition, the survival score is defined as follows:

$$S(k) = \frac{1}{|C(k)|}$$

We consider that an author can look for either one keyword (e.g., $k$) or a set of them (e.g., $K$). When looking for multiple keywords at the same time, the community $C(K)$ is described as the set of manuscripts that contains every keyword in $K$ simultaneously.

We can compute the survival score of $K$ as follows:

$$S_{\cup}(K) = \frac{\sum S(k_i)}{|K|}$$

where $K = \{k_1, k_2, \ldots, k_n\}$.

Another important concept in our work is keyword attention, $A(k)$, which is the level of interest shown by the community for a certain keyword. As discussed later in the paper, the attention of a word is a function of the number of times that word is used in a query. We derived this value using information provided by Google Trends.

Similarly, we can compute the attention of a set of keywords as the average value of the attention scores from every keyword in the input set. This is presented as

$$A_{\cup}(K) = \frac{\sum A(k_i)}{|K|}$$

We define the attention–survival score, $AS$, as the score of a manuscript defined by $K$ keywords. This score depends on the community and the attention of $K$:

$$AS_{\cup}(K) = \alpha \cdot S_{\cup}(K) + (1-\alpha) \cdot A_{\cup}(K)$$

where $\alpha \in [0,1], \alpha \in R$ is a weighting factor for $S_{\cup}(K)$ and $A_{\cup}(K)$.

Finally, it is relevant to consider that since survival scores range from 0 to 1, the attention score will need to be normalized.

***Keyword intersections***

Sometimes, information retrieval systems search for the intersection of each term introduced by the user instead of treating each term separately. In that case, we need to adjust our expressions. Firstly, we introduce the survival of an intersection as follows:

$$S_{\cap}(K) = \prod S(k_i)$$

while the attention of an intersection is defined in the following way:

$$A_{\cap}(K) = \prod A(k_i)$$

Attention scores for each keyword should be computed or extracted from a reliable data source. Finally, we can adapt the attention–survival metric previously defined as follows:

$$AS_{\cap}(K) = \alpha \cdot S_{\cap}(K) + (1-\alpha) \cdot A_{\cap}(K)$$

**Theoretical behavior**

***Proposals***

Here, we explain the theoretical behavior of survival and attention among the different levels of an ontology: from the root node to the very last child on the tree. The intuitive idea behind our model is that the number of manuscripts that use a particular term tends to be higher when the community's interest in that term is also high.

We present three propositions as follows:

**Proposition 1:** The survival score of a term depends on the specificity of that term inside the ontology structure so that the more specific a term, the greater the survival score.

**Proposition 2**: The attention level of a term depends on the specificity of that term inside the ontology structure so that the more specific a term, the less attention it will achieve.

**Proposition 3**: For every term, there is a point where survival and attention intersect, and that is the equilibrium point.

It is trivial to state that when the keyword does not have competitors, the survival score tends to be ∞ (i.e., $\lim_{C(k)\to 0} S(k) = \infty$), as opposed to when we have infinite competitors, where the survival score would be zero (i.e., $\lim_{C(k)\to \infty} S(k) = 0$). Concerning the attention score, if a term attracts the attention of infinite competitors, the attention will be at the maximum. In contrast, attention is zero when nobody is interested in that term.

Figure 3 graphically represents the *equilibrium point*. The equilibrium point is the level of specificity where the attention and survival scores intersect. Thus, the closest keyword to the equilibrium point would be the *equilibrium keyword*. The equilibrium point depends on various factors, such as the α value and $f_1$ and $f_2$, which refer to the minimum level of keywords and the minimum interest in terms that can be used to find the maximum depth of an ontology.

So, if we choose a more generic keyword than the equilibrium keyword, we are reducing the survival score, and with that, the AS score will also decrease. While choosing a keyword that is less generic than the equilibrium keyword will reduce the AS score due to a reduction in attention.

***Properties***

We can use the idea behind the dynamics of supply and demand (Whelan, Msefer, and V.Chung 2001) from econometrics to better understand the behavior of the attention and survival functions. However, we must take into consideration that behavior tends to be slightly different between the two concepts.

**Attention**

Attention is a dynamic function that fluctuates over time, as it describes the behavior of people. Therefore, the popularity of a keyword is constantly changing. For example, Figure 1 illustrates the behavior of the keyword "support vector machine" vs. "naive Bayes" from 2004 to 2021, according to Google Trends. As we can see, "support vector machine" seems to be surpassed by "naive Bayes" over time. Attention can play the same role as demand in economics theory. We also can interpret demand as the expected income a researcher hopes to receive when using a particular keyword.

A change in the popularity of a keyword produces a variation in the same sign in the attention function.

**Survival**

Survival is also dynamic and changes over time, as it describes the behavior of keywords. Survival can only grow to a certain fixed level based on the finite number of manuscripts that use a specific keyword. The role of survival might be similar to that of supply, but its dynamics are quite different. One approach can be to analyze attention and survival within a specific window of time so that survival would also be able to increase or decrease in response to tendencies. Survival can also be interpreted as the fixed price that a researcher must pay to use certain keywords.

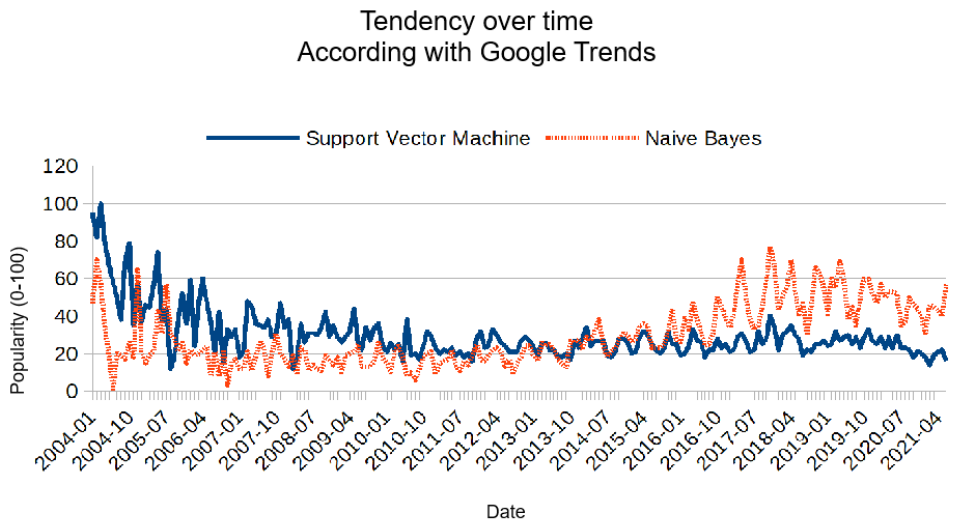
A change in the number of manuscripts that contains a specific keyword produces a variation in the opposite sign in the survival function.

**Complementary and substitutive keywords**

Beyond the relationships so far discussed, it should also be noted that keywords have relationships among themselves as well. In addition, sometimes, people start using a new keyword to refer to an existing concept. For example, in economic theory, a complementary keyword is a keyword of which the popularity can, in turn, affect the popularity of the related keyword. At the same time, when a complementary keyword is affected in terms of survival, the complemented one is affected in the same way. This relationship is common in the case of synonyms, semantic parents or children, or keywords significantly correlated to one another. For example, the term "machine learning" is highly connected with the term "artificial intelligence", according to Google Trends (see Figure 2). With this information, we can see that complementary keywords have a positive correlation. When the complemented keyword gains popularity, the complementary keyword also increases in popularity. When the researcher community increases the use of one keyword, the other keyword also experiences an increase.

If one keyword completely replaces another one, then we are talking about substitutive keywords. When one substitute candidate keyword experiences an increase in attention, the attention received by the replaced keyword decreases. Similarly, when one keyword decreases its survival score, the other one experiences a reduction in the rate that its survival decreases. If we use the window-in-time approach, there is a negative correlation between both keywords' survival scores. An example of possible substitutive keywords is "support vector machine" vs. "naive Bayes" (see Figure 1) or "C++" vs. "Python".

As always, it is important to note that correlation does not imply causation. For example, both keywords "Digimon" and "hip-hop" experienced a similar tendency on Google Trends, but there was no clear relationship between these




two concepts. While

correlation can help us identify associations between keywords, we are required to further analyze the information to make decisive conclusions. For example, ontologies, lists of synonyms and antonyms or analyses of social trends can help us to identify these associations.

**Outsiders, Outlier Keywords, and Local Maximums**

Not every keyword is part of an ontology relationship. For example, the "Me Too" movement and the hashtag #MeToo have an important attention score (France 2017), and there are many academic manuscripts that use "MeToo" as a keyword

*Figure 1: Global historic tendency of "Support Vector Machine" vs "Naive Bayes"*

(Blumell and Huemmer 2021). “Me Too”, for example, is an outlier keyword if we are using WordNet, where this concept is not represented.

Often, child concepts have better attention or survival than their parents. For example, “AIDS” has stronger popularity than its parent, “immunodeficiency”, in WordNet. Even if these local maximums’ existence is quite common within the ontology, the general tendency should follow the theoretical model posed in the previous section of our paper.

**Candidate generation**

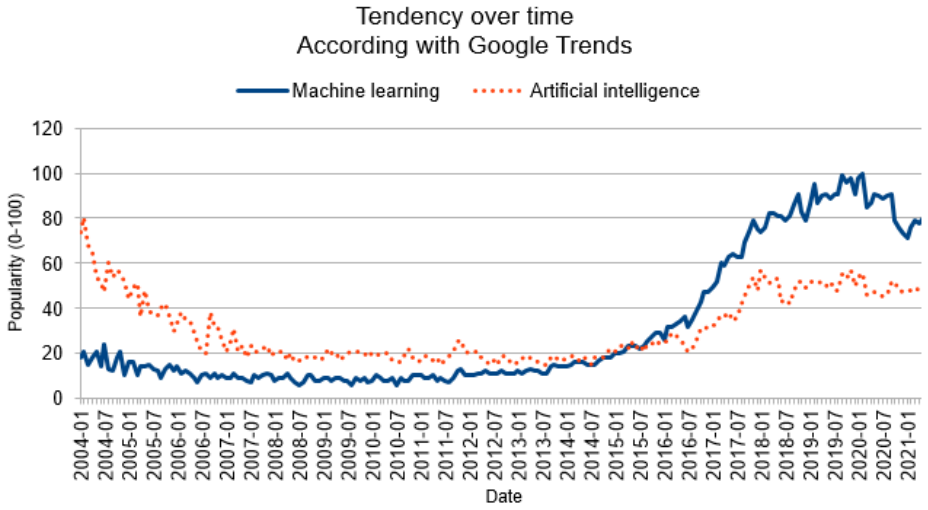



*Figure 2. Global historic tendency of “Machine learning” vs. “Artificial Intelligence”.*

We propose an iterative process to improve the attention–survival score of a manuscript’s keywords by using their keyword “neighbours”. To explore the neighborhood, we can use a variety of techniques. In our paper, we propose using ontologies, as we can use human knowledge to determine the meaningful relationships of a keyword. Often, keywords have a very specific meaning, and it is important to change their semantic role as little as possible.

As ontologies are represented and defined as a tree, we must assume a trade-off between being general and being specific. By generalizing, we will often be able to increase the attention score, but it will also increase the size of the community. Thus, an increase in $A(k)$ will often imply a decrease in $S(k)$. Conversely, moving to more specific keywords will increase $S(k)$ and decrease $A(k)$, as specific keywords are searched for less often than generic ones. For this reason, $\alpha$ and 1-$\alpha$ play an essential role in the refining process.

It is important to consider that ontologies can also contain synonyms (i.e., brother nodes). In relation to synonyms, it is difficult to predict their effects on survival and attention.

Let $g(k_i, k_j) = \frac{1}{d(k_i,k_j)}$ represent the benefit of selecting the keyword $k_i$ instead of $k_j$ in terms of the distance between nodes. When $k_i = k_j$, g is 1. We define the evaluation function $f(k_i, k_j, k_s)$ as

$$f(k_i, k_j, k_s) = AS(k_i) \cdot g(k_i, k_s) + AS(k_j) \cdot g(k_i, k_j)$$

where $k_s$ is the starting candidate keyword.

We aim to perform an iterative process to discover new candidate keywords and estimate whether paying the distance cost is worth increasing the AS score. Our iterative process is shown in Algorithm 1.

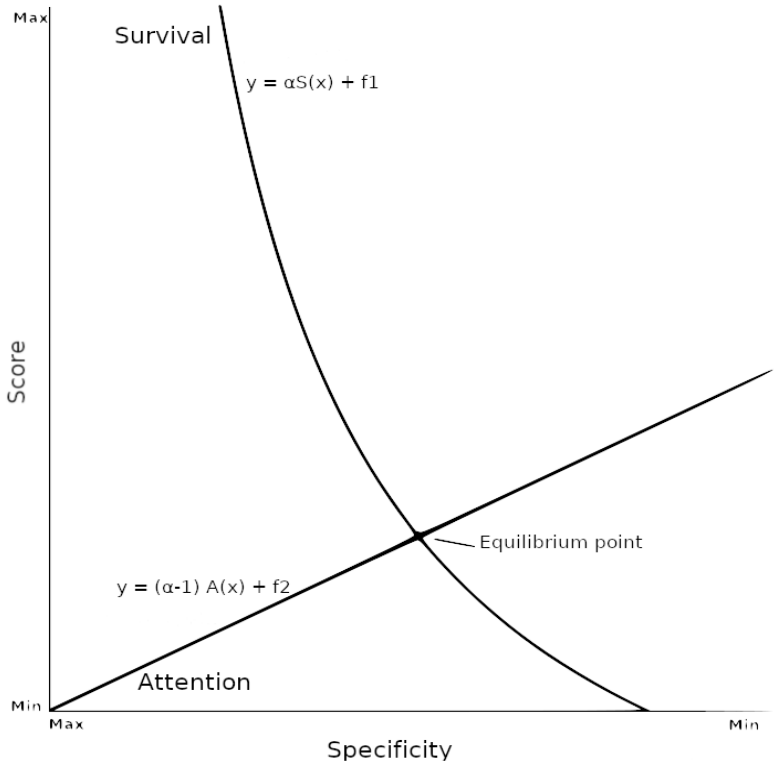



*Figure 3. Graphical representation of an equilibrium point.*

**Toy example**

To illustrate our method, we chose a real paper that only has three keywords: “Chlorofluorocarbons”, “sorption”, and “computer simulation” (George 1996).

The first step is to choose one keyword for the queue iteration, for example, “Chlorofluorocarbons”, and retrieve all their neighbors according to WordNet. Then, for each neighbor, we compute their attention–survival (AS) score as follows:

**Fluorocarbon → 4,95**; HCFC → 0,000264; Freon → 1,62;

hydrochlorofluorocarbon → 0,40535; Chlorofluorocarbons → 0

*Algorithm 1: Keyword refinement algorithm.*

1. We start with an initial candidate set , $K_{initial}$
2. $K_{candidates} = \{K_{initial}\}$
3. For each keyword, *k*, in $K_{initial}$:
   - 3.1. $k_{start} = k$ #$k_{start}$ is the starting candidate keyword.
   - 3.2. $K_{successors} = \{k_{start}\}$ #$K_{successors}$ is the set of potential replacements for $k_{start}$.
   - 3.3. $K_{queue} = \{k_{start}\}$ #$K_{queue}$ is a queue of candidates that have not been explored yet.
   - 3.4. While $K_{queue}$ is not empty:
     - 3.4.1. $k_{source} = next(K_{queue})$ #gets the element out in front
     - 3.4.2. $K_{neighbours}$ is the list of neighbours of $k_{source}$, including $k_{source}$.
     - 3.4.3. $AS_{neighbours}$ is the list of AS scores for each keyword in $K_{neighbours}$.
     - 3.4.4. $F_{neighbours} = \{f(ks_{ource}, k_j, k_{start})$ for each $k_j$ in $K_{neighbours}\}$
     - 3.4.5. $k_{best}$ is the keyword with the maximum $f$ value in $F_{neighbours}$.
     - 3.4.6. If $k_{best}$ is not in $K_{successors}$:
       - 3.4.6.1. $K_{successors} = K_{successors} \cup \{k_{best}\}$
       - 3.4.6.2. $K_{queue} = K_{queue} \cup \{k_{best}\}$
   - 3.5. For each keyword, $k_{candidate}$, *in* $K_{successors}$ *:*
     - 3.5.1. For each candidate set, $K_{candidate\_set}$, in $K_{candidates}$ :
       - 3.5.1.1. We create a new set, $K_{new\_candidate\_set}$ by replacing $k_{candidate}$ by *k*
       - 3.5.1.2. $K_{candidates} = K_{candidates} \cup K_{new\_candidate\_set}$
4. We return the set in $K_{candidates}$ that maximizes the AS score.

with "Fluorocarbon" being the neighbor with the best score. The next step is to compute ƒ, which considers the gain by moving from the original term to one of their neighbors. In this example, we consider that the benefit of moving to a direct parent, children, or other brother terms in the ontology will always be the same distance, 1. For example, the ƒ value of moving from "Chlorofluorocarbons" to their parent, "Fluorocarbon", can be expressed as follows:

$$f(\text{chlorofluorocarbons}, fluorocarbon, chlorofluorocarbons) = AS(chlorofluorocarbons) \cdot \cdot g(chlorofluorocarbons, chlorofluorocarbons) + AS(fluorocarbon) \cdot g(chlorofluorocarbons, fluorocarbon)$$

with

- $g(chlorofluorocarbons, chlorofluorocarbons) = 0;$

- $g(chlorofluorocarbon, fluorocarbon) = 1$;
- $AS(chlorofluorocarbons) = 0$;
- $AS(fluorocarbon) = 4{,}95$;

Afterwards, we compute f.

$$f(\text{chlorofluorocarbons}, fluorocarbon, chlorofluorocarbons) = 0 \cdot 1 + 4{,}95 \cdot 1 = 4.95$$

For the first step of the algorithm, retrieving the $f$ values is trivial, as the gain from moving to a direct neighbour is always one, and $k_i$ is the same as $k_{start}$, so all $f$ values will ultimately coincide with their AS scores.

$$f(\text{chlorofluorocarbons}, HCFC, chlorofluorocarbons) = 0{,}002$$

$$f(\text{chlorofluorocarbons}, freon, chlorofluorocarbons) = 1{,}62$$

$$f(\text{chlorofluorocarbons}, hydrochlorofluorocarbon, chlorofluorocarbons) = 0{,}41$$

$$f(\text{chlorofluorocarbons}, chlorofluorocarbons, chlorofluorocarbons) = 0$$

“Fluorocarbon” was the best scored term, so we added “Fluorocarbon” to the candidate set as well as to the queue for the following iteration.

We repeated the process, obtaining “Fluorocarbon” from the candidate set $k_i$ = “Fluorocarbon”. Note that our starting keyword in the algorithm, $k_{start}$, is “Chlorofluorocarbons”. When looking for neighbors and scores, we obtained 13 neighbors this time, with “Fluorocarbon” being the best. As “Fluorocarbon” was in the candidate set, we did not add it to the queue iteration.

The next step was to generate new candidate sets from the new candidate keywords. We proceeded by replacing the original keyword with the new candidate one. Thus, our new list of candidates was as follows:

{“Chlorofluorocarbons”, “sorption” and “computer simulation”}

{“Fluorocarbon”, “sorption” and “computer simulation”}

The next keyword to refine was sorption, which only had one good neighbor, “attention”. This meant we needed to add it to the new candidate sets:

{“Chlorofluorocarbons”, “sorption”, “computer simulation”}

{“Fluorocarbon”, “sorption”, “computer simulation”}

{“Chlorofluorocarbons”, “attention”, “computer simulation”}

{“Fluorocarbon”, “attention”, “computer simulation”}

Finally, it was time to refine “computer simulation”, which was a local maximum, meaning that this term did not have any neighbors with an AS score higher than its own AS score, so we did not add any new candidate sets. In conclusion, the best scored candidate set was **{“Fluorocarbons”, “attention”, “computer simulation”}**.

This entire process was firmly based on the knowledge from the ontology used (in our case, WordNet) and should be seen as a decision support system to help humans refine their keywords’ impact. In the example, “Chlorofluorocarbons” was replaced by “Fluorocarbon”, which is a more generic concept that includes all “Chlorofluorocarbons". Authors must then judge whether it is worth the cost to accept the loss of specific information to use a more attractive keyword for the audience.

In the case of “attention”, “sorption” is the generic form of “absorption”, so our algorithm moved to a child concept in order to finally end up with “attention”, which is another meaning of the keyword “absorption”. In the context of the article, it seems that “attention” is not a good choice to replace “sorption”, because the manuscript context seems

to be related to chemistry, not concentration. In this case, maybe the author would prefer to keep "sorption" or replace it with "absorption", which is a slightly more competitive term that receives more attention.

**Experimental design**

**Theoretical model validation**

***Data source***

To corroborate the validity of our proposals and gain a better understanding regarding the behavior of attention and survival on an ontology, we extracted data from a few different ontologies: WordNet (Miller 1995) and The Computer Science Ontology (CSO) (Salatino et al. 2018). WordNet is a lexical database that gathers words into groups of cognitive synonyms and defines relationships in terms of hypernymy and hyponymy. Thus, despite not being strictly an ontology, we can benefit from the WordNet structure, which also resembles the form of a tree, where each concept is a node.

For their part, CSO is an ontology automatically generated from 16 million publications focused on the computer science field. The CSO model includes eight different semantic relations (relatedEquivalent, superTopicOf, contributsTo, preferntialEquivalent, rdf:type, owlSameAs, and schema:relatedLink). We only used the first two relations mentioned.

All terms from the ontologies are lightly preprocessed to avoid ambiguity and to prepare them to be sent to the APIs in a proper format. The preprocessing steps are the following:

1. Replace "-" and "_" with spaces.
2. Remove all characters, except letters, spaces, and "&".
3. Replace "&" with "and".

For extracting the number of papers according to Scopus, we used the Scopus Search API. [4] We sent requests to the Scopus Search API using the following filters:

- We looked for terms inside the keywords' list of manuscripts: KEY("term");
- We filtered all manuscripts, except articles, reviews, and conference papers. DOCTYPE("ar"), DOCTYPE("re"), DOCTYPE("cp").

An example of a query search is as follows:

**KEY ( 'SCIENCE' ) AND ( LIMIT-TO ( DOCTYPE , 'ar' )**

**OR LIMIT-TO ( DOCTYPE , 're' ) OR LIMIT-TO ( DOCTYPE , 'cp' ) )**

[4] Elsevier Developer Portal: https://dev.elsevier.com/.

From Google Trends, we extracted the popularity results per country and computed the average popularity.

We aimed to analyze the differences between a generic source of terms such as WordNet and a more specific collection related to the computer science field. Scopus provided us with information on the number of papers per keyword necessary to infer survival, while Google Trends gave us information regarding the attention and popularity of a term. Unfortunately, it is important to state that extracting information from a specific academic search engine such as Google Scholar was impossible. Instead, Google Trends gave us results for Google Search, which is used by the general population. Nevertheless, using results from Google can provide us with additional information, such as altmetrics and the social interest in science topics. This paper used attention as an alternative to the topic popularity (TP) measure.

***Ontology analysis***

Our purpose was to map the terms in our datasets onto the ontology structure as defined by WordNet and CSO. Before mapping terms, we first wanted to perform an exploratory task on both WordNet and CSO to explore the ontologies' behavior and clarify whether the theoretical process of attention and survival metrics over an ontology was close to our assumptions. WordNet consists of different synsets outside a hierarchy, but we only studied those connected to the root synset (i.e., entity). A synset can have one or more lemmas, so we used the median value of the attention and survival scores across all lemmas. Using the median value instead of the mean to determine the score of a synset is based on the fact that the distribution scores of lemmas tend to present a skewed result. However, we used the mean attention and survival numbers to compute the scores per level. Table 1 shows the distribution of the Scopus and Google values' overall levels of depth in WordNet ($\alpha = 0.5$). We only analyzed depth levels that were greater than five, because, for prior levels, the number of synsets was reduced too much, which could lead to incorrect conclusions and inconsistent results.

After computing the attention and survival scores and extracting the mean values per level, we noted that the scores were on different scales, so we needed to perform a min–max normalization to keep all values in the same range [0,1]. We did this to make the analysis of the effect of both scores in the ontologies more accessible.

Figure 4 shows the survival and attention score evolution across the WordNet structure ($\alpha = 0.5$). As we can see, survival starts close to 1.0 at depth 17 and is in a continuous decline until being surpassed by attention at level 6, where the equilibrium point is located. Meanwhile, attention continually grows until it reaches its maximum value at level 2. The equilibrium point is located at the coordinates (6.38, 0.48), that is, in level 6, producing an AS score of 0.48, while the maximum score is achieved at level 2, with an AS score of 0.62.

If we check the results from Table 1, we can see that the average number of articles per level is significantly reduced until level 7, while attention tends to grow uniformly.

The observed behavior over the WordNet ontology is in consonance with our proposals. Moreover, we have empirically corroborated that more specificity is related to low attention and high survival scores.

Concerning the WordNet dynamic, we can see how most depth levels contain very specific terms, which are quite unattractive according to Google searches and the number of articles retrieved by a Scopus search.

From these results, one should not deduce that the best option is to choose keywords from levels 7 or 2. WordNet is a generic ontology that contains many terms that are not common in the academic field. Therefore, a domain-specific ontology would be a better option for choosing keywords. A good approach could be to choose an ontology according to the criteria described by Yu, Thom, and Tam (2007) (for example, the authors mention clarity, consistency, conciseness, expandability, correctness, completeness, minimal ontological commitment, and minimal encoding bias, among other criteria). Our purpose in employing WordNet was to use a generic and widely validated ontology to analyze the distribution of attention and survival scores. The case for CSO is illustrated in Figure 5.

In CSO, the equilibrium point was reached at level 10, while the maximum AS score was at level 2. In CSO, survival fell very fast, while attention had both fast-growing periods and periods of slow growth. In CSO, the equilibrium point score was very close to the maximum value of AS (the difference was less than 0.02).

Both ontologies show a sudden drop on the first level. It is important to state that the first level is not the root node, which was removed from our data. The upper levels of both ontologies contained such few words that their result could introduce noise into the graph and, thus, should be carefully interpreted.

***Keyword refinement***

In this section, we randomly chose keywords from 20 manuscripts, and we ran them through Algorithm 1. The attention results came from Google Trends and were normalized with the interval [0-100] for this refinement process. Thus, we did not need to perform a normalization step.

We limited the distance to the target keywords to two levels to prevent large differences in their conceptual meanings.

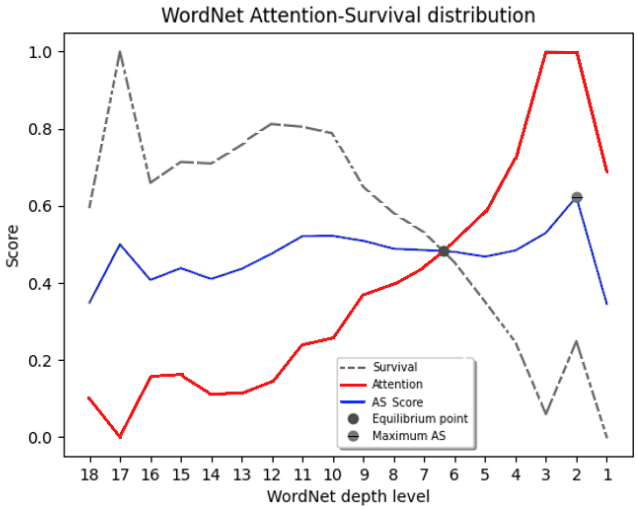



*Figure 4. Evolution of attention, survival, and attention–survival score (AS) throughout the WordNet structure.* $\alpha = 0.5$.

***WordNet and CSO refinements***

Table 2 shows real examples of the author's keywords refined using WordNet and CSO ontologies.[5] Keywords were randomly selected from the intersection of the terms contained in both ontologies. As CSO is generated from academic literature, all keywords in the example are real keywords. On the one hand, many keywords were not replaced by others. This can happen for two reasons: The algorithm is limited to exploring at only a distance of two. If the neighbors' attention–survival score is low, the algorithm decides not to replace the keyword.

On the other hand, if the keyword is not defined in the ontology, then we cannot use this ontology to refine the keyword, as we do not have enough information about the neighborhood. For WordNet, the distance between terms is provided by the path distance similarity, a metric that denotes how similar two synsets are based on the shortest path that connects the two nodes. This metric is provided by the Python Natural Language Toolkit (NLTK) library.[6] For CSO, we determined the distance between two words using the lowest common ancestor (LCA) algorithm (Aho, Hopcroft, and Ullman 1973).

[5] An extended version of Table 2 can be found in the Supplementary Materials of this paper.

[6] https://www.nltk.org/.

**Scopus and Google Scores in WordNet Ontology**

| Level | Scopus | Google |
|---|---|---|
| 1 | 22.823,0 | 5,504 |
| 2 | 13.961,5 | 2,756 |
| 3 | 39.043,0 | 3,626 |
| 4 | 31.037,5 | 3,538 |
| 5 | 1.203,0 | 2,608 |
| 6 | 263,0 | 2,160 |
| 7 | 95,0 | 1,900 |
| 8 | 49,0 | 1,676 |
| 9 | 34,5 | 1,580 |
| 10 | 25,0 | 1,494 |
| 11 | 11,5 | 1,200 |
| 12 | 9,5 | 1,044 |
| 13 | 10,0 | 0,836 |
| 14 | 11,5 | 0,776 |
| 15 | 16,0 | 0,774 |
| 16 | 17,3 | 0,920 |
| 17 | 14,8 | 0,882 |
| 18 | 3,0 | 0,544 |
| 19 | 16,5 | 1,044 |

*Table 1. Scopus and Google scores per depth level in WordNet. All values are the average scores of each synset with the same distance to the root node, and the value of each synset was computed by considering the median value of all lemmas within the synset. Note that level 1 only has the root synset.*

As we mentioned before, WordNet is probably not the best option for use as an ontology, and a knowledge-specific ontology should be used instead (e.g., CSO for computer science or The Ontology for Biomedical Investigations for biological or medical domains) (Bandrowski et al. 2016). We can see some replacements of keywords that perhaps are not the best option for manuscripts (Correspondence → card, Testing → Watch…).

The best way to use our method is inside an interactive system that allows the author to know the survival and attention scores from specific keywords and propose alternatives. Of course, the author should always make the final decision.

***WordNet and CSO Hierarchy***

WordNet and CSO have different purposes as ontologies. While some terms are present in both ontologies, the knowledge structure is different between them. This can generate very different results from one ontology to another, as reflected in Table 2. In our paper, we also compared the hierarchy of both ontologies. Figure 6 and Figure 9 present the structure of the same set of keywords according to both the WordNet and CSO ontologies. As these terms are included in both ontologies, we can suppose that these keywords are closely related to the computer science field. Since CSO is an ontology focused on computer science terminology, we can see how these keywords are connected to one another and have fewer isolated nodes. For WordNet, however, most of these keywords are completely isolated, and there are no strong clusters of keywords. Therefore, CSO represents terms with a greater granularity than WordNet, and this situation directly impacts the refinement algorithm.

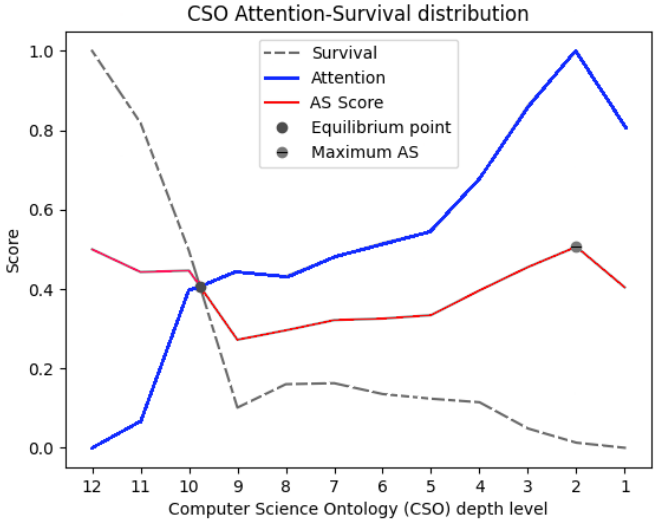



*Figure 5. Evolution of the attention, survival, and attention–survival score (AS) throughout the CSO structure.* $\alpha = 0.5$.

541

**Words Refinements: WordNet and CSO**

| Initial Word | WordNet Refinement | CSO Refinement |
|---|---|---|
| robotics | robotics | **robots** |
| telecommunication_equipment | **television** | **sensors** |
| electromagnetism | **acoustics** | **electromagnetic** |
| memory_access | memory_access | memory_access |
| computer-aided_design | **software** | **computer-aided** |
| gateway | gateway | **routing_protocols** |
| lexical_database | lexical_database | **artificial_intelligence** |
| speckle | speckle | **radar** |
| telecommunication_equipment | **television** | **sensors** |
| data_mining | **data_processing** | **clustering** |
| computer_science | **plan** | **software** |
| ergonomics | **technology** | **human_computer_interaction** |

| | | |
|---|---|---|
| cosmic_microwave_background | cosmic_microwave_background | **polarimeter** |
| buffer_storage | **fund** | **bandwidth** |
| white_noise | **impediment** | white_noise |
| relational_database | relational_database | **database** |
| electrical_energy | **AC** | electrical_energy |
| mobile_phone | **cell** | **sensors** |
| binoculars | binoculars | **binocular** |
| object-oriented_programming | **hack** | **java** |
| user_interface | **CLI** | **sensors** |
| authentication | **validation** | **security_of_data** |
| remote_control | **device** | **robotics** |
| spline | **remove** | **computer-aided_design** |

*Table 2. List of the different words before and after refining, using WordNet and CSO as ontologies.* [5]

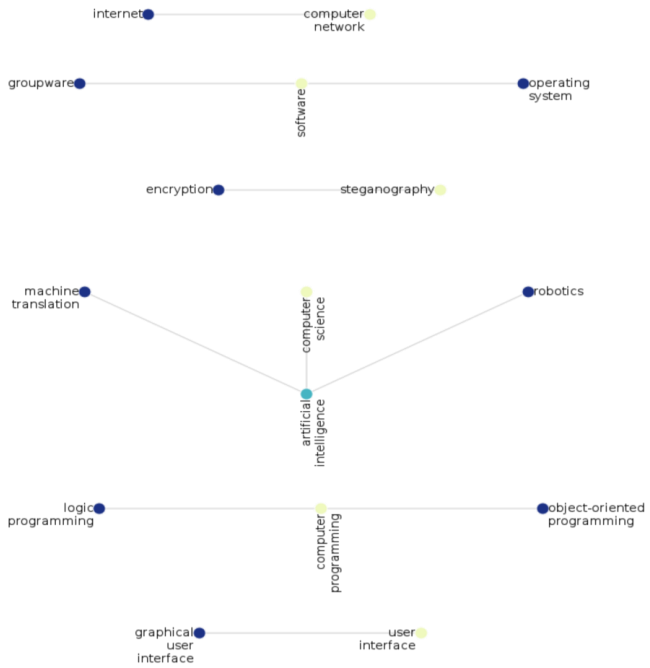



*Figure 6. WordNet hierarchy: the illustration represents the connections between different keywords in WordNet.*

**TF-IDF**

TF-IDF is probably the most popular measure to determine the relevance of a term inside of a corpus. In this work, we propose an alternative measure of popularity. In the following, we have listed some differences between the attention–survival score and TF-IDF.

- **Static vs. dynamic**: TF-IDF, applied over a corpus, is a static measure that is not liable to change while attention–survival is updated in real time. Even when applied over a static corpus, attention–survival will return a score that represents the current popularity of a measure according to the social attention given to a particular term. For this reason, the AS score is a recommended alternative to positioning a term taking into consideration the social behavior of a community around a term.
- **Source of information**: while TF-IDF relies on statistical information about the distribution of a term in a corpus, AS extracts its information from both statistical and temporal data.
- **Type of information**: On the one hand, TF-IDF is a technique that comes from the text mining field. This means it requires a corpus to extract the popularity score. On the other hand, the AS score does not require a corpus of text for it to work. It would be a suitable measure to categorize multimedia information, such as videos or photos.

***Alternative popularity measures***

As mentioned before, Google Trends provides information regarding the general behavior of internet users instead of the habits of researchers. Ya-Han et al. (2020) propose a measure named topic popularity (TP), which is represented as follows:

$$TP_Y^p = \sum_m \sum_n top_{m,n} \cdot \theta_m^p \cdot \varphi_n^m$$

where $\boldsymbol{p}$ is an article, Y is the year, $\boldsymbol{M}$ is the number of LDA topics, $\boldsymbol{N}$ is the number of keywords ($\boldsymbol{m} \in \{\mathbf{1}, \mathbf{2}, \dots, \boldsymbol{M}\}$,$\boldsymbol{n} \in \{\mathbf{1}, \mathbf{2}, \dots \boldsymbol{N}\}$), $\boldsymbol{\theta_m^p}$ is the probability of topic $\boldsymbol{m}$ occurring in paper $\boldsymbol{p}$, $\boldsymbol{\varphi_n^m}$ is the probability of keyword n occurring in topic m, and $\boldsymbol{top_{m,n}}$ is the number of questions retrieved by ResearchGate.

We aimed to analyze the differences between the Google Trends score and topic popularity. On the one hand, topic popularity collects data from ResearchGate[7], an academic source of scientific information. This guarantees that the popularity is not biased by data from outside of the scientific community, which is the main disadvantage of Google Trends. On the other hand, Google Trends represents popularity from the present, as it is based on current searches from the Internet, while ResearchGate could be counting information from the past (the minimum temporal unit of time used by ResearchGate is the year).

To perform the comparison, we used the same dataset mentioned in Section "Data Source", but we extracted the title and abstract from the articles and applied the LDA model as described by the authors in Ya-Han et al.( 2020). First, we defined 100 LDA topics with an asymmetric alpha parameter. After, we selected 400 random words from WordNet and used them as a target keyword, computing topic popularity and attention score for 2022. The results are shown in Figure 7. According to our results, TP always obtained higher attention values. This result can be explained by the fact that specific computer science terms are used more often by the academic public rather than the general audience. On the other hand, ResearchGate retrieves documents, links, and information generated before 2022 but uploaded in 2022, while Google Trends always shows results from 2022. Nevertheless, both metrics show a similar tendency, where popularity and attention decrease at higher WordNet levels.

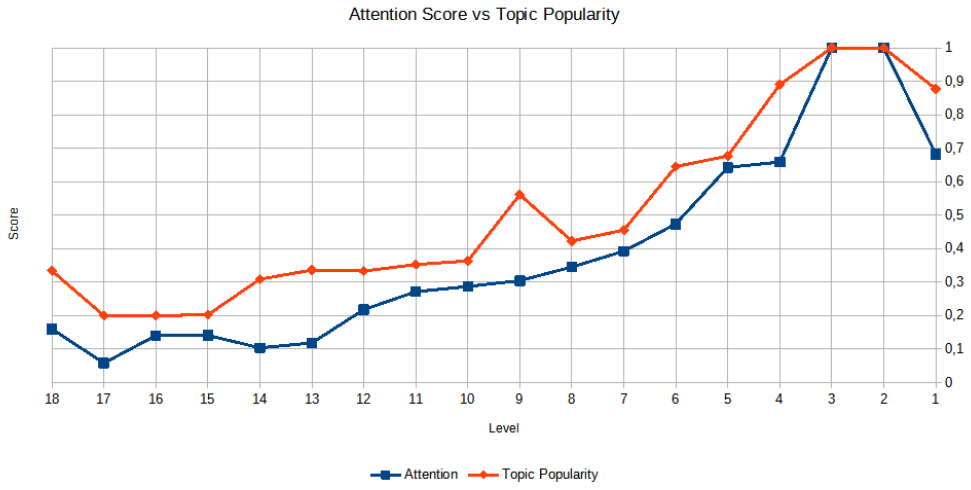



*Figure 7. Attention score vs topic popularity.*

[7] To collect data from ResearchGate, we used the online search engine Microsoft Bing, filtering by the domain researchgate.net.

**α parameter**

The α parameter allows us to calibrate the attention and survival scores.

Figure 8 represents the value of the AS score obtained with the words in the CSO ontology at different depth levels. Thus, at level 1, the words were more generic and less specific. At level 12, they were less generic and more specific. Figure 8 shows the analysis of the behavior of the AS score for different α values.

When α = 0, we only consider attention, while only survival is considered when α = 1. The green line in Figure 8 represents a low level with less specificity, and the survival score was low, whereas, at a high level, the score achieved the highest value. In addition, the blue line in Figure 8, with the opposite behavior, represents the attention score.

From Figure 8 and setting a word depth level, the most appropriate α can be chosen to maximize the AS score. Thus, for example, if the level of the words is 7, looking at Figure 8, the most suitable α would be equal to 0.6.

Moreover, by analyzing Figure 8, we can characterize the words of the CSO ontology. As shown, the words from level 1 to level 9 present a substantial attention value, meaning that they are generic words and are frequently used. From a depth level of 10, the words in the ontology are returned with a higher attention value, indicating that they are more specific.

Figure 8 also confirms Propositions 1, 2, and 3. Thus, for example, for α = 1, the green line shows that high survival values correspond to greater specificity (Proposition 1). At the same time, low survival values correspond to a lower specificity. In addition, when also analyzing the signal for α = 0, where only the attention component is active in the AS score, high specificity values correspond to low attention values. Following Proposition 3, the equilibrium point would be between levels 9 and 10.

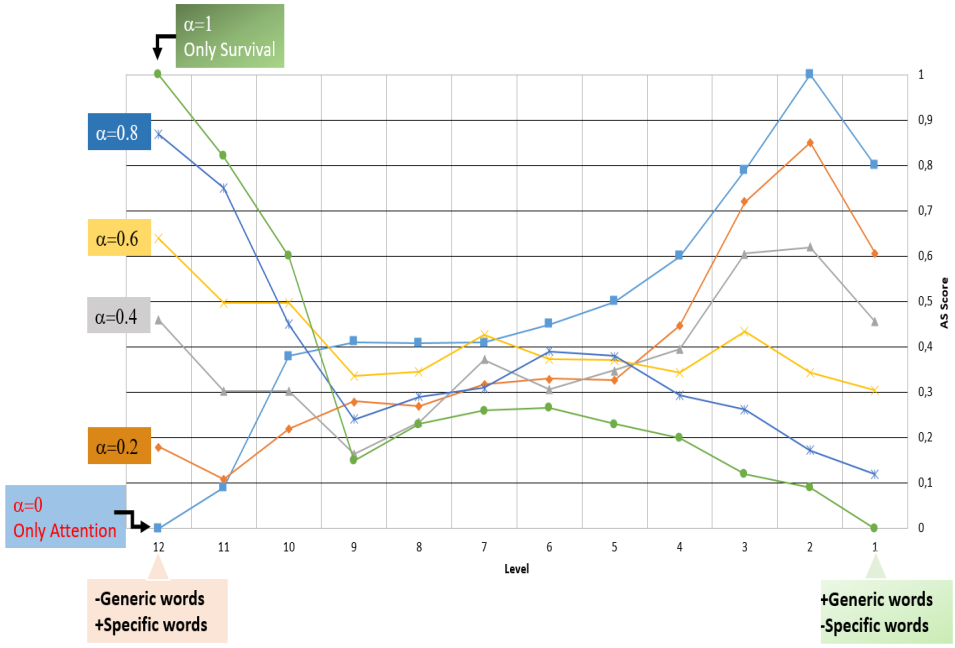



*Figure 8. Behavior of the AS score on CSO based on different α parameter values.*

**Refinement algorithm validation**

To validate Algorithm 1, we launched three experiments to test the adequacy of our model.

Experiments 1 and 2 used a survey to receive feedback from human subjects. The survey in both experiments was the same, but the target participants were different.

*Survey preparation*

First, we extracted 500 articles from Web of Science by selecting articles with the query "Computer Science" , including only papers. The complete dataset can be found at Kaggle.[8] The keywords from selected articles have been refined using Algorithm 1.

[8] https://www.kaggle.com/datasets/jorgechamorropadial/scie-2019.

To deal with unknown terms, we first looked at CSO, which is the ontology most related to the computer science dataset we used. If the keyword was not present in CSO, we the tried to find the concept in WordNet. When WordNet also failed, DBpedia[9] was used as a last attempt. Finally, if the term was not present in any ontology, we discarded the article and chose another one.

After obtaining a set of refined articles, we **randomly** chose ten papers to include in our survey. For each selected manuscript, we created a question in the survey where participants were allowed to see the title and abstract of the article and a table with two

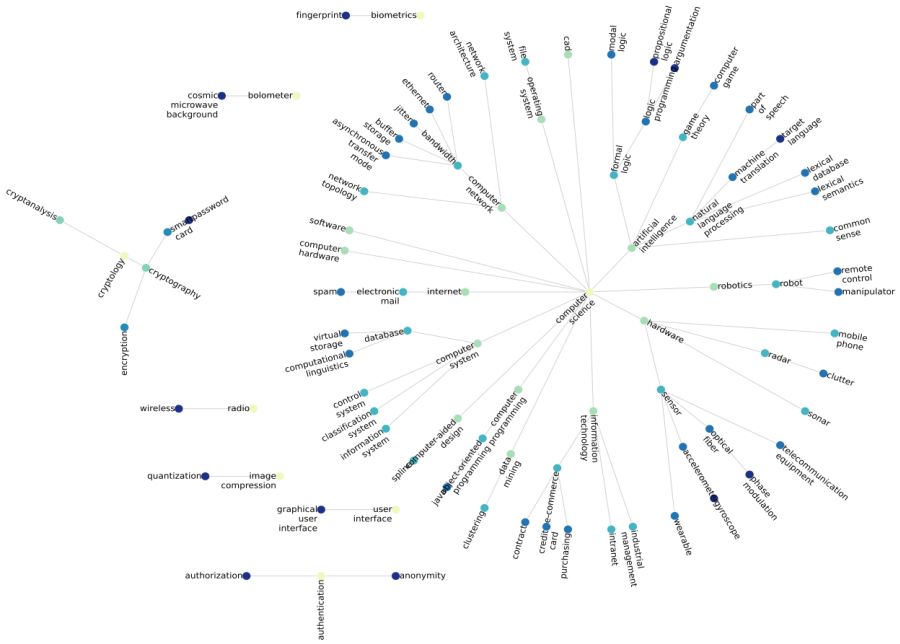



*Figure 9. CSO Hierarchy: the illustration represents the connections between the different keywords present in CSO.*

columns: one for the original set of keywords and another for the refined set. The user had to select one of the following answers:

[9] https://www.dbpedia.org

- **R1**: The refined set can describe the title with almost the same precision as the initial set;
- **R2**: The refined set cannot describe the title with almost the same precision as the initial set.

The Supplementary Materials contain all questions asked in the survey.

*Experiment 1: Survey for nonexperienced users*

Our survey was completed by 51 participants from Amazon Mechanical Turk. The participants were economically reimbursed for their participation and had to meet the following preliminary requirements:

- Live in an English-speaking country;
- Have a bachelor's degree;
- Working experience in IT.

Regarding the survey results, for all questions, most users answered R1. Figure 10 describes the results obtained per question. The worst performance was obtained for q7, where 50.98% of the participants chose R1, while the best results were obtained for q1, where 88.24% of the participants chose R1. In the entire survey, the answer R1 was chosen by 67. 85% (standard deviation: 10.43).

Five participants chose R1 for all questions, while nobody selected R2 more times than R1. In global terms, the participants chose R1 for 6.78 questions (standard deviation: 1.62).

As shown in Figure 11, the refined set of keywords deeply improved the AS score compared to the initial ones. The average improvement ratio is 22.5644 (standard deviation: 24.06).

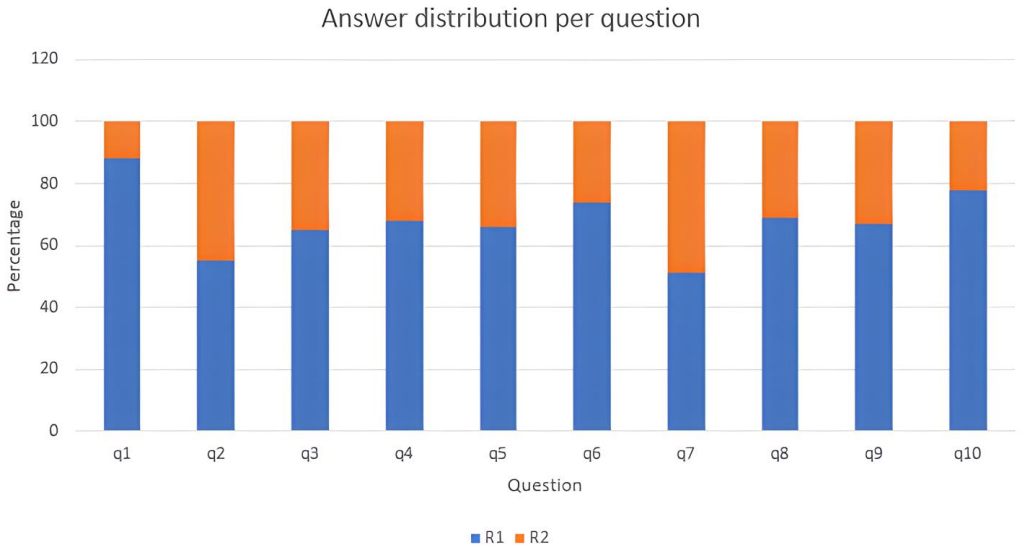



*Figure 10. Experiment 1: answer distribution per question, where the y-axis illustrates the percentage of R1 and R2 per question.*

*Experiment 2: Survey for experienced users*

Participants had to meet the following preliminary requirements:

- Live in an English-speaking country or have a C1 level of English according to The Common European Framework of Reference for Languages or higher;
- A master's thesis and a minimum of two scientific publications;
- Work experience in IT or in research.

Our survey was completed by 46 participants. A set of demographic questions were presented, with the following responses:

- English level:
    - Nineteen participants lived in an English-speaking country or had a native level of English;
    - Twenty-four participants had a C1 level or equivalent;

- Three participants had a C2 level.

- Education:
  - Twenty-five participants completed a master's thesis;
  - Twelve participants had a PhD;
  - Nine participants were university professors or researchers.
- Background:
  - Thirty-two participants had two scientific publications;
  - Fourteen participants had more than two scientific publications.

With respect to the survey results, for all questions, most of the users answered R1. Figure 11 describes the results obtained per question. The worst performance was obtained for q7, where 45.20% of the participants chose R1, while the best results were obtained for q6, where 87.95% of the participants chose R1. In the entire survey, the answer R1 was chosen by 63.52% (standard deviation: 11.82).

Three participants chose R1 for all questions, while nobody selected R2 more times than R1. In global terms, participants chose R1 for 6.20 questions (standard deviation: 2.01).

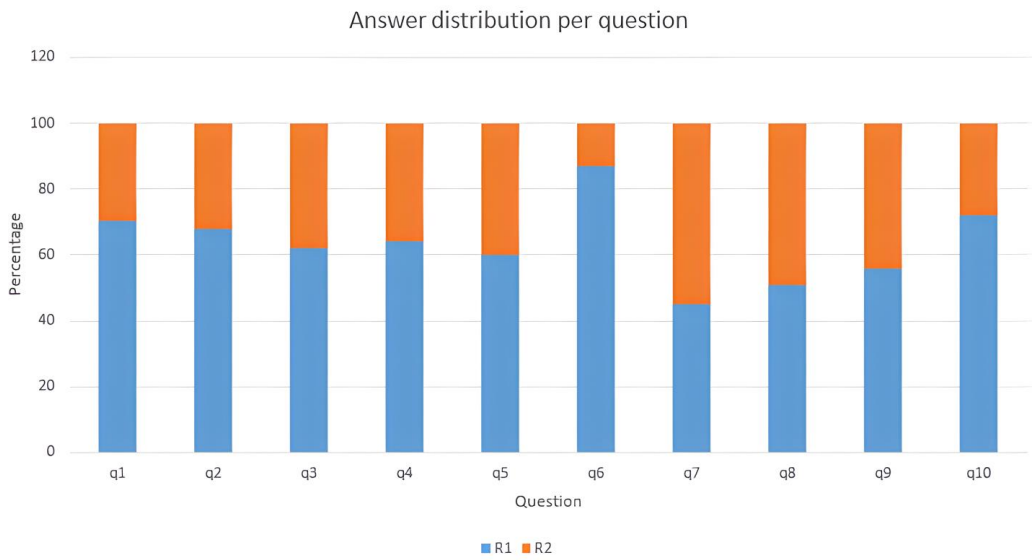



*Figure 11. Experiment 2: answer distribution per question, where the y-axis illustrates the percentage of R1 and R2 per question.*

As we can see in Figure 12, the refined set of keywords deeply improved the AS score in comparison with the initial ones. The average improvement ratio was 22.5644 (standard deviation: 24.06).

*Experiment 3*

For this experiment, we randomly selected 330 articles from the same dataset downloaded for the survey. For each article, every keyword had a 50% probability of being changed by another term that was related but with a lower AS score.

After changing the keywords of the 30 articles, we applied our algorithm to each of them and checked the intersection over union (IoU) metric. IoU is defined as follows:

$$IoU = \frac{\text{Proposed keywords } \cap \text{ Original Keywords}}{\text{Proposed keywords } \cup \text{ Original Keywords}}$$

where Proposed Keywords are the new keywords suggested by the algorithm, and Original Keywords is the original set of keywords chosen by the article's authors.

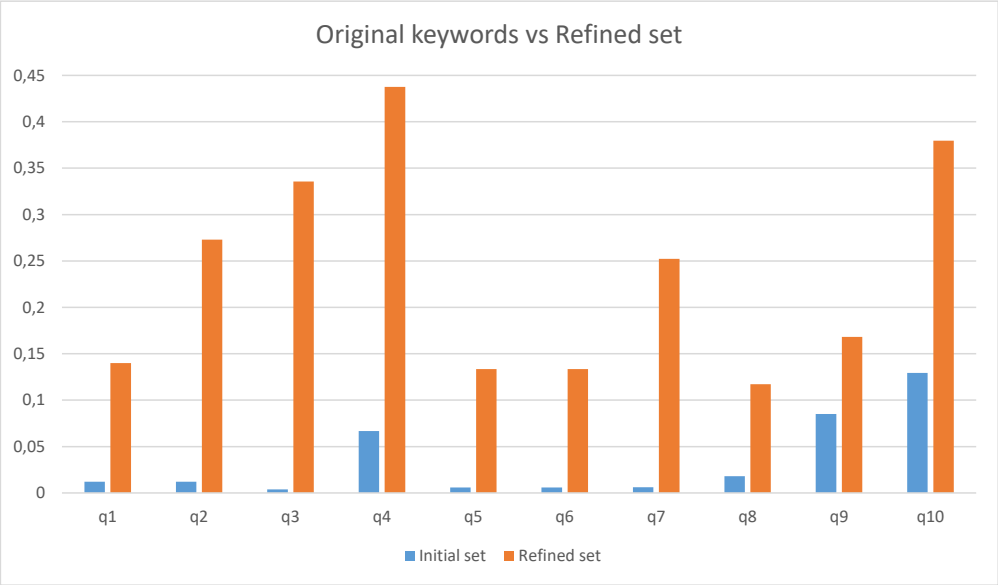



*Figure 12. This figure compares the AS score obtained using the original set of keywords (blue bars) with the AS score obtained using the refined keywords (red bars).*

We obtained the following results:

- IoU score: 0.594 (standard deviation: 0.307);
- Precision: 0.694 (standard deviation: 0.261);
- Recall: 0.696 (standard deviation: 0.258);
- Accuracy: 0.690 (standard deviation: 0.265);
- F1 score: 0.69 (standard deviation: 0.100);
- Rate of changed keywords: 0.431.

Note that not every keyword is susceptible to being changed. Keywords were not replaced if they were not defined in any of the three ontologies (WordNet, DBPedia, and CSO), even if they should be modified because of the probability.

**Future work**

Apart from the theoretical model tested in this work, as well as the three experiments performed and described below, the algorithm is expected to be put into production in order to serve other researchers and general users in an intuitive way. For this reason, a web application is being built to provide users with an intuitive and interactive user interface. The application is being designed to support the most common ontology formats (e.g., n3, owl, and xml) and is expected to be published in future works along with more experimental results to examine the performance of our algorithm.

**Conclusions**

Typically, keywords are selected by an author's intuition or sometimes without applying any method at all. This can lead to bias, errors, and loss of opportunity.

The goal of our paper was to emphasize the importance of knowing the results of choosing different keywords. Choosing a keyword implies putting a future manuscript in competition with others, which all work to gain a certain amount of attention from the community that varies and depends on external factors. For this reason, keywords are constantly changing in terms of attention and survival rates. Concerning survival, we can conclude that all keywords decrease their survival possibilities over time. However, in general terms, survival tends to decrease when moving from specific concepts to generic ones. At the same time, attention tends to decrease when moving from generic terms to specific ones. Sometimes, attention and survival intersect at certain equilibrium points. A keyword with both survival and attention scores that are simultaneously high characterizes a keyword that will be used across time and will continue to be of interest to the community. To establish the survival and attention values of a keyword, we defined the AS score.

We presented an algorithm to refine keywords using ontologies to find alternative keywords with high survival and attention scores. Ontologies can be used as an essential source of knowledge that can help us organize keywords along the generic–specific axis. We analyzed WordNet and The Computer Science Ontology (CSO) both ontologies but with different backgrounds. CSO is a field-specific ontology, while WordNet provides good comparison data to examine how our model works.

Implicitly, our method uses state-of-the-art strategies to reduce the probability of choosing a poor keyword (Zhang et al. 2016, Lozano et al. 2019) thanks to the implementation of ontologies and the possibility of moving into general or specific terms, according to the score obtained through the concepts hierarchy.

Another important topic is human validation. We performed a survey where 51 participants answered positively in relation to the results attained using our algorithm, which used WordNet, CSO and DBpedia as ontological sources.

Our method can be generalized and applied to other fields, for example, marketing. In addition, it can be helpful as a system to extract keywords when text mining is not an option, for example, if we want to categorize images or videos.

In conclusion, it is important to state that our algorithm is intended to provide authors with additional information regarding how to choose keywords for a manuscript and to propose suggestions. However, the author is ultimately responsible for making the decision. Finally, our algorithm is not a keyword suggester; if an author makes a bad decision in choosing a keyword, the refinement process likely will not help very much because it can only explore the

related context of a keyword. In that case, the author must use some other methodology or information to select a good starting keyword candidate set.

**Author contributions**

Conceptualization, Jorge Chamorro-Padial; Formal analysis, Jorge Chamorro-Padial; Funding acquisition, Rosa Rodríguez-Sánchez; Methodology, Jorge Chamorro-Padial; Project administration, Rosa Rodríguez-Sánchez; Software, Jorge Chamorro-Padial; Supervision, Rosa Rodríguez-Sánchez; Validation, and Rosa Rodríguez-Sánchez; Visualization, Rosa Rodríguez-Sánchez; Writing—original draft, Jorge Chamorro-Padial; Writing – review and editing, Rosa Rodríguez-Sánchez.

**Appendix**

**Biased models**

The basic model is useful for helping introduce our proposal and allows us to study the behavior of survival and attention scores without bias. Nowadays, academic information retrieval systems usually tend to return results using a concrete aspect of the manuscript (date of publication, impact, number of citations, journal, relevance, altmetrics, etc.), so there are plenty of biases to take into consideration to better estimate survival and attention.

A straightforward case of a biased information retrieval system is one that keeps a prioritized list of papers according to a certain parameter (date, relevance, etc.). The list is ordered in descending order of survival scores so that the first paper in a prioritized list of n papers has n times greater survival score than the last document on the list. In this situation, we need to redefine survival to consider the position of the paper on the list. The biased survival score of a manuscript, $S_{biased}$, could be expressed as follows:

$$S_{biased}(p_t, K) = \frac{|\{p: Pos(p,K) \leqslant Pos(p_t,K)\}|/p \in C(K)}{\sum_{i=1}^{|C(K)|} i}$$

where $p_t$ is the target paper the survival score of which we want to study, $K$ is the set of keywords introduced by the user, and $Pos(p,K)$ is the position of the paper on the biased list returned by the information retrieval system so that the first paper returned by looking for papers that contain the keywords in the set $K$ is $p_1$.

**References**

Aho, A. V., J. E. Hopcroft, and J. D. Ullman. 1973. “On Finding Lowest Commong Ancestors in Trees.” In *Http://Dx.Doi.Org/10.1137/0205011*, ACM, 253–65.

Bandrowski, Anita et al. 2016. “The Ontology for Biomedical Investigations.” *PLoS ONE* 11(4).

Blumell, Lindsey E., and Jennifer Huemmer. 2021. “Reassessing Balance: News Coverage of Donald Trump’s Access Hollywood Scandal before and during #metoo.” *Journalism* 22(4): 937–55. http://journals.sagepub.com/doi/10.1177/1464884918821522 (June 5, 2021).

Dong, Sicong, Yike Yang, He Ren, and Chu-Ren Huang. 2021. “Directionality of Atmospheric Water in Chinese: A Lexical Semantic Study Based on Linguistic Ontology.” *SAGE Open* 11(1): 215824402098829. http://journals.sagepub.com/doi/10.1177/2158244020988293 (May 9, 2021).

Fernandes, Kelwin, Pedro Vinagre, and Paulo Cortez. 2015. “A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News.” In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 535–46.

France, Lisa Respers. 2017. “#MeToo: Social Media Flooded with Personal Stories of Assault.” *CNN*. https://web.archive.org/web/20171016002502/http://www.cnn.com/2017/10/15/entertainment/me-too-twitter-alyssa-milano/index.html (June 5, 2021).

George, A. R. 1996. “A Computational Investigation of Zeolite-Chlorofluorocarbon Interactions.” *Zeolites* 17(5–6): 466–72.

Gil-Leiva, Isidoro, and Adolfo Alonso-Arroyo. 2007. “Keywords given by Authors of Scientific Articles in Database Descriptors.” *Journal of the American Society for Information Science and Technology* 58(8): 1175–87. http://doi.wiley.com/10.1002/asi.20595 (January 15, 2020).

González, Luis Millán et al. 2018. “An Author Keyword Analysis for Mapping Sport Sciences.” *PLoS ONE* 13(8): 1–22.

Grant, Maria J. 2010. “Key Words and Their Role in Information Retrieval.” *Health Information & Libraries Journal* 27(3): 173–75. http://doi.wiley.com/10.1111/j.1471-1842.2010.00904.x (May 7, 2021).

Gruber, Thomas R. 1993. “A Translation Approach to Portable Ontology Specifications.” *Knowledge Acquisition* 5(2): 199–220. https://linkinghub.elsevier.com/retrieve/pii/S1042814383710083 (May 9, 2021).

Guarino, Nicola, Daniel Oberle, and Steffen Staab. 2009. “What Is an Ontology?” In *Handbook on Ontologies*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1–17. http://link.springer.com/10.1007/978-3-540-92673-3_0 (May 9, 2021).

Haribabu, S., Kumar, P. S. S., Padhy, S., Deepak, G., Santhanavijayan, A., & Kumar, N. (2019, December). A novel approach for ontology focused inter-domain personalized search based on semantic set expansion. In 2019 fifteenth international conference on information processing (ICINPRO) (pp. 1-5). IEEE.

Hartley, James, and Ronald N. Kostoff. 2003. “How Useful Are ‘key Words’ in Scientific Journals?” *Journal of Information Science* 29(5): 433–38. http://journals.sagepub.com/doi/10.1177/01655515030295008 (May 7, 2021).

Hasany, N, Jantan, A.B., Selamat, M.H.B. and Saripan M.I., 2010. "Querying Ontology using Keywords and Quantitative Restriction Phrases". *Information Technology Journal, 9: 67-78.*

Huang, M., Kong, H., Baek, S., & Kim, P. 2007.TSM. Topic Selection Method of Web Documents. In *First Asia International Conference on Modelling & Simulation (AMS'07)* (pp. 369-374). IEEE.

ISO 5963. 1985. “ISO/IEC 5963:1985 Documentation - Methods for Examining Documents , Determining Their Subjects , and Selecting Indexing Terms.” *Iso 5963:1985*: 3–5. https://www.iso.org/standard/12158.html.

Jerath, Kinshuk, Liye Ma, and Young-Hoon Park. 2014. “Consumer Click Behavior at a Search Engine: The Role of Keyword Popularity.” *Journal of Marketing Research* 51(4): 480–86. http://journals.sagepub.com/doi/10.1509/jmr.13.0099 (May 9, 2021).

Jose, V., Jagathy Raj, V.P., George, S.K. 2021. Ontology-Based Information Extraction Framework for Academic Knowledge Repository. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Fifth International Congress on Information and Communication Technology. *Advances in Intelligent Systems and Computing*, vol 1184. Springer, Singapore. https://doi.org/10.1007/978-981-15-5859-7_6

Khan, L., McLeod, D., & Hovy, E. 2004. Retrieval effectiveness of an ontology-based model for information selection. T*he VLDB Journal*, 13(1), 71-85.

Kong, H., Hwang, M., Hwang, G., Shim, J., & Kim, P. 2006, November). Topic selection of web documents using specific domain ontology. In *Mexican International Conference on Artificial Intelligence* (pp. 1047-1056). Springer, Berlin, Heidelberg.

Liu, Hanwen, Huaizhen Kou, Chao Yan, and Lianyong Qi. 2020. “Keywords-Driven and Popularity-Aware Paper Recommendation Based on Undirected Paper Citation Graph.” *Complexity* 2020.

Liu, Xueqing, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. “Automatic Taxonomy Construction from Keywords.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, New York, USA: ACM Press, 1433–41. http://dl.acm.org/citation.cfm?doid=2339530.2339754 (May 7, 2021).

Liu, M., Lang, B., & Gu, Z. (2017). Calculating Semantic Similarity between Academic Articles using Topic Event and Ontology. *arXiv*. https://doi.org/10.48550/arXiv.1711.11508

Lozano, S., L. Calzada-Infante, B. Adenso-Díaz, and S. García. 2019. “Complex Network Analysis of Keywords Co-Occurrence in the Recent Efficiency Analysis Literature.” *Scientometrics* 120(2): 609–29. https://doi.org/10.1007/s11192-019-03132-w.

Lu, Wei et al. 2020. “How Do Authors Select Keywords? A Preliminary Study of Author Keyword Selection Behavior.” *Journal of Informetrics* 14(4): 101066.

Miller, George A. 1995. “WordNet: A Lexical Database for English.” *Communications of the ACM* 38(11): 39–41. http://portal.acm.org/citation.cfm?doid=219717.219748 (October 23, 2018).

Pearce, Patricia F., Rodney W. Hicks, and Charon A. Pierson. 2018. “Keywords Matter: A Critical Factor in Getting Published Work Discovered.” *Journal of the American Association of Nurse Practitioners* 30(4): 179–81.

Purohit, L., & Kumar, S. 2016. "Web service selection using semantic matching". In *Proceedings of the International Conference on Advances in Information Communication Technology & Computing* (pp. 1-5).

Salatino, Angelo A. et al. 2018. “The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas.” In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 187–205.

Sesagiri Raamkumar, Aravind, Schubert Foo, and Natalie Pang. 2017. “Using Author-Specified Keywords in Building an Initial Reading List of Research Papers in Scientific Paper Retrieval and Recommender Systems.” *Information Processing and Management* 53(3): 577–94.

Wei Lu, Shengzhi Huang, Jinqing Yang, Yi Bu, Qikai Cheng, Yong Huang. 2021. "Detecting research topic trends by author-defined keyword frequency, ".Information Processing & Management 58 (4) https://doi.org/10.1016/j.ipm.2021.102594.

Whelan, Joseph, Kamil Msefer, and Celeest V.Chung. 2001. *Economic Supply & Demand*. Cambridge, Mass. : MIT, 2001.

Hu, Y. H., Tai, C. T., Liu, K. E., and Cai, C. F. 2020. Identification of highly-cited papers using topic-model-based and bibliometric features: The consideration of keyword popularity. *Journal of Informetrics*, 14(1), 101004.

Yu, Jonathan, James A. Thom, and Audrey Tam. 2007. “Ontology Evaluation Using Wikipedia Categories for Browsing.” In *International Conference on Information and Knowledge Management, Proceedings*, New York, New York, USA: ACM Press, 223–32. http://portal.acm.org/citation.cfm?doid=1321440.1321474 (June 3, 2021).

Zhang, Juan et al. 2016. “Comparing Keywords plus of WOS and Author Keywords: A Case Study of Patient Adherence Research.” *Journal of the Association for Information Science and Technology* 67(4): 967–72.

# Attention–Survival score: A metric to choose better keywords and improve visibility of information

## Supplementary Material

### WordNet and CSO comparison

Down below, the Table 1 summarizes a list of common words included in WordNet and CSO ontology. The algorithm is limited to explore at only a distance of two neighbours from the original word.

| Original word | WordNet refinement | Computer Science Ontology (CSO) refinement |
|---|---|---|
| argumentation | discussion | computer_programming_languages |
| accelerometer | measuring instrument | sensors |
| administrative_data_processing | data processing | database |
| anonymity | anonymity | privacy |
| artificial_intelligence | robotics | artificial_intelligence |
| authentication | validation | security_of_data |
| binoculars | binoculars | binocular |
| biometrics | eugenics | access_control |
| blog | diary | internet |
| broadcasting | radio | communication_channels_ %28information_theory%29 |
| buffer_storage | fund | bandwidth |
| cad | scoundrel | computer_aided |
| channel_capacity | channel_capacity | communication_channels_ %28information_theory%29 |
| chaotic_attractor | attractor | cryptography |
| classification_system | edition | control_system |
| clustering | meet | clustering |
| computer_programming | computer_programming | computer_programming |
| computer_science | plan | software |
| computer_virus | computer_virus | security_of_data |
| computer-aided_design | software | computer_aided |
| computerized_tomography | tomography | medical_imaging |
| contract | flex | contract |
| control_system | servo | control_system |
| cosmic_microwave_background | cosmic_microwave_background | polarimeter |
| cryptography | cryptography | cryptography |
| cryptology | psychology | cryptography |
| data_mining | data_processing | clustering |
| demodulation | reception | communication_channels_ |

| | | |
|---|---|---|
| | | information_theory |
| e-commerce | e-commerce | internet |
| edutainment | entertainment | e-learning |
| electrical_energy | AC | electrical_energy |
| electromagnetism | acoustics | electromagnetic |
| electronic_mail | spam | internet |
| elementary_education | pedagogy | e-learning |
| ergonomics | technology | human_computer_interaction |
| facial_expression | frown | facial_expression |
| file_system | file_system | computer_science |
| flash_memory | flash_memory | embedded_system |
| gateway | gateway | routing_protocols |
| handover | release | mobile_telecommunication_systems |
| holmium | lu | rare_earth |
| industrial_management | technology | information_technology |
| knowledge_base | realm | artificial_intelligence |
| lexical_database | lexical_database | artificial_intelligence |
| logic_gate | logic_gate | hardware |
| logic_programming | programming | computer_programming_languages |
| machine_translation | robotics | artificial_intelligence |
| magnetic_storage | magnetic_storage | microprocessor_chips |
| memory_access | memory_access | memory_access |
| mobile_phone | cell | sensors |
| network_architecture | spec | computer_network |
| neural_network | rf | neural_network |
| nuclear_physics | crystallography | nuclear_physics |
| object-oriented_programming | hack | java |
| optical_fiber | loofa | sensors |
| outage | breakdown | outage |
| owl | raptor | semantic_web |
| phase_modulation | fm | sensors |
| privacy | privacy | privacy |
| programming_language | prolog | programming_language |
| purchasing | purchasing | internet |
| reconstruction | makeover | reconstruction |
| relational_database | relational_database | database |
| remote_control | device | robotics |
| robotics | robotics | robots |
| search_engine | program | internet |
| sip | ingestion | internet_protocol |
| source_code | source_code | software_engineering |
| speckle | speckle | radar |
| spin | revolution | spin |
| spline | remove | computer-aided_design |
| surveillance_system | surveillance_system | cryptography |
| tactics | tactics | software_design |
| target_language | language | machine_translation |
| telecommunication_equipment | television | sensors |
| teleconferencing | discussion | telecommunication_services |
| transmission_control_protocol | http | internet_protocol |
| user_interface | CLI | sensors |
| validation | validation | validation |
| verification | checksum | verification |

| | | |
|---|---|---|
| virtual_storage | memory | database |
| visual_communication | graphics | image_coding |
| web_page | web_page | internet |
| white_noise | impediment | white_noise |
| wireless | wireless | cellular |

Table S1: WordNet vs CSO refinement. The list of original words is included in both ontologies (WordNet and CSO).

## Survey structure

### *Question 1*

**Title:** HUBBLE: an optical link management system for dense wavelength division multiplexing networks

**Abstract**

Timely detection of Dense Wavelength Division Multiplexing (DWDM) link quality and service performance problems of fiber deployment are important and critical for telecommunication operators. In this paper, we propose a new methodology for network fault detection inside optical transmission systems deployed in a real-operator environment and present the working principles of the system. Our new calculation methodology is used for joint fiber and DWDM link quality evaluation inside the proposed High-level Unified BackBone Link Examiner (HUBBLE) platform. At the end of the paper, we also detail some of the benefits, challenges, and opportunities of automation in DWDM networks using the proposed HUBBLE platform.

| Initial set | Refined set |
|---|---|
| Dense wavelength division multiplexing | D.W.D.M |
| service provider | service provider |
| fault management | network services |
| link quality | link quality |

### *Question 2*

**Title**: Feature Completion for Occluded Person Re-Identification

**Abstract**

Occluded person re-identification (Re-ID) is a challenging problem due to the destruction of occluders. Most existing methods focus on visible human body parts through some prior information. However, when complementary occlusions occur, features in occluded regions can interfere with matching, which affects performance severely. In this paper, different from most previous works that discard the occluded region, we propose a Feature Completion Transformer (FCFormer) to implicitly complement the semantic information of occluded parts in the feature space. Specifically, Occlusion Instance Augmentation (OIA) is proposed to simulates real and diverse occlusion situations on the holistic image. These augmented images not only enrich the amount of occlusion samples in the training set, but also form pairs with the holistic images. Subsequently, a dual-stream architecture with a shared encoder is proposed to learn paired discriminative features from pairs of inputs. Without additional semantic information, an occluded-holistic feature sample-label pair can be automatically created. Then, Feature Completion Decoder (FCD) is designed to complement the features of occluded regions by using learnable tokens to aggregate possible information from self-generated occluded features. Finally, we propose the Cross Hard Triplet (CHT) loss to further bridge the gap between complementing features and extracting features under the same ID. In addition, Feature Completion Consistency (FC2) loss is introduced to help the generated completion feature distribution to be closer to the real holistic feature distribution. Extensive experiments over five challenging datasets demonstrate that the proposed FCFormer achieves superior performance and outperforms the state-of-the-art methods by significant margins on occluded datasets.

| Initial set | Refined set |
|---|---|
| Training | Training |
| Task analysis | Task analysis |
| Feature extraction | Feature extraction |
| Image retrieval | Image retrieval |
| Optimization | Local convergenge |
| Semantics | Linguistics terminology |
| Person re-identification | Person re-identification |
| Data mining | virtual_learning |

### *Question 3*

**Title**: Robust Adaptive Fault-Tolerant Control of Multiagent Systems With Uncertain Nonidentical Dynamics and Undetectable Actuation Failures

**Abstract**

This paper studies the distributed consensus problem of multiagent systems (MASs) in the presence of nonidentical unknown nonlinear dynamics and undetectable actuation failures. Of particular interest is the development of a robust adaptive fault-tolerant consensus protocol capable of compensating uncertain dynamics/disturbances and time-varying yet unpredictable actuation failures simultaneously. By introducing the virtual parameter estimation error into the artfully chosen Lyapunov function, the consensus problem is solved with a robust adaptive fault-tolerant control scheme based upon local (neighboring) agent state information. It is shown that the proposed method is user friendly in that there is no need for detail dynamic information of the agent or costly detection/diagnosis of the actuation faults in control design and implementation, resulting in a structurally simple and computationally inexpensive solution for the leaderless consensus problem of MAS. Simulation results illustrate and verify the benefits and effectiveness of the proposed scheme.

| Initial set | Refined set |
|---|---|
| Fault tolerant control | Fault tolerant control |
| Adaptative control | Mobile agent |
| Multiagent systems | Speed control |

### *Question 4*

**Title:** Mobile Robot Obstacle Avoidance Based on Neural Network with a Standardization Technique

**Abstract**

Reactive algorithm in an unknown environment is very useful to deal with dynamic obstacles that may change unexpectantly and quickly because the workspace is dynamic in real-life applications, and this work is focusing on the dynamic and unknown environment by online updating data in each step toward a specific goal; sensing and avoiding the obstacles coming across its way toward the target by training to take the corrective action for every possible offset is one of the most challenging problems in the field of robotics. This problem is solved by proposing an Artificial Intelligence System (AIS), which works on the behaviour of Intelligent Autonomous Vehicles (IAVs) like humans in recognition, learning, decision making, and action. First, the use of the AIS and some navigation methods based on Artificial Neural Networks (ANNs) to training datasets provided high Mean Square Error (MSE) from training on MATLAB Simulink tool. Standardization techniques were used to improve the performance of results from

the training network on MATLAB Simulink. When it comes to knowledge-based systems, ANNs can be well adapted in an appropriate form. The adaption is related to the learning capacity since the network can consider and respond to new constraints and data related to the external environment.

| Initial set | Refined set |
|---|---|
| Obstacle avoidance | Autonomous vehicles |
| Simulation-based learning | Simulation-based learning |
| Neural networks | Radial basis function |
| Autonomous mobile robots | Autonomous robots |

### *Question 5*

**Title**: Deep temporal motion descriptor (DTMD) for human action recognition

**Abstract**

Spatiotemporal features have significant importance in human action recognition, as they provide the actor's shape and motion characteristics specific to each action class. This paper presents a new deep spatiotemporal human action representation, the deep temporal motion descriptor (DTMD), which shares the attributes of holistic and deep learned features. To generate the DTMD descriptor, the actor?s silhouettes are gathered into single motion templates by applying motion history images. These motion templates capture the spatiotemporal movements of the actor and compactly represent the human actions using a single 2D template. Then deep convolutional neural networks are used to compute discriminative deep features from motion history templates to produce the DTMD. Later, DTMD is used for learning a model to recognize human actions using a softmax classifier. The advantage of DTMD are that DTMD is automatically learned from videos and contains higher-dimensional discriminative spatiotemporal representations as compared to handcrafted features; DTMD reduces the computational complexity of human activity recognition as all the video frames are compactly represented as a single motion template; and DTMD works effectively for single and multiview action recognition. We conducted experiments on three challenging datasets: MuHAVI-Uncut, iXMAS, and IAVID-1. The experimental findings reveal that DTMD outperforms previous methods and achieves the highest action prediction rate on the MuHAVI-Uncut dataset.

| Initial set | Refined set |
|---|---|
| Human activity recognition | Human activity recognition |
| Deep convolutional neural network | Deep convolutional neural network |
| Motion history images | Motion history images |
| Deep temporal motion descriptor | Deep temporal motion descriptor |

| | |
|---|---|
| Computer vision | Linear motor |

### *Question 6*

**Title**: Well placement optimization using metaheuristic bat algorithm

**Abstract**

The design of an optimal field development and production management is a complicating task because of influencing various factors on decision-making process. Typical factors include number and type of wells, well locations, production constraints, economic factors like capital expenditure, operating costs, and oil sale price. The situation is further complicated due to the uncertainty associated with various effective engineering and geological parameters.

In this study, three meta–heuristics algorithms of genetic algorithm (GA), particle swarm optimization (PSO) and bat inspired algorithm (BA) are used for optimal determination of six production well locations. Net present value (NPV) is used as an objective function in optimization process. PUNQ-S3 benchmark model is simulated in MATLAB environment in order to search the entire complex reservoir during optimization. Next, the effectiveness of algorithms will be compared in terms of convergence rate and NPV improvement over iterations.

The simulation results show that the BA is superior since it reduces the number of functional evaluations and thus improving the computational efficiency. In addition, the BA provides better NPV improvement over PSO and GA. The results indicate that the BA increases NPV by 7.5% and 21.7% over PSO and GA respectively.

| Initial set | Refined set |
|---|---|
| Fractional calculus | Fractals |
| Bat algorithm | Bat algorithm |
| Levy flight | Nonlinear equations |
| Non-parametric statistical tests | Non-parametric statistical tests |

### *Question 7*

**Title:** An Iterated Multi-stage Selection Hyper-heuristic

**Abstract**

There is a growing interest towards the design of reusable general purpose search methods that are applicable to different problems instead of tailored solutions to a single particular problem. Hyper-heuristics have emerged as such high level methods that explore the space formed by a set of heuristics (move operators) or heuristic components for solving computationally hard problems. A selection hyper-heuristic mixes and controls a predefined set of low level heuristics with the goal of improving an initially generated solution by choosing and applying an appropriate heuristic to a solution in hand and deciding whether to accept or reject the new solution at each step under an iterative framework. Designing an adaptive control mechanism for the heuristic selection and combining it with a suitable acceptance method is a major challenge, because both components can influence the overall performance of a selection hyper-heuristic. In this study, we describe a novel iterated multi-stage hyper-heuristic approach which cycles through two interacting hyper-heuristics and operates based on the principle that not all low level heuristics for a problem domain would be useful at any point of the search process. The empirical results on a hyper-heuristic benchmark indicate the success of the proposed selection hyper-heuristic across six problem domains beating the state-of-the-art approach.

| Initial set | Refined set |
|---|---|
| Hyperheuristics | Cognitive systems |
| Simultaneous multithreadings | Simultaneous multi-threading |
| Resource partitioning | Resource partitioning |
| Fuzzy partition | Optimization techniques |

## *Question 8*

**Title:** An alternative method of biomedical signal transmission through the GSM voice channel

**Abstract**

In this work, a new solution for online and accurate biomedical data transmission is presented. For this purpose, a global system for mobile (GSM) communication voice channel is, for the first time, used as a communication link between the patient and healthcare provider. Biomedical signals are converted into speech-like signals before being transferred over a GSM voice channel. On the receiver side, speech-like symbols are stored in a symbols bank, and constructed using random stochastic signals. On the receiver end, the index of the symbol with the most similarity to the received signal is selected as the identified sample. This method

enables us to communicate with an accuracy of 99.8% at a transfer rate of 110 samples per second and signal-to-noise ratio (SNR) of 10. By utilizing a GSM voice channel, any voice channel, such as a cell phone, can be used for data transmission. The transmitted signal is encoded; therefore, the connection is secured. GSM technology has benefits such as availability, reliability, and robustness. Additionally, GSM can be used as a backup or service for transmitting vital physiological signals in emergency situations (e.g. in an ambulance). This technology can also be used to transmit other physiological signals as well as nonphysiological generic data.

| Initial set | Refined set |
|---|---|
| Electrocardiography | Mathematics in medicine |
| GSM | GSM |
| Telemedicine | Telemedicine |
| Voice | Voice |
| Speech codecs | Speech codecs |

### *Question 9*

**Title**: Characterizing Generalized Rate-Distortion Performance of Video Coding : An Eigen Analysis Approach

**Abstract**

Rate-distortion (RD) theory is at the heart of lossy data compression. Here we aim to model the generalized RD (GRD) trade-off between the visual quality of a compressed video and its encoding profiles (e.g., bitrate and spatial resolution). We first define the theoretical functional space W of the GRD function by analyzing its mathematical properties. We show that W is a convex set in a Hilbert space, inspiring a computational model of the GRD function, and a method of estimating model parameters from sparse measurements. To demonstrate the feasibility of our idea, we collect a large-scale database of real-world GRD functions, which turn out to live in a low-dimensional subspace of W. Combining the GRD reconstruction framework and the learned low-dimensional space, we create a low-parameter eigen GRD method to accurately estimate the GRD function of a source video content from only a few queries. Experimental results on the database show that the learned GRD method significantly outperforms state-of-the-art empirical RD estimation methods both in accuracy and efficiency. Last, we demonstrate the promise of the proposed model in video codec comparison.

| Initial set | Refined set |
|---|---|
| Rate-distortion function | Rate-distortion function |
| Video quality assessment | Video quality |
| Quadratic programming | Quadratic programming |

### *Question 10*

**Title**: WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions

**Abstract**

In many areas, practitioners need to analyze large datasets that challenge conventional single-machine computing. To scale up data analysis, distributed and parallel computing approaches are increasingly needed.

Here we study a fundamental and highly important problem in this area: How to do ridge regression in a distributed computing environment? Ridge regression is an extremely popular method for supervised learning, and has several optimality properties, thus it is important to study. We study one-shot methods that construct weighted combinations of ridge regression estimators computed on each machine. By analyzing the mean squared error in a high dimensional random-effects model where each predictor has a small effect, we discover several new phenomena.

1. Infinite-worker limit: The distributed estimator works well for very large numbers of machines, a phenomenon we call "infinite-worker limit".

2. Optimal weights: The optimal weights for combining local estimators sum to more than unity, due to the downward bias of ridge. Thus, all averaging methods are suboptimal.

We also propose a new Weighted ONe-shot DistributEd Ridge regression (WONDER) algorithm. We test WONDER in simulation studies and using the Million Song Dataset as an example. There it can save at least 100x in computation time, while nearly preserving test accuracy.

| Initial set | Refined set |
|---|---|
| Distributed learning | Learning methods |
| High-dimensional statistics | High-dimensional statistics |
| Ridge regression | Optimization techniques |
| Random matrix theory | Random variable |

## 4.4. Corner Centrality of Nodes in Multilayer Networks: A Case Study in the Network Analysis of Keywords

### 4.4.1. Datos generales

1. **Autores**: Rosa Rodríguez-Sánchez, Jorge Chamorro-Padial.
2. **Revista**: Algorithms.
3. **Datos sobre la publicación**:
   - **Referencia**: Rodríguez-Sánchez y Chamorro-Padial (2022)
   - **Año**: 2022.
   - **Editorial**: MDPI.
   - **DOI**: https://doi.org/10.3390/a15100336
4. **Estado**: Publicado.
5. **Métricas**:
   - **Ranking**:
     - *Emerging Sources Citation Index (ESCI)*[10]:
       - *Computer Science, Artificial Intelligence*: Q3 - 121/190 (año 2021).
       - *Computer Science, Theroy & Methods*: Q3 - 81/143 (año 2021).

### 4.4.2. Contribuciones principales

- Hemos propuesto un modelo para determinar la importancia de un nodo en una red multicapa considerando la importancia tanto de su nodo como de su vecindario.
- Nuestro modelo ha sido validado frente a dos medidas de centralidad en redes multicapa bien conocidas: Pagerank versatilty (Domenico et al., 2015) y APABI (Agryzkov et al., 2019).
- Frente a PageRank versatility, reducimos los requisitos de memoria necesarios, obteniendo unos resultados muy positivos.

[10] A fecha de depósito de esta tesis, aún no se disponen de datos del año 2022.

### 4.4.3. Resumen

En nuestro trabajo, mediante el uso de redes complejas (formadas, por ejemplo, por agentes que participan y se relacionan en diferentes contextos), proponemos un método para la identificación de nodos que ocupan una posición central dentro de una red. Esta identificación se realiza mediante un análisis en redes multicapa: En primer lugar, se determina la importancia de un nodo en cada una de las capas de la red. El segundo paso consiste en integrar la importancia que recibe un nodo en cada una de las capas, determinando así el rol del nodo en la red general. Nuestro análisis no se detiene exclusivamente en analizar la importancia de un nodo de forma aislada, sino que tomamos en cuenta su vecindario, analizando la importancia de los nodos directamente enlazados. De esta manera, los nodos con mayor puntuación y el vecindario más relevante consiguen las mejores puntuaciones de centralidad de acuerdo a nuestro método.

Nuestra medida de centralidad se denomina *Corner centrality*. Entrando en más detalle, un nodo recibe una puntuación elevada de centralidad si es reconocido como central en todas las capas. Para que un nodo reciba ese reconocimiento en una capa determinada, debe destacar por encima de su vecindario. En el artículo, se emplea un ejemplo muy representativo de nuestro escenario: *Si dos países quieren negociar para poner fin a un conflicto por la vía diplomática, se debe buscar una persona que sea reconocida por ambos países. Si una de las dos partes no acepta a esta persona, la negociación no puede comenzar.*

Cada nodo recibe un valor de importancia inicial (ejemplos de puntuaciones iniciales pueden ser el factor de impacto de una revista, el índice h de un autor, la centralidad de grado (Zhang y Luo, 2017/03)...). Este valor de importancia es nuestro punto de referencia, pero para determinar el rol de un nodo, utilizamos el concepto de *importancia relativa*: un nodo es percibido como importante si está rodeado de nodos que tienen un menor valor de importancia. Es decir, si el nodo destaca en su vecindario. Esta importancia se puede determinar tomando en consideración los enlaces en la misma capa o bien, acudiendo a los enlaces entre capas. Para el primer caso, la importancia relativa se mediría de la siguiente manera:

$$I_L(u) = \sum_{\forall v \in N_L(u)} i_L(u) - i_L(v)$$

En el segundo escenario, tomando en consideración los enlaces entre capas, la importancia relativa se expresa así:

$$I_L(u) = \sum_{\forall v \in N_L(u)} i_L(u) - i_{L*}(v)$$

Donde:

- $u$ es un nodo perteneciente a la capa $L$.
- $N_u$ es el conjunto de vecinos del nodo $u$.
- $L*$ es una capa de la red, diferente a $L$.

Una vez tenemos definida la importancia relativa, podemos construir el concepto de *importancia global* mediante la siguiente expresión:

$$I(u) = (\sum_{i=0}^{M} I_{L_i}(u))^2$$

Donde $M$ hace referencia al número total de capas presentes en la red. Esta medida de importancia global aún debe ser normalizada, tal y como se explica en nuestro artículo. Si un nodo tiene un importancia relativa elevada, el valor se transmite entre los nodos de su vecindario.

Una vez definido nuestro modelo a nivel teórico, realizamos diferentes pruebas experimentales a fin de medir el rendimiento de nuestro modelo en diferentes escenarios. Los experimentos realizados han sido cuatro:

1. **Equipo de fútbol**: Utilizando una red formada por dos capas de veinte nodos cada una, representando cada nodo a un jugador de un equipo de fútbol. En una capa se representa las conexiones entre jugadores en redes sociales, y en otra capa se representan los pases de balón que ocurren entre jugadores. Se realiza una comparación entre APABI y Corner Centrality.

2. **Familias de Florencia**: Utilizando la red multicapa definida por las conexiones de negocios entre miembros de familias de Florencia del siglo XV (Breiger y Pattison, 1986). Para validar nuestro método, en este experimento hemos realizado una comparación con otras medidas de centralidad, como Pagerank versatilty (Domenico et al., 2015) y APABI (Agryzkov et al., 2019).

3. **Scopus vs Google Trends**: A partir de un corpus de 69.000 palabras clave relacionadas con las Ciencias de la Computación, medimos la importancia de cada palabra clave utilizando información de Scopus y de Google Trends.

4. **Author Keywords vs KeyWords Plus**: Es un experimento similar al anterior, partiendo del mismo corpus, se busca determinar las palabras clave centrales tomando en consideración las relaciones entre las palabras clave escogidas por los propios autores y aquellas palabras clave designadas por el algoritmo descrito por Eugene Garfield en 1990 basado en el título y las referencias de un documento (Garfield, 1990).

Finalmente, queda la validación del modelo, para la cual hemos comparado *corner centrality* con Pagerank versatilty. Aquí nuestro objetivo no era evaluar si una métrica rinde igual que otra sino el verificar que nuestra métrica se comporta de una forma similar a otra métrica para la cual ya existe consenso y validación de la Comunidad Científica. Para ello, hemos calculado el Coeficiente de correlación de Spearman y obteniendo unos resultados que indican una correlación elevada, con un $p$ valor bajo.

El articulo ha sido publicado en la modalidad de Acceso Abierto.

# Corner Centrality of Nodes in Multilayer Networks: A case study in the network analysis of keywords

Rosa Rodriguez-Sánchez
Jorge Chamorro-Padial

**Abstract** In this paper, we present a new method to measure the nodes' centrality in a multilayer network The multilayer network represents nodes with different relations between them. The nodes have an initial relevance or importance value. Then, the node's centrality is obtained according to this relevance along with its relationship to other nodes. Many methods have been proposed to get the node's centrality by analyzing the network as a whole. In this paper, we present a new method to get the centrality in which, in the first stage, every layer would be able to define the importance of every node in the multilayer network. In the next stage, we would integrate the importance given by each layer to each node. As a result, the node that is perceived with a high level of importance for all of its layers, and the neighborhood with the highest importance, gets the highest centrality score. This score has been named the Corner Centrality.

As an example of how the new measure works, suppose we have a multilayer network with different layers, one per research area, and the nodes are authors belonging to an area. The initial importance of the nodes (authors) could be their h-index. A paper published by different authors generates a link between them in the network. The authors can be in the same research area (layer) or different areas (different layers). Suppose we want to get the centrality measure of the authors (nodes) in a concrete area (target layer). In the first stage, every layer (area) receives the importance of every node in the target layer. And in the second stage, the relative importance given for every layer to every node is integrated with the importance of every node in its neighborhood in the target layer. This process can be repeated with every layer in the multilayer network. The method proposed has been tested with different configurations of multilayer networks receiving excellent results. Moreover, the proposed algorithm is very efficient regarding computational time and memory requirements.

**Keywords** networks centrality; multilayer networks; PageRank centrality; Corner centrality; author's keywords

---

Rosa Rodriguez-Sánchez
Departamento de Ciencias de la Computación e I.A. CITIC-UGR. Universidad de Granada, 18071 Granada, Spain.
ORCID: 0000-0001-7886-9329
E-mail: rosa@decsai.ugr.es

Jorge Chamorro-Padial (corresponding author)
CITIC-UGR.Universidad de Granada.18071 Granada, Spain
ORCID:0000-0002-6334-3786
E-mail: jorgechp@correo.ugr.es

## 1 Introduction

Which countries are most relevant in the world for being the largest producers of raw materials (food, minerals, etc.) for the rest of the world?; what are the most effective drugs for a set of diseases?; which diplomats are the most relevant to carry out deals between countries in conflict?; or given a set of scientific articles, which keywords are the most relevant? These questions are examples where we need to discover the most relevant agents (i.e., countries, drugs, diplomatic persons, or keywords) in a complex system composed of agents and the relations between them.

These complex systems often uses networks to represent these relations. To this aim, network theory is an important area that offers solid tools for describing the complex system in different environments such as biology, social networks, information technology, and engineering [1]. Most of these complex systems use graphs to represent these relations and the characteristics between the entities representing the system.

Many problems have historically been modeled with simple graphs, e.g., traveling salesman problem [2], minimum spanning tree [3], etc. And so, these simple situations need only be represented with a single graph.

However, complex systems need more than one graph to be properly represented, and in many cases, a multilayer network is used for these representations. A multilayer network is made up of a set of layers, each one represented by a graph [4],[5]. The nodes in a multilayer network can have different states. For example, a person can be analyzed by her/his friendships, jobs, or relationships in different social networks. Thus, a layer can be described as a set of state nodes and the edges between these state nodes [30].

An initial idea that could be used to represent these complex systems would be to break the system down into independent graphs and analyze each graph. However, not taking the potential dependency into consideration for these graphs can lead to a cascade of failures and a misinterpretation of the system's reality [23]. If the system is represented by a multilayer network composed of a set of layers (each layer representing a portion of the system's information using a graph) then that allows us to describe intra-relationships between members of the same layer as inter-relationships between members of different layers.

Thus, the multilayer network may have intralinks and interlinks between network nodes. For example, in Figure 1 Multilayer Network 2 (MLN2) contains only intralinks while MLN3 has intralinks and interlinks connecting nodes between the layers L1 and L2.
Multilayer networks are used to represent complex systems, such as in traffic control, social networks, or biological systems [23, 24]. Traffic control systems can describe traffic dynamics [6], for example, when passengers are transported by public transport. Every layer can represent the ways in which citizens travel such as by bus, subway, or tram in a city. Also, in air traffic, the layers correspond to flight routes operated by different airline companies.

Recent works, fox example, use complex network to analyse the role of conformist and profiteers players in evolutionary games [25], and to simulate how asymptomatic people can spread COVID-19 with a high level of accuracy [26]. For urban networks, in [17], the PageRank

algorithm (APA) was adapted, providing a model to establish a ranking of nodes in spatial networks according to their importance within it. Furthermore, this model was modified to obtain a measure of the centrality of the nodes in a biplex network [18].

In social networks, data sets need a description to support different relationships between different types of entities, namely, a system with information from researchers, articles, and institutions [8].

In biology, the different interactions in a system can be modeled as a multilayer network, such as interactions between genes and proteins, genes and diseases, or diseases and drugs [7]. Multilayer networks have also made it possible to represent and study intercellular communications in tissues. For example, a disruption in intercellular communications can trigger diseases [24].

In these examples, every type of entity is grouped into a layer.

We can distinguish two types of multilayer networks: multiplex networks and networks with interconnected layers [9]. In multiplex networks, layers have the same set of vertices, and interlayer edges are defined between the same vertices at different layers. A particular case of this is when the network has the same links in each layer, and the only difference is the importance of the entity in each layer. A type of multiplex network is a *temporal multilayer network* in which each node is connected to itself over discrete layers that represent time periods (for example, a multilayer network describing the temporal evolution of Facebook). In contrast, in interconnected networks, the interlayer edges would connect between different entities (for example, documents and researchers)

In Figure 1 MLN1 and MLN2 are multiplex networks, and MLN3 and MLN4 are interconnected networks.

Studies of multilayer networks show the importance of obtaining the centrality nodes [10],[11], [27]. The centrality of a node is defined as a ranking amoung the nodes of a multilayer network. For example, with this information, a user can establish the influencers in a social network, the more effective drugs for different genetic problems, or even the airport where a large number of aerial enterprises are taking off. Computing central nodes is also very important when designing routes in Wireless sensor networks (WSNs) in order to reduce delays and energy consumption, as well as improving the routing overlap [27, 28]. Central nodes are also important when determining the size of a network for different fields such as computer science, social networks, or mathematical modeling, among others [29].

There are different techniques that can be used to get the centrality of a node in multiplex networks without interlinks. Multiplex Pagerank [12] calculates the centrality of a node across the different layers. Other algorithms such as Multiplex Eigenvector Centralities [10] and Functional Multiplex PageRank [11] associate a different influence with the links of different layers that weigh their contribution to the centrality of the nodes. Additionally, classical centrality measures have been redefined to be applied to multiplex networks. Thus, in [13] they redefined the betweenness centrality measure to apply to a multiplex network. In general multilayer networks the algorithms based on Versatility and Communicability can be applied to get the centrality of a node [14].

### 1.1 Main Contribution

This paper proposes a new nodes' centrality measure in a multilayer network. This new centrality measure has been called the "Corner centrality", (the name is inspired by the Harris corner detector for images [15]). The corner centrality algorithm assigns a high centrality value to a node in the target layer when all reference layers recognize that node as a node with high relative importance value. We say that a node in the target layer has a high relative importance from the point of view of a reference layer, when the node has an initial value of importance that stands out over the importance of the nodes of its neighborhood in that reference layer. For example, if we are looking for diplomats to be the negotiators between countries in conflict, these diplomats must be positively recognized by all the countries in conflict. If only one of them does not like that diplomatic person, deals cannot occur.

The initial importance value of each node is an implicit characteristic of the node. For example, in a multilayer network where the relationships between authors who publish in certain scientific journals are represented. In turn, scientific journals are related if they are in the same area of research. The initial level of importance for each author could be their h-index. And the initial level of importance for each journal could be its impact factor.

In the event that this initial information of importance is unknown, an alternative is to take the degree of the node as the initial value of importance.

The Corner Centrality measure can be applied to networks with intralinks and/or interlinks, and the algorithm is independent from the multilayer network structure. Additionally, the Corner Centrality measure can be used with disconnected graphs. Section 3.1.2 will provide an example where we will apply the method on a multilayer network with layers containing disconnected graphs

The only restriction is that two nodes in different layers connected (by an interlink) must have an initial value of importance, and the importance must contain information from the same source.

In particular, we suppose that a set of researchers are related because they are authors of the same paper, and the researchers are grouped per research area, defining a layer. However, the paper's authors can be in different areas and in that case, interlinks would represent their relations. Moreover, the initial importance of these nodes can be, for example, the h-index value of the researchers.

When the multilayer network is a multiplex network, both the initial value of importance of the nodes as well as the relationships between them can be different.

To test the performance of the Corner Centrality measure, we performed three experiments. In the first and second experiments we used multiplex networks. The results were compared with the APABI method [18] and Pagerank versatility method [22]. The APABI method is characterized as an Adapted Pagerank measure that incorporates the informational features of the nodes as the measure of the initial importance. Since the APABI method can only be used on multiplex networks, in the first experiment, we replicated a multiplex network defined in [18] to test our centrality measure. Additionally, the method was compared to the Pagerank versatility method in this first experiment. The Pagerank versatility method expanded on the idea of Google's Pagerank centrality [16] for multilayer networks with interlinks. In this case, we used the Florentine Family multilayer network to compare Pagerank versatility and Corner Centrality. To apply the Corner Centrality to a Florentine Family multilayer network, we took the initial value of importance as the degree of the node.

The second experiment aimed to deduce the most relevant author keywords in the area of Computer Science. To this aim, we used the information from Scopus and Google Trend to define the initial importance of the nodes.

In the third experiment, we created a multilayer network with interlinks between the nodes of different layers. We wanted to get the Corner Centrality measure of author keywords in the "Computer Science" area by using the set of author keywords and KeyWords Plus keywords (from Web of Science). The initial importance of the keywords was the number of documents in which the keywords appeared.

These results allow us to present the research community with a new mechanism to define the centrality or importance of an agent that starts with an initial value of importance and is embedded in a complex system made up of intra-group and/or inter-group relationships.

### 1.2 Structure of the Paper

The paper is organized as follows:

- Section 2 describes the mathematical model used to get the new centrality measure called the "corner centrality" of nodes in a multilayer network. To best understand the model, we have illustrated it with a toy example.
- In Section 3 we applied the model to different multilayer networks. For this, we selected three experiments to test the utility of the Corner Centrality measure.
- Section 4 presents the main conclusions of the paper and new research lines.

## 2 Mathematical model for the corner centrality of nodes in a multilayer network

We consider that a multilayer network $\mathcal{G}$ is composed of a set of layers $\{G_1, G_2, \cdots, G_M\}$. Every layer $G_i = (X_i, E_i)$ is a directed simple graph with $X_i = \{e_{i1}, e_{i2}, \cdots, e_{iN_i}\}$ being the set of nodes, and with $E_i = \{(e_{il}, e_{jk})\}$ being the set of edges such that $l \in \{0,1, \cdots, N_i\}$ and $k \in \{0,1, \cdots, N_j\}$with $N_i$ and $N_j$ being the number of nodes in the graphs $G_i$ and $G_j$, respectively. Different examples of multilayer networks are shown in Figure 1.

Also, every node $u$ in a layer $L$ has an initial importance value $i_L(u)$. For example, in Figure 1, in the multilayer network $MLN1$, the $a$ node in layer one has an importance of 2 while in layer two it has a value of 1. This type of situation can be seen in a team project, when a researcher has a more relevant contribution in one task while in other tasks, the importance of that researcher is not as high.

A node should be perceived with more relative importance in a layer if its importance is higher than the importance of the nodes in its neighborhood in that layer. In this way, we can define the relative importance of a node $u$ in the layer $L$ as:

$$I_L(u) = \sum_{\forall v \in Neighbour_L(u)} i_L(u) - i_L(v) \quad (1)$$

Equation 1 uses the intralinks defined in layer $L$.

On the other hand, we can also define the relative importance in other layers, separate from the layer of the node, by using the interlinks as:

$$I_L(u) = \sum_{\forall v \in Neighbour_L(u)} i_{L*}(u) - i_L(v) \quad (2)$$

With $L$ being different from the node's layer $L_*$. In this case, Equation 2 uses the interlinks defined between the $L$ and $L_*$ layers. $L_*$ is defined as the *target layer* while $L$ is the *reference layer*.

For example, by using Equation 1 for the node $b$, in the multilayer network MLN3 of Figure 1, we obtained the relative importance $I_{L_1}(b)$ and by using Equation 2 we got the relative importance in the layer $L_2$, $I_{L_2}(b)$.

The relative importance of a layer is normalized across the nodes to get a mean and standard deviation value of 0 and 1, respectively.

Finally, when we analyze the multilayer network, the relative global importance of a node will be higher if high relative importance is met in every layer for that node. In this case, we can define the relative global importance of a node as:

$$I(u) = \left(\sum_{i=0}^{M} I_{L_i}(u)\right)^2 \quad (3)$$

In Figure 1 in MLN1, the importance for node $a$ would be: $I(a) = (I_{L_1}(a) + I_{L_2}(a))^2$

To simplify, we are going to suppose that we have a multilayer network with only two layers: $L_1$ and $L_2$. Then Equation 3 would be:

$$I(u) = (I_{L_1}(u) + I_{L_2}(u))^2 \quad (4)$$

Also, if a node has a high global relative importance, this value must be transmitted across the nodes in its neighborhood, in the layer to which it belongs. In this sense, let $L_*$ be the target layer to which the node $u$ belongs, then the $u$ node would have a high global relative importance across all the layers and also in layer $L_*$ the nodes that surround it would also have a high global relative importance:

$$CC_{L_*}(u) = \sum_{v \in Neighbour_{L_*}(u) \cup u} I(v) = \sum_{v \in Neighbour_{L_*}(u) \cup u} (I_{L_1}(v) + I_{L_2}(v))^2 \quad (5)$$

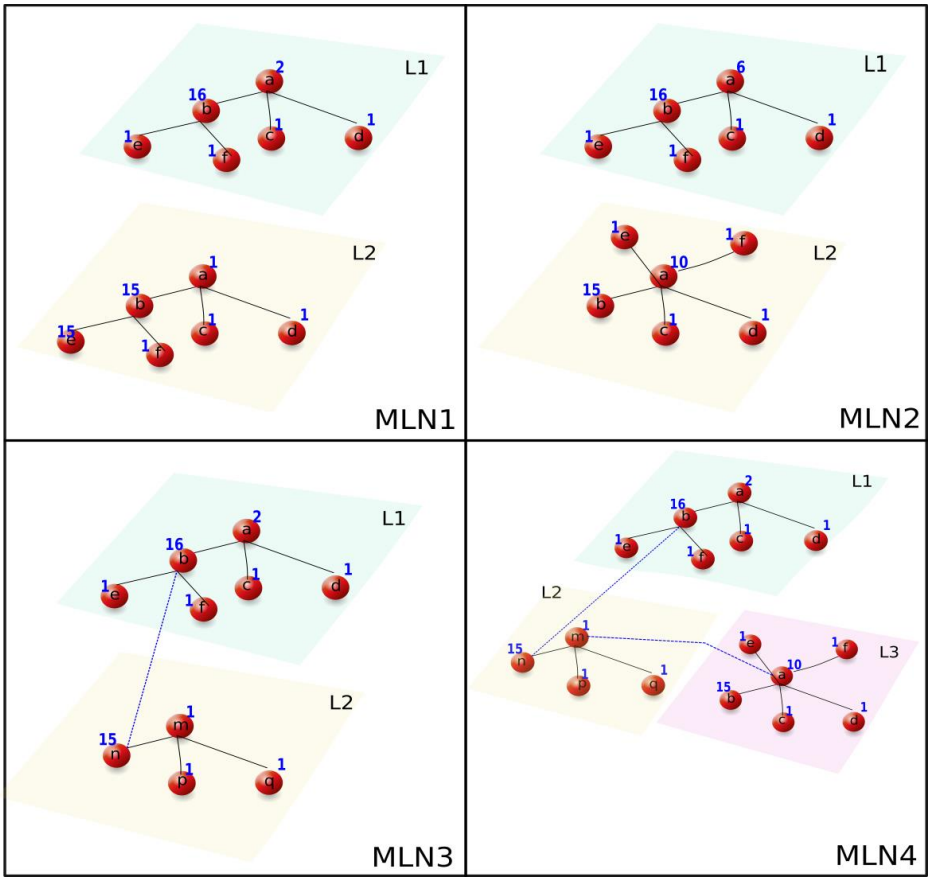



*Figure 1 Examples of different multilayer networks. MLN1 is a multilayer network with two layers that have the same nodes and links. However, the nodes in each layer have different levels of importance. MLN2 has two layers with the same nodes but different links, and the nodes have different levels of importance. MLN3 has two layers with different nodes and links and the two layers are connected by links. MLN4 has three layers.*

In Equation 5, the summation that iterates over the neighbors of $u$ (including u) in the target layer adds to the relative importance of its neighbors in the target layer. Taking into account that $(I_{L_1}(v) + I_{L_2}(v))^2 = I_{L_1}(v)^2 + 2I_{L_1}(v)I_{L_2}(v) + I_{L_2}(v)^2$, Equation 5 can be written in matrix form as:

$$CC_{L_*}(u) = \sum_{v \in Neighbour_{L_*}(u) \cup u} (1, \quad 1) \begin{pmatrix} I_{L_1}(v)^2 & I_{L_1}(v) I_{L_2}(v) \\ I_{L_1}(v) I_{L_2}(v) & I_{L_2}(v)^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (6)$$

$CC_{L*}(u)$ can be rewritten as :

$$CC_{L_*}(u) = (1, \quad 1) M(u) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (7)$$

where M is the structure tensor, and it is defined as:

$$M(u) = \sum_{v \in Neighbour_{L_*}(u) \cup u} \begin{pmatrix} I_{L_1}(v)^2 & I_{L_1}(v) I_{L_2}(v) \\ I_{L_1}(v) I_{L_2}(v) & I_{L_2}(v)^2 \end{pmatrix} \quad (8)$$

Note that the matrix M is derived from the differentials of the initial value of nodes' importance in the target layer with respect to the neighboring nodes' initial importance in the reference layers. For a number of layers $N$ bigger than two, the matrix $M$ is defined as:

$$M(u) = \sum_{v \in Neighbour_{L_*}(u) \cup u} \begin{pmatrix} I_{L_1}(v)^2 & I_{L_1}(v) I_{L_2}(v) & \cdots & I_{L_1}(v) I_{L_N}(v) \\ \vdots & & \ddots & \\ I_{L_N}(v) I_{L_1}(v) & I_{L_N}(v) I_{L_2}(v) & \cdots & I_{L_N}(v)^2 \end{pmatrix} \quad (9)$$

Equation 8 can be rewritten as:

$$M(u) = \begin{pmatrix} \sum_{v \in Neighbour_{L_*}(u) \cup u} I_{L_1}(v)^2 & \sum_{v \in Neighbour_{L_*}(u) \cup u} I_{L_1}(v) I_{L_2}(v) \\ \sum_{v \in Neighbour_{L_*}(u) \cup u} I_{L_1}(v) I_{L_2}(v) & \sum_{v \in Neighbour_{L_*}(u) \cup u} I_{L_2}(v)^2 \end{pmatrix} \quad (10)$$

To summarize the distribution of the relative importance of a node across different layers, we get the eigenvalues $\lambda_1$ and $\lambda_2$ of the matrix $M$.

In the case that $\lambda_2 \ll \lambda_1$ the node has a high relative importance in layer one, but the relative importance in layer 2 is low. The opposite situation is achieved when $\lambda_1 \ll \lambda_2$. In this paper, we want to get the nodes with highest relative importance across all layers, so for this we need $\lambda_1 \approx \lambda_2$ (the two eigenvalues are large and similar in magnitude). Therefore to get the minimum between $\lambda_1$ and $\lambda_2$ we use

$$\lambda_{min} \approx \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} = \frac{det(M)}{trace(M)} \quad (11)$$

Where $det$ and $trace$ are the determinant and trace operators of the matrix M.

For example if $\lambda_1 \gg \lambda_2$ then $\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} = \frac{\lambda_2}{\frac{\lambda_2}{\lambda_1}+1} \approx \lambda_2$. To summarize, our method generates a score for each node in layer $L_*$ as:

$$S_{L_*}(u) = \frac{det(M(u))}{trace(M(u))+\epsilon} \quad (12)$$

where $\epsilon$ is a small constant to avoid divisions by zero. The score $S_{L_*}$ defines the grade of the corner centrality of a node.

---

**Algorithm 1:** Corner Centrality. Let $\mathcal{G} = \{L_1, L_2, \cdots, L_M\}$ a multilayer network with $M$ layers. Let $L_*$ be the target layer from $\mathcal{G}$.

---

1: **for** every node $u$ in $L_*$ **do**
2: **for** every layer $L$ in G **do**
3: **if** $u$ is in $L$ **then**
4: Get $I_L(u)$ by using Equation 1 ▷ by using intralinks within L
5: **else if** $u$ is not in $L$ **then**
6 Get $I_L(u)$ by using Equation 2 ▷by using interlinks between $L_*$ and $L$ layers.
7: **end if**
8: **end for**
9: **end for**
10: **for** every layer $L$ in G **do**
11: Normalize $I_L$
12: **end for**
13: **for** every node $u$ in $L_*$ **do**
14: Get the matrix $M(u)$ by using Equation 9 . ▷with only two layers by using Equation 10
15: Get the score of the node $S_{L*}(u)$ ▷by using Equation 12
16: **end for**
17: Obtain the rank of every node in $L_*$ by sorting $S_{L*}$

---

Algorithm 1 shows the steps to obtain the ranking proposed in this paper. The algorithm can be used for multilayer networks complying with the following conditions:

1. The multilayer network can contain layers with the same set of nodes. In this case, the layers do not have interlinks between them.
2. The multilayer network can contain layers with different nodes. In this case, the layers can be connected using interlinks between them.
3. The multilayer network can contain interlinks and intralinks between layers.

With a multiplex network having every layer with the same set of nodes, without interlinks, the corner centrality value for every node is defined as the minimum value of corner

centrality obtained across the layers. For example, in Figure 1, MLN2 get $S_{L_1}$ and $S_{L_2}$ for the node $a$ and the final score will be the minimum between $S_{L_1}$ and $S_{L_2}$.

The computational time of the algorithm would be:

1. If the number of nodes n in every layer $L$ is bigger than the number of layers $M$, the computational time is $O(n^2 \times \mathrm{M})$. Usually, $n \gg \mathrm{M}$ and $O(n^2 \times \mathrm{M}) \approx O(n^2)$.
2. If the number of nodes n in every layer $L$ is less than the number of layers $M$, the computational time is $O(n * M^{2.373})$.

A more detailed analysis of the computational efficiency has been described in Appendix. Also, Appendix describes the memory needs of the algorithm proposed.

## 2.1 A toy example

In this section, we will show the main steps of Algorithm 1 for the multilayer network MLN4 in Figure 1. If we want to get the Corner Centrality value for the nodes in layer $L_1$,then, according to Algorithm 1 $L_*$ will be $L_1$ . Following the steps 1-9 in Algorithm 1, we obtained the relative importance of every node in $L_1$, analyzed by every layer (see Table 1):

| **Relative Importance for every node** | | | |
|---|---|---|---|
| **Node** | $\boldsymbol{I_{L_1}}$ | $\boldsymbol{I_{L_2}}$ | $\boldsymbol{I_{L_3}}$ |
| a | -12 | 0 | 31 |
| b | 44 | 1 | 5 |
| c | -1 | 0 | -9 |
| d | -1 | 0 | -9 |
| e | -15 | 0 | -9 |
| f | -15 | 0 | -9 |

*Table 1: Relative importance of every node in layer $L1$ from MLN4 in in Figure 1*

To get the relative importance for Layer $L_1$ (see column $I_{L_1}$ in Table 1) for every node in $L_1$ we used the intralinks information. Thus using a fixed node in the L1 layer, we consider its neighbors in $L_1$ and compute the differences of the initial value of importance between the node and its neighbors. In the same way, in Layer $L_3$ we obtained the relative importance by using Equation 1 (see column $I_{L_3}$ in Table 1). But for Layer $L_2$ we used the interlinks information to establish the relative importance of the nodes in $L_1$ for layer $L_2$ by using Equation 2 (see column $I_{L_2}$ in Table 1). In this case, the neighbors of the node in the $L_2$ layer are considered. For example, the node $b$ has a relative importance of 44 in the layer $L_1$. This value was obtained by adding:

$$I_{L_1}(b) = i_{L_1}(b) - i_{L_1}(a) + i_{L_1}(b) - i_{L_1}(e) + i_{L_1}(b) - i_{L_1}(f) = 14 + 15 + 15$$

Concerning layer $L_2$ the relative importance of the node $b$ was obtained using the interlinks between layers $L1$ and $L2$:

$$I_{L_2}(b) = i_{L_1}(b) - i_{L_2}(n) = 16 - 15 = 1$$

And to get the relative importance for the node $b$ in the layer $L_3$ we used the intralinks in layer $L_3$

$$I_{L_3}(b) = i_{L_3}(b) - i_{L_3}(a) = 5$$

Analyzing Table 1, we can see that node $a$ has a high relative importance in layer $L_3$ but very low in layers $L_1$ and $L_2$. Moreover, though node $a$ has a bigger initial importance than the importance of nodes $c$ and $d$ in layer $L_1$ ; this relative importance stays low in regards to the importance of node $b$. Also, we see in Table 1 that for all the layers, node $b$ is essential.

In steps 11-12 we normalized the relative importance by subtracting the mean and divided by the standard deviation. The result of the normalization is shown in Table 2:

| **Normalized Relative Importance for every node** | | | |
|---|---|---|---|
| **Node** | $I_{L_1}$ | $I_{L_2}$ | $I_{L_3}$ |
| a | -0.584 | -0.447 | 2.098 |
| b | 2.142 | 2.236 | 0.338 |
| c | -0.049 | -0.447 | -0.609 |
| d | -0.049 | -0.447 | -0.609 |
| e | -0.730 | -0.447 | -0.609 |

*Table 2: Normalized relative importance of every node in layer $L1$ from MLN4 in Figure 1.*

In steps 12-14, we calculated the Corner Centrallity value for each node in layer $L_1$. In step 13 we got the matrix $M(u)$ which is $3 \times 3$ and symmetrical. We must recall that the relative global importance of a node (see Equation 9) must attend to its relative importance as well as the relative importance of the nodes in its neighborhood within layer $L_1$.

For the node $b$ in layer $L_1$, it is defined as:

$$M(b) = \begin{pmatrix} M_{11}(b) & M_{12}(b) & M_{13}(b) \\ M_{21}(b) & M_{22}(b) & M_{23}(b) \\ M_{31}(b) & M_{32}(b) & M_{33}(b) \end{pmatrix}$$

with

$$M_{11}(b) = I_{L_1}(b)^2 + I_{L_1}(a)^2 + I_{L_1}(e)^2 + I_{L_1}(f)^2$$
$$M_{12}(b) = M_{21}(b) = I_{L_1}(b)I_{L_2}(b) + I_{L_1}(a)I_{L_2}(a) + I_{L_1}(e)I_{L_2}(e) + I_{L_1}(f)I_{L_2}(f)$$
$$M_{13}(b) = M_{31}(b) = I_{L_1}(b)I_{L_3}(b) + I_{L_1}(a)I_{L_3}(a) + I_{L_1}(e)I_{L_3}(e) + I_{L_1}(f)I_{L_3}(f)$$
$$M_{22}(b) = I_{L_2}(b)^2 + I_{L_2}(a)^2 + I_{L_2}(e)^2 + I_{L_2}(f)^2$$
$$M_{23}(b) = M_{32}(b) = I_{L_2}(b)I_{L_3}(b) + I_{L_2}(a)I_{L_3}(a) + I_{L_2}(e)I_{L_3}(e) + I_{L_2}(f)I_{L_3}(f)$$
$$M_{33}(b) = I_{L_3}(b)^2 + I_{L_3}(a)^2 + I_{L_3}(e)^2 + I_{L_3}(f)^2$$

Each value in the matrix evaluates the importance of each node in each layer, and further integrates the importance of the nodes in its neighborhood in the target layer.

Finally, we got the Corner Centrality measure by using Equation 12. For our example, the score was

$$S_{L_1}(a) = 0.345$$
$$S_{L_1}(b) = 0.323$$
$$S_{L_1}(e) = 1.75e - 16$$
$$S_{L_1}(f) = 1.75e - 16$$
$$S_{L_1}(f) = -1.167e - 17$$
$$S_{L_1}(d) = -1.16e - 17$$

This result shows that node $a$ had the highest centrality, followed by $b$. However, node $a$ was not the most important when we analyzed layer by layer. Nevertheless, adding the importance of the nodes in its neighborhood, the $a$'s importance grew because of the node's importance to $b$.

# 3 Experimentation

## 3.1 Experiment 1: Biplex Networks

Here we focus our method on a particular type of multilayer network: biplex networks. A biplex network is composed of two layers, and each layer only has intralinks. In this experiment, we wanted to test the goodness of our method compared to the APABI centrality [18] and Pagerank versatility [22]. The APABI method also uses an initial value of importance for the nodes in the multilayer network but it is only applied to multiplex networks.

The Pagerank versatility can be applied to a multilayer network with intralinks and interlinks, but it does not utilize an initial value of importance of the nodes. Compared with this method, the Corner Centrality value takes the nodes degree as its initial value of importance. Next, we described the behavior of our method for the two different multiplex networks used in [18] and in [22].

### 3.1.1 Football Team

This experiment uses the example proposed in [18]. In this example, a biplex network was analyzed. The biplex is composed of two layers, each one with 20 nodes. Each node represents a player of a football team. The multilayer network has two layers that connect the players differently, and an undirected graph represents each layer. The first layer is constructed with the 20 nodes and the relationships between the team members are analyzed from the point of view of their social or virtual relationships. Thus, two nodes are joined by an edge if they are related or linked through a social network. The initial importance of every node in this layer is the number of messages that each player receives from their teammates within a certain period. With the same 20 players, the second layer shows how the players relate to each other within the game. Thus, two players are connected in the graph if they pass the ball with some consistency during a match. The players' initial importance is the number of games played in a season, seen in layer two.

| Node | Social Network Links | Messages | Game Links | Games |
|---|---|---|---|---|
| 1 | 2, 5, 7, 9, 16, 17, 19, 20 | 15 | 2, 4, 5, 6, 9, 12, 13, 14, 18, 19 | 33 |
| 2 | 1, 5, 7, 9, 20 | 9 | 1, 4, 8, 10, 13, 18, 19 | 26 |
| 3 | 7, 9, 11, 13, 14, 15, 17 | 12 | 4, 5, 6, 12, 14, 15, 17, 20 | 18 |
| 4 | 5, 9, 11, 14, 15, 16, 18, 20 | 19 | 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 20 | 32 |
| 5 | 1, 2, 4, 6, 7, 11, 12, 14, 18, 20 | 28 | 1, 3, 4, 6, 7, 8, 11, 14, 17, 19, 20 | 20 |
| 6 | 5, 7, 10, 20 | 7 | 1, 3, 4, 5, 8, 11, 12, 19 | 12 |
| 7 | 1, 2, 3, 5, 6, 8, 9, 18, 20 | 20 | 4, 5, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20 | 32 |
| 8 | 7, 10, 12, 17, 18 | 7 | 2, 4, 5, 6, 9, 12, 13, 14, 18, 19 | 6 |
| 9 | 1, 2, 3, 4, 7, 10, 15, 18, 20 | 16 | 1, 4, 8, 10, 13, 16, 18, 19 | 18 |
| 10 | 6, 8, 9, 11, 13, 14, 15, 16, 18, 20 | 21 | 2, 4, 7, 9, 13, 15, 18, 19, 20 | 25 |
| 11 | 3, 4, 5, 10, 13, 18, 19, 20 | 14 | 4, 5, 6, 7, 12, 14, 15, 17, 20 | 24 |
| 12 | 5, 8, 14, 17, 20 | 8 | 1, 3, 4, 6, 7, 8, 11, 14, 19, 20 | 18 |
| 13 | 3, 10, 11, 15, 19, 20 | 11 | 1, 2, 4, 7, 8, 9, 10, 16, 17, 19, 20 | 6 |
| 14 | 3, 4, 5, 10, 12, 16, 18, 19 | 13 | 1, 3, 4, 5, 8, 11, 12, 19 | 26 |
| 15 | 3, 4, 9, 10, 13, 17, 20 | 11 | 3, 7, 10, 11, 16, 17, 20 | 38 |
| 16 | 1, 4, 10, 14, 17, 18, 19 | 14 | 4, 7, 9, 13, 15, 18, 19, 20 | 6 |
| 17 | 1, 3, 8, 12, 15, 16, 20 | 12 | 3, 4, 5, 7, 11, 13, 15, 18, 19 | 12 |
| 18 | 4, 5, 7, 8, 9, 10, 11, 14, 16, 19, 20 | 35 | 1, 2, 4, 7, 8, 9, 10, 16, 17, 19, 20 | 30 |
| 19 | 1, 11, 13, 14, 16, 18, 20 | 15 | 1, 2, 5, 6, 7, 8, 9, 10, 12, 13, 14, 16, 17, 18, 20 | 8 |
| 20 | 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 13, 15, 17, 18, 19 | 27 | 3, 4, 5, 7, 10, 11, 12, 13, 15, 16, 18, 19 | 25 |

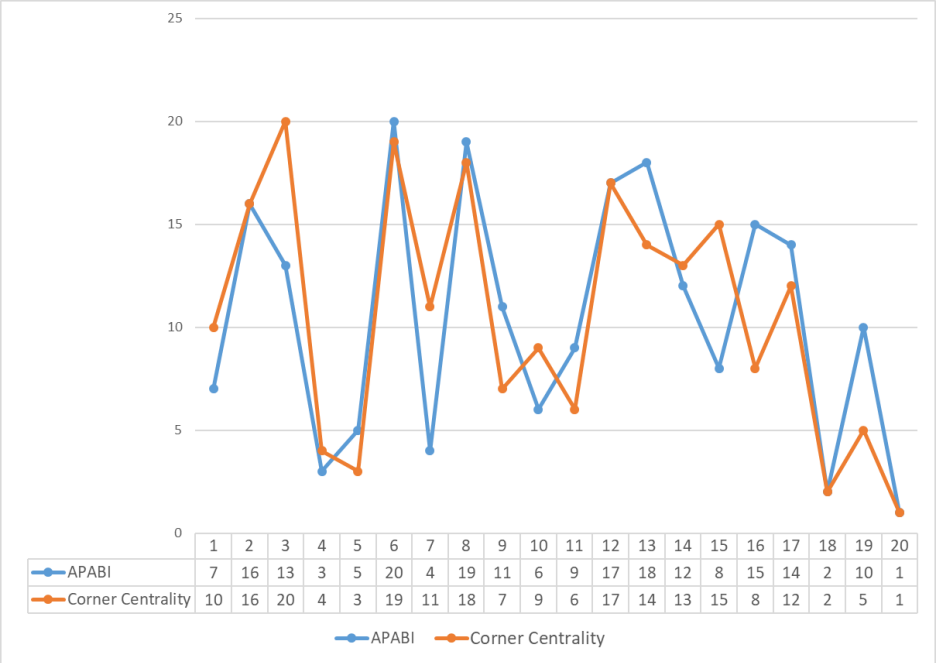
*Table 3:Data associated with the biplex network constructed by the team. This table was presented in [18]*

Table 3 shows the relationships for each node with other nodes (second column) when the social network connection is analyzed. The third column shows the number of messages received by a player in a period. The fourth column shows the relations between players in regards to the number of times they passed the ball to each other. And, the fifth column is the number of games the player participated in during a season.

We have applied our method to this multilayer network in order to get the corner centrality of each node. In this case, the multilayer is a multiplex network with the same set of nodes but without interlinks. In this case, the corner centrality value of every node is defined as the minimum value of corner centrality obtained across the layers. Also, the centrality measure defined for biplex multilayers in [18], called APABI, has been used to compare against our proposed corner centrality measure.

| | **APABI** | | **Corner Centrality** | |
|---|---|---|---|---|
| **Nodes** | **Value** | **Rank** | **Value** | **Rank** |
| 1 | 0.05581 | 7 | 4.309008 | 10 |
| 2 | 0.03777 | 16 | 2.550988 | 16 |
| 3 | 0.04193 | 13 | 1.429319 | 20 |
| 4 | 0.06517 | 3 | 5.171486 | 4 |
| 5 | 0.06440 | 5 | 5.5569 | 3 |
| 6 | 0.02862 | 20 | 1.520039 | 19 |
| 7 | 0.06477 | 4 | 3.883725 | 11 |
| 8 | 0.03071 | 19 | 1.865937 | 18 |
| 9 | 0.04836 | 11 | 4.475997 | 7 |
| 10 | 0.05902 | 6 | 4.412212 | 9 |
| 11 | 0.05013 | 9 | 4.668175 | 6 |
| 12 | 0.03727 | 17 | 1.979066 | 17 |
| 13 | 0.03684 | 18 | 3.198247 | 14 |
| 14 | 0.04820 | 12 | 3.2216934 | 13 |
| 15 | 0.05085 | 8 | 3.068847 | 15 |
| 16 | 0.03781 | 15 | 4.450555 | 8 |
| 17 | 0.04033 | 14 | 3.3716 | 12 |
| 18 | 0.07590 | 2 | 6.405673 | 2 |
| 19 | 0.04838 | 10 | 4.726761 | 5 |
| 20 | 0.07775 | 1 | 8.063785 | 1 |

*Table 4: APABI centrality [18] versus Corner Centrality*




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| APABI | 7 | 16 | 13 | 3 | 5 | 20 | 4 | 19 | 11 | 6 | 9 | 17 | 18 | 12 | 8 | 15 | 14 | 2 | 10 | 1 |
| Corner Centrality | 10 | 16 | 20 | 4 | 3 | 19 | 11 | 18 | 7 | 9 | 6 | 17 | 14 | 13 | 15 | 8 | 12 | 2 | 5 | 1 |

*Figure 2 Comparision between APABI ranking and Corner Centrality ranking for the Football Team multiplex network. In this figure we can see the degree of consensus between APABI and Corner Centrality. According to APABI, the leaders of the group are nodes 20 and 18, showing an overlapwith the Corner Centrality measure.*

In Table 4 the value of centrality and rank for each node for both the APABI method and our proposed method is presented. Also, in Figure 2 we have illustrated the ranking for both methods.

The APABI centrality shows that the nodes classified as the highest values of centrality, the leaders of the group, are nodes 20 and 18. On the one hand, node 20 was third in messages received and eighth in the number of games in the season. On the other hand, node 18 was the node that received more messages but was fifth among players that played games in the season. Corner Centrality also established that the nodes with the highest values of importance were: first place node 20 and second place node 18.

Our method generated a ranking similar to the APABI ranking. The main disadvantage of the APABI method is the tremendous amount of memory that it needs, as well as the method's susceptibility to the parameter $\alpha$.

### 3.1.2 Florentine Families

In this experiment, we applied our method to the Florentine Families Multiplex Network [19]. This multiplex network is composed of two layers. One layer describes the business dealings between sixteen florentine families in the XV century, and the other layer illustrates their alliances due to marriage. Figure 3 shows the married and business relationships between the Florentine families. In Table 5 the code associated with the name of every family

member is presented. To apply the Corner Centrality measure to this network, we assigned the node's degree value as the node's initial importance value in every layer. The method was then compared to the Pagerank versatility [22].

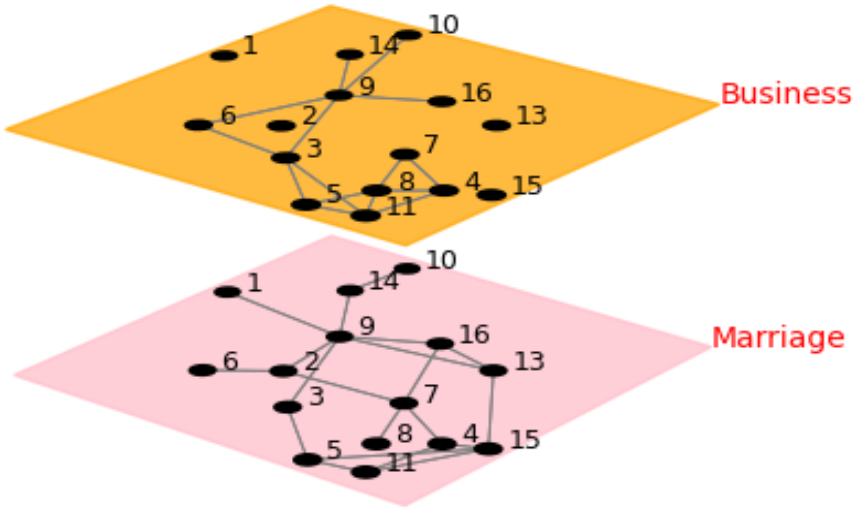



*Figure 3 This figure illustrate the Florentine Families multiplex network. For the experiment described in section 3.1.2, we built a multiplex network using information about business alliances in one layer, and marriage in the another one.*

| | | Corner Centrality | | PageRank versatility | |
|---|---|---|---|---|---|
| Family Name | Node Id | Value | Rank | Value | Rank |
| ACCIAIUOL | 1 | 0 | 15 | 0.8638167 | 6 |
| ALBIZZI | 2 | 0 | 12 | 0.8389033 | 13 |
| BARBADORI | 3 | 0.87675385 | 2 | 0.8567245 | 8 |
| BISCHERI | 4 | 0.250263979 | 5 | 0.8732232 | 4 |
| CASTELLAN | 5 | 0.268755428 | 4 | 0.8557555 | 7 |
| GINORI | 6 | 0.083220632 | 8 | 0.8347692 | 15 |
| GUADAGNI | 7 | 0.223183775 | 6 | 0.8823537 | 3 |
| LAMBERTES | 8 | 0.109123972 | 7 | 0.8561786 | 9 |
| MEDICI | 9 | 1.229130987 | 1 | 1 | 1 |
| PAZZI | 10 | 0.041761825 | 10 | 0.830739 | 16 |
| PERUZZI | 11 | 0.278145671 | 3 | 0.8975903 | 2 |
| PUCCI | 12 | 0 | 16 | 0.8638167 | 5 |
| RIDOLFI | 13 | 0 | 13 | 0.8370955 | 14 |
| SALVIATI | 14 | 0.071072416 | 9 | 0.8374698 | 12 |
| STROZZI | 15 | 0 | 14 | 0.8496274 | 10 |
| TORNABUON | 16 | 0.025402425 | 11 | 0.8430032 | 11 |

*Table 5: Corner Centrality values and Pagerank versatility for the Florentine Families multiplex network*

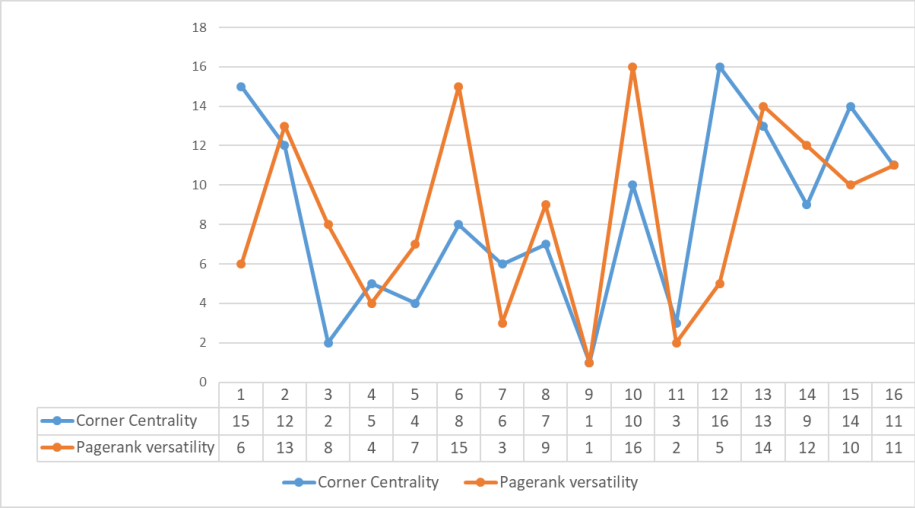



*Figure 4: Comparision between Pagerank versality ranking and Corner Centrality ranking for the Florentine Families multiplex network. In this comparison, we wanted to check the level of consensus between Corner Centrality and PageRank versatility measures. Both measures selected node 9 as the one with the highest centrality values.*

In Table 5 and Figure 4 the value and ranking of the Corner Centrality and Pagerank versatility methods are shown. For both methods, the Medici family (node 9 in Figure 4) is the node with the highest centrality value. Unlike PageRank versatility, the Corner Centrality method assigns a zero value to a node with an initial value of importance equal to zero in at least one layer. This is so because Corner Centrality looks for those nodes that maintain high importance throughout all the layers. It is a logical concept, for example, when we need an expert in different fields or a peace mediator in the different countries of a conflict.

### 3.2 Experiment 2: Scopus vs Google Trend

In this experiment, our objective was to analyze the importance of a set of author's keywords. The set comes from a corpus of 69,000 documents as a result of a search in the Web of Science, using "Computer Science" as the search criteria but without applying any filters to the documents, neither by type nor by area, to obtain academic literature that discusses computer science but is not limited exclusively to this area.

We obtained two values for each keyword in this dataset: Scopus and Google. According to the Scopus Database, the Scopus value refers to the number of articles that contain specific keywords. Thanks to the Scopus value, we can get information about how frequently researchers use a keyword.

Concerning the Google value, we used the information provided by Google Trends to extract the global popularity of a term. We used trends data from 2020 to 2021. The Google value can help us to understand the social popularity of a keyword and the use of that keyword outside the world of academia. We built our dataset by using the unofficial PyTrends API.

To obtain the value for global popularity, the average popularity of the results for each of the 250 countries for which Google provides information was calculated. Then, the Scopus

search API was used to find the number of articles that contained a particular word so as to get the Scopus value.

Many authors' keywords were not presented in Scopus or Google Trend. In this case, the author keyword was erased. Thus the dataset of keywords consisted of 27,704 words.

In this experiment, we built a multiplex network with the author's keywords as the nodes. Thus, we would say that a link existed between two keywords if they appeared in the same paper. The multiplex layer had two layers. The layers had the same nodes and same links, but the initial importance of every node was the Google Trend score in the first layer and the Scopus score in the second layer.

Let $MN_{ak}$ be this multiplex network. To compare our centrality score, Corner Centrality, with APABI centrality [18] we obtained different subnetworks from $MN_{ak}$. The number of nodes for the subnetworks were $\{100, 150, 200, \cdots 1000\}$. For each size, we generated a set of thirty subnetworks from $MN_{AK}$, where the nodes were chosen randomly. In Table 6, we computed the mean value of Spearman's Correlation Rank between Corner Centrality and the APABI method for each set, and we show the p-value of the correlation. A p-value close to 0 was obtained for every set. Therefore, a strong correlation exists between the Corner Centrality and APABI methods.

| | **Spearman's Rank-Order Correlation** | | **p-value** | |
|---|---|---|---|---|
| Nodes | Mean | Std | Mean | Std |
| 100 | 0.604 | 0.065 | 0.000 | 0.000 |
| 150 | 0.691 | 0.041 | 0.000 | 0.000 |
| 200 | 0.683 | 0.039 | 0.000 | 0.000 |
| 250 | 0.673 | 0.036 | 0.000 | 0.000 |
| 300 | 0.705 | 0.017 | 0.000 | 0.000 |
| 350 | 0.697 | 0.028 | 0.000 | 0.000 |
| 400 | 0.729 | 0.022 | 0.000 | 0.000 |
| 450 | 0.753 | 0.02 | 0.000 | 0.000 |
| 500 | 0.752 | 0.018 | 0.000 | 0.000 |
| 550 | 0.769 | 0.014 | 0.000 | 0.000 |
| 600 | 0.758 | 0.024 | 0.000 | 0.000 |
| 650 | 0.769 | 0.022 | 0.000 | 0.000 |
| 700 | 0.763 | 0.015 | 0.000 | 0.000 |
| 750 | 0.756 | 0.019 | 0.000 | 0.000 |
| 800 | 0.753 | 0.02 | 0.000 | 0.000 |
| 850 | 0.772 | 0.013 | 0.000 | 0.000 |
| 900 | 0.756 | 0.022 | 0.000 | 0.000 |
| 950 | 0.769 | 0.021 | 0.000 | 0.000 |
| 1000 | 0.765 | 0.025 | 0.000 | 0.000 |

*Table 6: Comparision between APABI centrality and Corner Centrality. For each set of networks (row), the number of nodes, the Spearman's Correlation Rank mean value, and the standard deviation across the set is shown. Also, the mean values of the p-value and standard deviations are presented. Every set had 30 networks with the number of nodes given in the first column. The nodes were chosen from the $MN_{AK}$ multiplex network.*

### 3.3 Experiment 3: Author Keywords vs. KeyWords Plus Keywords

In this experiment, we wanted to obtain the Corner Centrallity value of a set of author keywords using the relationship with the KeyWords Plus keywords from Web of Knowledge. Typically, authors select keywords, so we will call these words 'author keywords'. Although it is more likely that authors do not have the freedom to choose keywords and automatic algorithms and predefined categories are used instead [20]. KeyWords Plus is one such alternative to author keywords. KeyWords Plus keywords are automatically generated from the titles of the articles that are referenced in a paper. According to [21], KeyWords Plus tries to reduce the problems generated by letting authors select their own keywords.

In the same way as Experiment 2, we got a set of author keywords from a corpus of 69,000 documents as a result of a search in Web of Science, using "Computer Science" as the search criteria but without applying any filters to the documents, neither by type nor by area, to obtain academic literature that discusses computer science but is not limited exclusively to this area. The set of author's keywords was composed of 48,115 words, and the set of KeyWords Plus keywords contained 24,040 words. In this multilayer network, we had two layers. The first layer was composed of the author keywords as nodes, and between the nodes a link existed if they appeared in the same paper. The initial importance of the nodes was the number of papers in which the author keywords appeared. In the second layer, the nodes were the KeyWords Plus keywords.

Similarly to the other experiments, the nodes had an initial importance equal to the number of papers where they appeared. In this second layer, two nodes were connected if they were in the same paper. Additionally, between nodes in layer 1 (author keywords) and nodes in the second layer (KeyWords Plus) a link could exist if they shared the same paper. Our aim in this multilayer was to obtain the Corner Centrality value of the set of author keywords. In this case, an author keyword would have a higher centrality if the relative importance was higher in the layer of author keywords and the layer of KeyWords Plus keywords.

In Figure 5, on the left the 50 authors' keywords that appeared in the highest number of papers are presented. Moreover, the 50 author's keywords with the highest Corner Centrality are on the right. In these wordclouds, the size of the word is related to the value given. Therefore, the ranking given by Corner Centrallity distinguishes the words with high importance, such as education (first place), machine_learning (second place), and computer_science (third place).

Likewise, we obtained the KeyWords Plus keywords with the highest centrality and compared them with the KeyWords Plus keywords most frequently used in the papers. Thus, in Figure 6 on the left, the 50 KeyWords Plus keywords with the highest frequency are shown, and on the right, the 50 KeyWords Plus keywords with the highest values of Corner Centrality are shown.

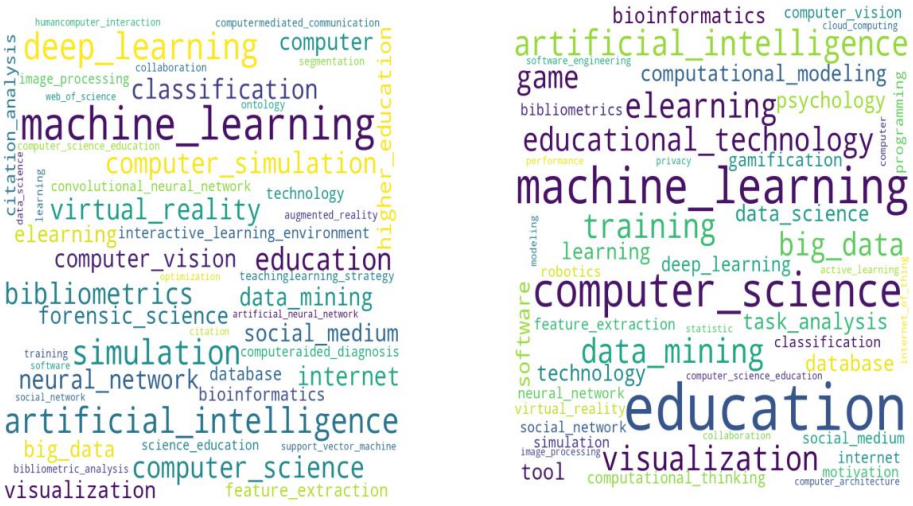



*Figure 5: Two word clouds that compare results of the Corner Centrality value with the 50 most frequent keywords. Left: The 50 author keywords with the highest frequency. Right:The 50 author keywords with the highest Corner Centrallity value.*

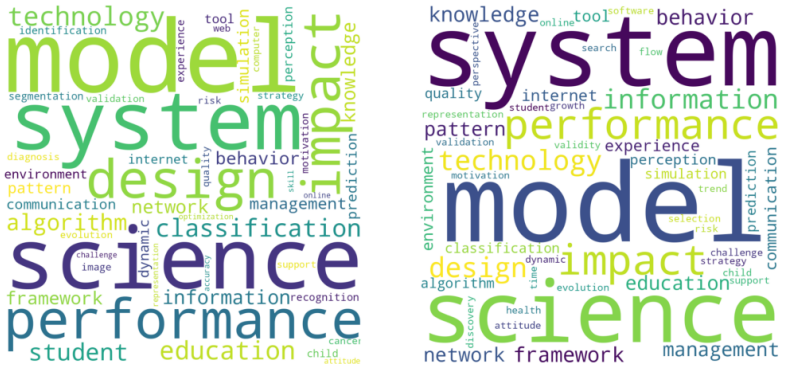



*Figure 6: Two word clouds that compare results of the Corner Centrality value with the 50 most frequent keywords. Left: The 50 KeyWords Plus keywords with the highest frequency. Right:The 50 Keywords Plus keywords with the highest Corner Centrallity value.*

### 3.3.1 Validation

To validate the Corner Centrality measure, we compared it with the Pagerank versatility method [22]. However, the Pagerank versatility method has restrictions on the amount of memory that a multilayer needs. Thus, we generated different subnetworks from the author keywords and KeyWords Plus keywords (AK-KP multilayer network). All of them were undirected, containing intralinks, interlinks, and two layers: author keywords and

KeysWords Plus keywords. The multilayer networks had 50, 100, 200, and 300 nodes. And for every set of nodes, we obtained five multilayer networks. This dataset can be downloaded at https://www.kaggle.com/datasets/jorgechamorropadial/author-keywords-keywordsplus .

For our method, the initial value of the importance of every node is its degree, recalling that a link between two keywords means that they appeared in the same paper.

We created five sets of multilayer networks composed of 50, 100, 200, and 300 nodes. Every multilayer network obtained the node's ranking given by the Pagerank versatility method and the Corner Centrality measure. To test the correlation between the two rankings, we applied Spearman's Rank Correlation, finding the correlation ($c$) and how likely or probable it was that any observed correlation was due to chance ($p$).

In Figure 7 and Table 7 the correlation ($c$) and the p-value ($p$) are presented. Thus, for $Set1$ we obtained a minimum correlation of 0.70 for the network with 300 nodes and 0.82 for the network with 200 nodes. For $Set2$ the minimum value of correlation was achieved for the network of 50 nodes with a value of 0.71, and the maximum value was obtained for the network with 200 nodes (0.82). For $Set3$ we obtained a minimun and maximum correlation of 0.72 and 0.85, respectively. For $Set4$ we obtained the values of 0.64 and 0.80 as the minimum and maximum values of correlation, respectively. And finally, for $Set5$ we acheived a minumum value of 0.68 and a maximum value of 0.82. For all cases shown in Table 7, the probability that an observed correlation was due to chance($p$) is close to zero. The average correlation ($\hat{c}$) and the average probability p-value ($\hat{p}$) across the multilayer networks with the same number of nodes is shown in Table 8.

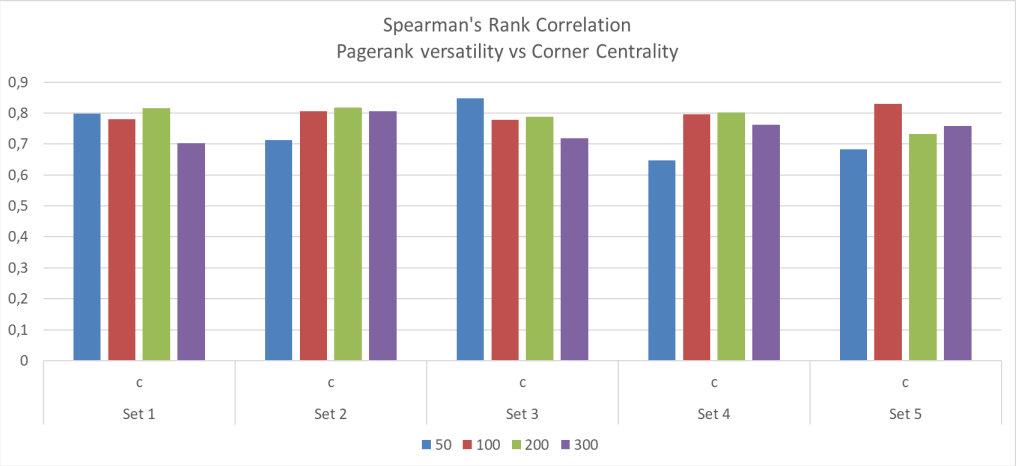



*Figure 7: Spearman's Rank Correlation between the centrality measures: Pagerank versatility and Corner Centrality. The multilayer networks compared have 50, 100, 200, and 300 nodes. For every set of nodes, we generated five multilayer networks and obtained the correlation and significance values (c and p) for each.*

| | Set 1 | | Set 2 | | Set 3 | | Set 4 | | Set 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\#Node$ | c | $p$ | c | $p$ | c | $p$ | c | $p$ | c | $p$ |
| 50 | 0.7977 | 4e-12 | 0.7130 | 6e-09 | 0.8475 | 8e-15 | 0.6463 | 3e-07 | 0.6835 | 4e-08 |
| 100 | 0.7798 | 1e-21 | 0.8048 | 6e-24 | 0.7784 | 1e-21 | 0.7951 | 5e-23 | 0.8295 | 1e-26 |
| 200 | 0.815 | 6e-49 | 0.8171 | 2e-49 | 0.7878 | 1e-43 | 0.8026 | 2e-46 | 0.7327 | 6e-35 |
| 300 | 0.7019 | 7e-46 | 0.8066 | 4e-70 | 0.7190 | 5e-49 | 0.7627 | 2e-58 | 0.7582 | 2e-57 |

*Table 7: Spearman's Rank Correlation between the centrality measures of the Pagerank versatility and Corner Centrality (c) and p-value (p)*

| | 50 | 100 | 200 | 300 |
|---|---|---|---|---|
| $\hat{c}$ | 0.7376 | 0.7975 | 0.7911 | 0.7497 |
| $\hat{p}$ | 8e-08 | 5e-22 | 1e-35 | 1e-46 |

*Table 8: $\hat{c}$= Average Spearman's Rank Correlation between the centrality measures: Pagerank versatility and Corner Centrality, across the multilayer networks with the same number of nodes. $\hat{p}$ = Average p-value (p) across the multilayer networks with the same number of nodes.*

## 4 Conclusions

In this paper, we have presented a new method to obtain the centrality of nodes in a multilayer network. The method can be applied to multilayer networks with nodes that have an initial value of importance. If the initial value of importance is not given, the method will use the degree of the node.

Our method deals with multilayer networks, which can be multiplex networks and multilayer networks with interlinks. The primary constraint is that the nodes in different layers with interlinks have an initial value of importance with information of the exact nature related to their source. For example, consider a multilayer network with different layers, one per area and the nodes are authors publishing in that area. The initial importance of the nodes (authors) can be their h-index. A paper published by different authors generates a link between them in the network. The authors can be in the same area (layer) or different areas (different layers), but the nodes' initial value of importance in different layers with interlinks between them is the same information (h-index value).

Two hypotheses support this method: 1) A node will be more important when all the layers establish that the node has a high relative importance. 2) The centrality of a node will be higher if nodes in its neighborhood are essential. To meet these two hypotheses, we presented a new algorithm with a low computational time. It is also important to note that the memory needs are very low compared to other algorithms, and the algorithm does not need to adjust any input parameter.

To test the performance of our method, we compared it with the APABI centrality and the Pagerank versatility methods by carrying out three experiments. The first experiment was applied to two small biplex networks. In both biplexes, the nodes in both layers were the same,

but the links were different

The second experiment was also applied to a set of biplex networks with several nodes per layer. This experiment used the relations between author keywords in the context of scientific papers. We aimed to get the author's keywords with the highest centrality in this multiplex network. We characterized the nodes with Scopus and Google Trends values.

For the third experiment, we used author keywords and KeyWords Plus keywords (from Web of Knowledge), which was defined as a multilayer network with interlinks between two layers. Also, in this experiment, we wanted to determine the author's keywords with the highest centrality. In this case, the keywords were characterized by the number of papers in which they appeared. The results of the Corner Centrality method were then compared with the results of the Pagerank versatility method.

In every experiment the results were outstanding.

As a line of future research, we would like to analyze the possibility of weighing the importance given by every layer to the nodes in the target layer. Furthermore, we want to study how the Corner Centrality of a node is affected when the importance of the nodes in the neighborhood is weighed.

## Source code

The source code of our experiments can be found in a GitHub repository located at https://github.com/rosadecsai/Corner-Centrality.git

## Author's Biography

**Rosa Rodríguez-Sánchez.** Associate Professor at the University of Granada**.** She received a B.S and a Ph.D. in Computer Science from the University of Granada, Spain in 1996 and 1999, respectively. Her research lines are 1) models of image representation; 2) image and video compression; and 3) science metrics and research evaluation.

Rosa has obtained results in: 1) the visual distinction of objects in biomedical applications and military objectives; 2) explanation of certain illusory forms; and 3) development of visual attention models. The results in this line gave rise to different publications in JCR journals of the highest prestige and the publication of academic books. In the second line of research, models for the transmission of images were developed in order to transmit the most visually relevant information compared to the rest.

**Jorge Chamorro-Padial**. PhD Student at the University of Granada. He received a B.S degree in computer science from the University of Granada and a M.S degree in Cognitive Systems and Interactive Media from the Universitat Pompeu Fabra. Since 2016, he has worked as Software Developer for T-Systems (Deutsche Telekom) where he is also responsible for training employees in Artificial Intelligence and Big Data. His research interests include Bibliometrics, Text Mining, Computational Sociology and Human-Computer Interaction.

## Conflict of Interest

The authors declare no conflicts of interest.

## References

[1] Newman, M. *Networks: An Introduction*; Oxford University Press: Oxford, UK, 2010.

[2] Applegate, D. L.; Bixby, R. M.; Chvátal, V.; Cook, W. J. *The Traveling Salesman Problem*; Princeton University Press: Princeton, US, 2006, ISBN 978-0-691-12993-8.

[3] Gabow, H. N.; Galil, Z.; Spencer, T.; Tarjan, R. E. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica* 1986. 6 (2): 109. doi:10.1007/bf02579168. S2CID 35618095.

[4] Cellai, D.; Bianconi, G (2016). Multiplex networks with heterogeneous activities of the nodes. Phys. Rev. ,93, 032302.

[5] De Domenico, M.; Solé-Ribalta, A.; Cozzo, E.; Kivelä, M.; Moreno, Y.; Porter, M.; Gómez, S.; Arenas, A. Mathematical Formulation of Multilayer Networks . *Physical Review X* 2013. 3 (4): 041022. doi:10.1103/PhysRevX.3.041022.

[6] J. Wu; C. Pu; L. Li.; G. Cao. Traffic dynamics on multilayer networks. *Digital Communications and Networks* 2020, ISSN: 2352-8648, Vol: 6, Issue: 1, Page: 58-63

[7] Lv, Y.; Huang, S.; Zhang, T.;Gao, B. Application of Multilayer Network Models in Bioinformatics. *Frontiers in Genetics* 2021, 12:664860. doi:10.3389/fgene.2021.664860

[8] McGee, F.; Ghoniem, M.; Melançon, G.; Otjacques, B.; Pinaud. The State of the Art in Multilayer Network Visualization. *Computer Graphics Forum* 2019, doi:10.1111/cgf.13610.

[9] Kinsley, AC.; Rossi, G; Silk, MJ.; VanderWaal, K. Multilayer and Multiplex Networks: An Introduction to Their Use in Veterinary Epidemiology. *Frontiers in Veterinary Science* 2020, 7:596. doi:10.3389/fvets.2020.00596

[10] Sola, L.; Romance, M.; Criado, R.; Flores, J.; Garcia del Amo, A.; Boccaletti, S. Eigenvector centrality of nodes in multiplex networks. *Chaos* 2013, 23, 033131.

[11] Iacovacci, J.; Rahmede, C.; Arenas, A.; Bianconi, G. Functional Multiplex PageRank. *EPL* 2016, 116 28004.

[12] Halu, A.; Mondragon, R.; Panzarasa, P.; Bianconi, G.(2013). Multiplex PageRank. PLoS ONE 8, e78293

[13] Solé-Ribalta, A.; De Domenico, M.; Gómez, S.; Arenas, A. Centrality Rankings in Multiplex Networks, Proceedings of the 2014 ACM Conference on Web Science, Bloomington, IN, USA, 23–26 June 2014; ACM:New York, NY, USA, ; pp. 149–155.

[14] Bianconi, G. Multilayer Networks. Structure and Functions; Oxford University Press: Oxford, UK, 2018

[15] Harris, C.; ,Stephens, M. A Combined Corner and Edge Detector. Alvey Vision Conference 1988. Vol. 15.

[16] Brin, S.; Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems 1998, 30(1-7), 107-117

[17] Agryzkov, T.; Oliver, J.L.; Tortosa, L.; Vicent, J.F. An algorithm for ranking the nodes of an urban network based on the concept of PageRank vector. *Applied Mathematics and Computation* 2012,219, 4, 2186-2193 doi:10.1016/j.amc.2012.08.064

[18] Agryzkov, T.; Curado, M.; Pedroche, F.; Tortosa, L.; Vicent, J.F. Extending the Adapted PageRank Algorithm Centrality to Multiplex Networks with Data Using the PageRank Two-Layer Approach. *Symmetry* 2019, 11, 284. doi:10.3390/sym11020284

[19] Padgett, JF.; Ansell, CK. Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology* 1993, 98(6):1259–1319. doi:10.1086/230190

[20] Lu, W.; Liu, Z.; Huang, Y.; Bu, Y.; Li, X.; Cheng, Q. How do authors select keywords? A preliminary study of author keyword selection behavior. *Journal of Informetrics* 2020, 14(4), 101066. doi:10.1016/j.joi.2020.101066.

[21] Garfield, E.; Sher, I. H. KeyWords Plus. Algorithmic Derivative Indexing. *Journal of the American Society for Information Science* 1993, 44(5), 298-299,

[22] De Domenico, M.; Solé-Ribalta, A.; Omodei, E.; Gómez, S.; Arenas, A. Ranking in interconnected multilayer networks reveals versatile nodes. Nature communications 2015, 6:6868. doi: 10.1038/ncomms7868. PMID: 25904405.

[23] Moreno, Y.; Perc, M. Focus on multilayer networks. *New Journal of Physics* 2019, *22*(1), 010201.

[24] Gosak, M., Markovič, R., Dolenšek, J., Rupnik, M. S., Marhl, M., Stožer, A., Perc, M. . Network science of biological systems at different scales: A review. *Physics of life reviews* 2018, 24, 118-135

[25] Pi, B.; Zeng, Z.; Feng, M.; Kurths, J. Evolutionary multigame with conformists and profiteers based on dynamic complex networks. *Chaos* 2022: An Interdisciplinary Journal of Nonlinear Science, 32(2), 023117.

[26] Stella, L.; Martínez, A.P.; Bauso, D.; Colaneri, P. The role of asymptomatic infections in the COVID-19 epidemic via complex networks and stability analysis. *SIAM Journal on Control and Optimization* 2022, 60(2), S119-S144

[27] Sharifi, S. S.; Barati, H. A method for routing and data aggregating in cluster-based wireless sensor networks. *International Journal of Communication Systems* 2021, 34(7), e4754.

[28] Oliveira, E. M.; Ramos, H. S.; Loureiro, A. A. Centrality-based routing for wireless sensor networks. IEEE, 2010 IFIP Wireless Days, 2010, 1-5.

[29] Kenyeres, M.; Kenyeres, J. Comparative Study of Distributed Consensus Gossip Algorithms for Network Size Estimation in Multi-Agent Systems. *Future Internet* 2021, 13(5):134. doi:10.3390/fi13050134

[30] Bazzi, M.; Lucas, G.; Jeub, S.; Arenas, A.; Howison, S. D. ; Porter, M. A. A framework for the construction of generative models for mesoscale structure in multilayer networks. *Physical Review Research* 2020. doi:10.1103/PhysRevResearch.2.023100

[31] Aho, Alfred V.; Hopcroft, John E.; Ullman, Jeffrey D. .The Design and Analysis of Computer Algorithms. 1974 Addison-Wesley, Theorem 6.6, p. 241

## Appendix

### Computational Time

The analysis carried out below assumes that the information on the neighbors of a particular node is obtained in lineal time n. n is the number of nodes in each layer.

The information *about* the node's neighbors is needed in Equations 1, 2 and 10.

To get the computational time of Algorithm 1, we have analyzed the computation time in every step.

---

**Algorithm 1:** Corner Centrality. Let $\mathcal{G} = \{L_1, L_2, \cdots, L_M\}$ a multilayer network with $M$ layers. Let $L_*$ be the target layer from $\mathcal{G}$.

---

1: **for** every node $u$ in $L_*$ **do**
2: **for** every layer $L$ in G **do**
3: **if** $u$ is in $L$ **then**
4: Get $I_L(u)$ by using Equation 1 ▷ by using intralinks within L
5: **else if** $u$ is not in $L$ **then**
6 Get $I_L(u)$ by using Equation 2 ▷by using interlinks between $L_*$ and $L$ layers.
7: **end if**
8: **end for**
9: **end for**
10: **for** every layer $L$ in G **do**
11: Normalize $I_L$
12: **end for**
13: **for** every node $u$ in $L_*$ **do**
14: Get the matrix $M(u)$ by using Equation 9 . ▷with only two layers by using Equation 10
15: Get the score of the node $S_{L*}(u)$ ▷by using Equation 12
16: **end for**
17: Obtain the rank of every node in $L_*$ by sorting $S_{L*}$

---

- In steps 1-8, the computational time is calculated by:

$$\sum_{i=1}^{n}\sum_{j=1}^{M} n = n^2 \times M$$

With $n$ the number of nodes in the layer $L_*$, $M$ the number of layers. We have assumed that each layer has the same number of nodes n. Within the second summation, we have $n$ which is the time consumed by the search for the neighbors of a node in the jth layer.

- In steps 10-12, the computational time is $n \times M$

- In steps 13-16, the computational time is $\mathrm{n} \times (\mathrm{M}^2 + \mathrm{n} + \mathrm{M}^{2.373} + \mathrm{M})$. $\mathrm{M}^2 + \mathrm{n}$ is the time to create the matrix in step 14. $\mathrm{M}^{2.373}$ is the computational time to apply the determinant for a matrix $\mathrm{M} \times \mathrm{M}$ [31]. And M is the computational time to get the trace for a matrix $\mathrm{M} \times \mathrm{M}$.
- And in step 17, the computational time is $\mathrm{n} \times \log_2 \mathrm{n}$

Adding all the terms

$$\mathrm{n}^2 \times \mathrm{M} + 2 \times \mathrm{n} \times \mathrm{M} + \mathrm{n}^2 + \mathrm{n} \times \mathrm{M}^2 + \mathrm{n} \times \mathrm{M}^{2.373} + \mathrm{n} \times \log_2 \mathrm{n} \quad (13)$$

As can be seen in Equation 13, the computational time is a function of the number of nodes in each layer (n) and the number of layers (M).
In the case that $\mathrm{M} \ll \mathrm{n}$ the computational time in the worst case is $\mathrm{O}(\mathrm{n}^2 \times \mathrm{M})$.
However, if $\mathrm{n} \ll \mathrm{M}$ the computational time in the worst case is $\mathrm{O}(\mathrm{n} \times \mathrm{M}^{2.273})$.
Usually, the number of layers is much less than the number of nodes; therefore, the computational time is set according to the number of nodes.

**Memory Requirements**

Algorithm 1 has as input for every layer an incidence matrix $A_{ij}^{\alpha\alpha}$, being $\alpha$ the layer and $i$ and $j$ the nodes in the layer. Thus $A_{ij}^{\alpha\alpha} = 1$ if there is an edge between $i$ and $j$ and $A_{ij}^{\alpha\alpha} = 0$ otherwise. Between two different layers $\alpha$ and $\beta$, Algorithm 1 needs as input an incidence matrix $A_{ij}^{\alpha\beta}$ with value one if between the node $i$ in layer $\alpha$ and node j in layer $\beta$ there is an edge. Also, Algorithm 1 takes as input the initial value of importance for every node in the multilayer network.

Analyzing the body of Algorithm 1, in steps 1-9 it is generated the arrays $I_L$ for every layer L in the multilayer network and every node $u$ in the target layer. In these steps, the memory requirement is $n \times M$ (with $n$ the number of nodes in the target layer and $M$ the number of layers). Every element in array $I_L$ stores a float.

In step 14 the temporal matrix $M(u)$ occupies $M \times M$. From this matrix, Algorithm 1 in step 15 calculates the node's score, which is stored in the array $S_{L*}$. $S_{L*}$ is n-dimensional. In summary, without taking into account the inputs' memory requirements, Algorithm 1 needs $n \times M + M \times M + n$ objects of type float.

# Capítulo 5

# Otras publicaciones

## 5.1. Clasificación de texto. Utilizando métricas de ganancia de información para categorizar disposiciones legales

### 5.1.1. Datos generales

1. **Autores**: Jorge Chamorro-Padial, Rosa Rodríguez-Sánchez.
2. **Revista**: Revista Internacional de Tecnología, Conocimiento y Sociedad.
3. **Datos sobre la publicación**:
   - **Referencia**: Chamorro-Padial y Rodríguez-Sánchez (2019).
   - **Volumen**: 7.
   - **Número**: 2.
   - **Páginas**: 37-48.
   - **Año**: 2019.
   - **Editorial**: Common Ground.
   - **DOI**: https://doi.org/10.18848/2474-588X/CGP/v07i02/37-48.
4. **Estado**: Publicado.
5. **Métricas**:
   - Indexada en:
     - Matriz de Información para el Análisis de Revistas (MIAR, 2022).
     - Academic Search Premier (EBSCO).
     - Fuente Académica Plus (EBSCO).

- Sherpa Romeo.
- Directory of Open Access Resources (ROAD).
- Directory of Open Access Journals (DOAJ).

### 5.1.2. Contribuciones principales

- Clasificamos documentos jurídicos escritos en español con unos resultados de precisión y exactitud muy elevados.
- Comparamos el rendimiento de TF-IDF con métricas de ganancia de información a la hora de clasificar textos provenientes de un corpus jurídico, observando cómo estos se comportan con unos resultados muy similares pero reduciendo el tiempo de computación con respecto al requerido por TF-IDF.
- Kullback-Leibler incluso supera a TF-IDF en resultados obtenidos en nuestros experimentos.

### 5.1.3. Motivación

Este artículo surge a raíz de la participación en el **XV Congreso Internacional de Tecnología, Conocimiento y Sociedad**, que tuvo lugar en Barcelona los días 11 y 12 de marzo del año 2019. Nuestro trabajo no ha sido publicado en una revista indexada en JCR o Scopus, sino que es fruto de una publicación para un Congreso de Ciencias Sociales y tecnología y fue una primera toma de contacto con la tesis Doctoral y una manera de aportar soluciones prácticas a una problema muy concreto. Los documentos jurídicos, en España, precisan de un mayor esfuerzo para que puedan ser accedidos y clasificados adecuadamente. En los últimos años, algunas administraciones públicas han hecho un esfuerzo considerable en hacer públicos en Internet diferentes corpus jurídicos. Sin embargo, cada administración lo ha realizado de una manera diferente utilizando, por ejemplo, diferentes categorías para documentos similares lo que hace complicada la tarea de identificar documentos similares si estos han sido publicados por diferentes administraciones.

### 5.1.4. Resumen

En nuestro trabajo, proponemos realizar clasificación de textos jurídicos escritos en lengua castellana y publicados en España. Para ello, hemos hecho uso de la iniciativa *Open Data* que ha emprendido el *Gobierno de Aragón* mediante la creación de un portal llamado *Aragón Open Data* [1] y que engloba

[1] https://opendata.aragon.es/ (Accedida el 2 de mayo del 2023).

catálogos de datos abiertos publicados por las instituciones públicas de esta comunidad autónoma.

Utilizando este portal, descargamos 1700 disposiciones jurídicas para construir un dataset con disposiciones legales de diferentes categorías (por ejemplo, Declaraciones de Impacto ambiental, Ayudas y Subvenciones, Becas...). y realizamos una clasificación mediante *Support Vector Machine (SVM)*. El motivo de utilizar SVM no fue aleatorio sino que fue el clasificador que dio mejores resultados en pruebas iniciales. Sin embargo, el objetivo de este trabajo no era encontrar el mejor clasificador sino el de analizar diferentes métricas de ganancia de información. Para ello, comparamos TF-IDF (Imran y Sharan, 2011), que se utilizó como métrica de referencia, con Kullback-Leibler (simétrica y asimétrica) (Garcia et al., 2001), Ganancia de Información Selectiva (Roobaert et al., 2006) y Ganancia Compuesta (simétrica y asimétrica) (Imran y Sharan, 2011). Los resultados fueron positivos, especialmente para Kullback-Leibler, que superó los resultados obtenidos por TF-IDF. Además, se observó que las métricas de ganancia de información asimétricas rendían ligeramente por encima de sus homónimas simétricas. Si bien estos datos están plasmados en nuestro artículo, no se ha realizado una mayor investigación en este sentido.

La ventaja de establecer métricas alternativas a TF-IDF radica en el coste de computación, obtener una *matriz término-documento* calculada mediante TF-IDF implica un coste de computación más elevado que utilizando otras métricas de calculo más sencillo, como las que hemos analizado en nuestro trabajo, llegando al a conclusión de que tienen un rendimiento similar o incluso superior al de TF-IDF en algunos casos.

# Clasificación de texto. Utilizando métricas de ganancia de información para categorizar disposiciones legales.

**(Text Classification. Using Gain Information Metrics to Categorize Legal Provisions)**

Jorge Chamorro-Padial,[1] Universidad de Granada, España

Rosa Rodríguez-Sánchez, Universidad de Granada, España

*Resumen: Dentro del ámbito de la clasificación de textos, en este trabajo hemos estudiado herramientas para clasificar textos en castellano pertenecientes al dominio jurídico. Concretamente, hemos estudiado diferentes métricas basadas en la ganancia de información y su rendimiento a la hora de clasificar las disposiciones legales que conforman un Boletín Oficial. Todas las métricas estudiadas han presentado unos buenos resultados de clasificación, incluso cuando la muestra de entrenamiento era de tamaño reducido. Los resultados presentados muestran las métricas basadas en la ganancia de información como una alternativa a tener en cuenta a la hora de abordar problemas de Aprendizaje Automático con este tipo de textos.*

*Palabras clave: aprendizaje automático, clasificación de texto, minado de texto, entropía, ganancia de información, textos jurídicos*

*Abstract: In the field of Text Classification, in this work we have studied tools to classify texts written in Spanish language from the legal field. Specifically, we have analyzed the performance of different gain information metrics when dealing with classify legal provisions from an Official Journal. All the studied metrics has presented good result, even when using a small sized training set. Our results denote information gain metrics as a relevant alternative to consider when coping with legal texts on Machine Learning problems.*

*Keywords: Machine Learning, Text Classification, Text Mining, Entropy, Gain Information, Legal Texts*

## Introducción

La clasificación de documentos cuenta con un largo recorrido en las Ciencias de la Computación. Podemos distinguir dos enfoques diferentes para afrontar este problema: 1) El análisis mediante referencias bibliográficas, el cual estudia la relación entre textos analizando las citas, recibidas o aportadas, por un documento. 2) El análisis de texto, comprendido por un conjunto de técnicas utilizadas para comparar mediante el contenido del texto con el fin de inferir la distancia existente entre dos o más documentos (Ahlgren 2009).

[1] Corresponding Author: Jorge Chamorro Padial, Calle Periodista Rafael Gómez Montero 2, CITIC-UGR, Universidad de Granada, Granada, 18100, Spain. email: jorgechp@correo.ugr.es

El análisis de referencias bibliográficas se basa en la idea de que las referencias bibliográficas son un reflejo del contenido de un documento (Artandi 1965). Entre las diferentes estrategias utilizadas en este tipo de análisis, se pueden subrayar las siguientes: citación directa, coocurrencia bibliográfica y cocitación. Una de sus ventajas más importantes que presenta este enfoque es que, al no tratarse de un análisis realizado directamente sobre el texto de un documento, sino sobre sus referencias, se consigue independencia del lenguaje. Es decir, dos textos en diferentes idiomas pero que compartan referencias podrían ser clasificados como similares entre sí (Weinberg 1974). El uso análisis bibliográfico ha ido decayendo frente al análisis del texto, cuyo uso es más común en la actualidad. Sin embargo, hoy por hoy siguen siendo reseñables diferentes técnicas para clasificar documentos basadas en las referencias de los mismos, como puede ser *SimRank* o el sistema *PageRank* elaborado por Google (Nguyen et al. 2015).

En cuanto al análisis del texto presente en un documento, la minería de datos proporciona técnicas para establecer la similitud existente entre documentos. Dichas técnicas se basan en la extracción de características relevantes en documentos de forma que se puedan establecer relaciones entre las mismas. Normalmente, cada palabra es una característica. Pero de igual forma, podemos utilizar caracteres, conceptos, oraciones o párrafos como características.

En este trabajo, proponemos una técnica de análisis de texto, analizando su desempeño a la hora de clasificar documentos de carácter jurídico.

## Análisis de la entropía

El análisis de la entropía es aplicado por diferentes autores para desarrollar técnicas de clasificación de texto. Mientras que las métricas basadas en la distancia Euclídea son utilizadas para la clasificación de textos por diferentes autores, no siempre pueden ser idóneas para problemas de clasificación de datos que no pertenezcan al espacio Euclídeo (Lin, Jiang y Lee 2014). La Divergencia de Kullback-Leibler ha sido utilizada como métrica para definir distancias existentes entre documentos (Bigi 2003). Igualmente, esta métrica es empleada para para medir la distancia de existente entre textos cortos de dominio restringido (Pinto 2007). En nuestro trabajo, nos vamos a centrar en la clasificación de textos de carácter jurídico. Este tipo de textos se caracterizan por presentar un dominio semántico restringido y una estructura generalmente estandarizada, uniforme y bien definida.

## Procesado de documentos

El *Modelo de Espacio Vectorial* (SVM) se utiliza frecuentemente como forma de representación de documentos por su versatilidad, ya que permite, de una forma sencilla, tratar los documentos como vectores facilitando el cálculo de distancias entre los mismos (Bondarchuk y Timofeeva 2015). Un documento, $d_j$, se representa como un vector de términos o características, $t_i$, ponderadas con un peso, $w_{ij}$.

$$D_j = \left( w_{1\,j}, \mathrm{w}_{2\mathrm{j}\ldots}, \mathrm{w}_{\mathrm{mj}} \right)$$

Esta representación facilita el cálculo de la distancia entre documentos al tratarlos como vectores. En cuanto a la ponderación de los términos de un documento, podemos recurrir al análisis de frecuencias, aplicando TF (*Term-frequency*), DF (*Document-frequency*), IDF (*Inverse Document-Frequency*) o TF-IDF así como métricas derivadas de estas (Sabbah et al. 2017).

También es frecuente aplicar una etapa inicial de preprocesado a los documentos con el fin de eliminar caracteres especiales o palabras que no aporten información relevante, previniendo así ruido o incertidumbre que pudieran perjudicar el análisis del documento y reduciendo la complejidad del problema de clasificación, al disminuir el número de características que tratar. El uso de palabras prohibidas, y la reducción de una palabra a su raíz son algunas de las técnicas de preprocesado más habituales (Balakrishnan y Ethel 2014). La importancia del preprocesado es clave para mejorar significativamente los resultados de clasificación (Srividhya y Anitha 2010).

Una vez que el documento ha sido preprocesado y convenientemente representado, se procede a realizar el análisis pertinente sobre él. En el caso de la clasificación de textos, se aplican técnicas de Aprendizaje Automático tales como el empleo de diferentes clasificadores, algunos de los cuales requieren la definición de métricas de distancia entre documentos. En este sentido, se han empleado multitud de clasificadores diferentes, siendo habituales k-NN, Naive Bayes o SVM amén de técnicas de agrupamiento (*clustering*) y de aprendizaje no supervisado (Pratama y Sarno 2015; Aggarwal 2018, 73–112; Shafiabady et al. 2016). Más allá de la clasificación de documentos, se pueden resolver otro tipo de problemas sobre documentos, como la extracción de la categoría gramatical de una oración (*part of speech*), el análisis morfosintáctico o la segmentación de texto.

Centrándonos en la clasificación de textos, en espacios euclídeos podemos definir la distancia entre documentos mediante diferentes métricas, siendo las más representativas la distancia Euclídea, la distancia del coseno o la distancia de Jaccard. Kullback-Leibler también es utilizado con frecuencia para esta tarea (Aggarwal 2018; Jain et al. 2017). Finalmente, es relevante mencionar el Análisis Semántico Latente (Latent Semantic Analysis, LSA), como técnica que permite conocer las relaciones entre documentos y eliminar ruido (Hofmann 2017).

## Métricas basadas en la Ganancia de Información

### Ganancia de Información

La *ganancia de información* hace referencia a la reducción de la entropía producida tras la observación de una determinada característica (Coppin 2004, 278). Existen diferentes métricas basadas en la ganancia de Información. Entre ellas, vamos a analizar en este trabajo Kullback-Leibler o *Ganancia de Información Selectiva* y *Ganancia Compuesta* (García et al. 2001).

Para obtener la ganancia de información entre dos documentos, establecemos una serie de características que pueden ser medibles en base a la ganancia de información obtenida entre un documento de referencia, R, y un documento de entrada, I. Para lograr esto, cada documento se representa como un conjunto de términos y cada término aparece en el documento con una probabilidad determinada. A partir de este punto, caracterizamos la ganancia de información existente entre las distribuciones de probabilidad de R y de I como se explica a continuación.

Definimos p(t/R) y p(t/I) como la probabilidad de que un término t aparezca en el documento de referencia y en el documento de entrada, respectivamente. Asumimos que cada término en el documento R puede aparecer en el documento I. Es decir, $p(t/R) \neq 0$ y $p(t(I) \geq 0$ (García et al. 2001).

Sean las distribuciones de probabilidad del documento de referencia y del documento de entrada, respectivamente, $P=\{p(t_i/R)\}$ y $Q=\{p(t_i/I)\}$, para cada uno de los diferentes términos presentes en R y en I, podemos definir la ganancia de información como:

$$\varepsilon(P,Q)=\sum_{t_i \in R} p(t_i/R)\log\left(\frac{p(t_i/R)}{p(t_i/I)}\right)$$

Esta definición de ganancia de información equivale a Kullback-Leibler (Garcia et al. 2001). La probabilidad de que un término aparezca en un documento hace referencia a cómo de inesperado es ese término en el documento. Podemos expresar cómo de inesperados son los términos del documento R cuando se utiliza la probabilidad de distribución del documento I, por tanto, de la siguiente manera:

$$\sum_{t_i \in R} p(t_i/R)\log\left(p(t_i/I)\right)$$

El documento de referencia R es menos predecible desde una distribución estimada que mediante su distribución *real*. Finalmente, ε(P,Q) es no negativa y no simétrica. Para que cumpla todas las propiedades de un espacio métrico, la Ganancia de Información debería ser una medida de distancia, y por lo tanto, ser simétrica (Bukatin et al. 2009). Por ello, definimos la ganancia de información simétrica de la siguiente forma (García et al. 2001).

$$D=(P,Q)=\varepsilon(P,Q)+\varepsilon(Q,P)$$

**Ganancia de Información Selectiva**

La *Ganancia de Información Selectiva* puede ayudar a reducir el error del clasificador al analizar información sobre el vecindario de los términos que constituyen un documento. En definitiva, se intenta lograr dos condiciones:

1) Las características que forman el documento deben ser relevantes, es decir, el documento debe estar constituido por términos que no añadan ruido.

2) Para comparar dos documentos, se debe definir una medida selectiva que utilice el contexto del término de interés que estamos analizando.

La información relevante de cada documento la podemos determinar aplicando una fase de preprocesado como hemos explicado anteriormente. Una vez eliminadas las palabras que generan ruido en el documento, filtramos aquellas que no tienen una frecuencia alta de aparición en el documento. Para ello, definimos un valor límite, T, de la siguiente manera:

$$T(\text{cuartil}) = \inf\left\{p(t_i) \in R : \text{cuartil} \leq F(t_i)\right\}$$

Siendo $F(t_i)$ la distribución acumulada del término en R, y $p(t_i)$ la probabilidad del término $t_i$. Una vez que se han seleccionado los términos significativos de un documento, debemos definir, dado un término, qué otros términos son similares a lo largo del conjunto de documentos que vamos a analizar (corpus). En este trabajo, empleamos para este propósito *Word2vec*, una red neuronal que procesa texto con el fin de generar *word embeddings* (Mikolov et al, 2013). *Word2vec* permite aprender relaciones entre palabras analizando su vecindario.

Definimos $P(W_i)$ como el conjunto de probabilidades de los términos que forman el vecindario del término $W_i$, $\left\{p(t_i/R) \vee t_i \in W_i\right\}$. Estos vecinos se caracterizan por tener una gran similitud (determinada por Word2vec) con respecto a $W_i$.

La *Ganancia de Información Selectiva* de un término $W_i$ se expresa de la siguiente manera:

$$\varepsilon^{W_i} = (\mathrm{P},\mathrm{Q}) = \varepsilon\left(P(W_i), \mathrm{Q}(W_i)\right)$$

Por lo tanto, la *Ganancia de Información Selectiva* entre dos documentos queda definida así:

$$\varepsilon^{W_1, W_2, \ldots, W_n}(\mathrm{P},\mathrm{Q}) = \sum_{i=1}^{n} \varepsilon^{W_i}(\mathrm{P},\mathrm{Q})$$

Siendo $\left\{W1, W2, \ldots, \mathrm{W}_n\right\}$ términos de interés en el documento de referencia. Nuevamente, esta métrica no cumple la propiedad de simetría.

**Ganancia de información compuesta**

En un documento podemos encontrarnos palabras con diferentes niveles de significancia que viene definida por la frecuencia de un término en un documento en relación con la frecuencia, a su vez, de este término en el corpus. Si un término tiene una frecuencia elevada en un documento, y una frecuencia reducida en el corpus, este término es relevante a la hora de identificar a un determinado documento. Se define la ganancia de información compuesta en un término de la siguiente manera:

$$\varepsilon_C^{W_i}(\mathrm{P},\mathrm{Q}) = \sum_{t \in R \wedge t \in N(W_i)} \text{idf}(W_i)\, p(t/\mathrm{Rw}_i) \log\left(\frac{\text{idf}(W_i)\, p(t/R)}{\text{idf}(W_i)\, p(t/I)}\right)$$

Siendo N ($W_i$) el conjunto de términos similares a $W_i$ de acuerdo con su vecindario e idf($W_i$) la frecuencia inversa de documento (Ao, Castillo y Huang 2011, 193).

La ganancia de información compuesta entre los documentos R e I se define de la siguiente manera:

$$\varepsilon_C^{W_1,W_2,\ldots,W_n}(P,Q) = \sum_{i=1}^{n} \varepsilon_C^{W_i}(P,Q)$$

En este caso, la ecuación anterior no cumple la propiedad de no negatividad. La ganancia asimétrica es una variación de esta ecuación, que sí cumple la no negatividad:

$$D_C^{W_1,W_2,\ldots,W_n}(P,Q) = \varepsilon_C^{W_1,W_2,\ldots,W_n}(P,Q) + \varepsilon_C^{W_1,W_2,\ldots,W_n}(Q,P)$$

## Diseño experimental

El objetivo del presente trabajo es analizar la bondad de las medidas de ganancia de información anteriormente presentadas a la hora de analizar disposiciones legales con el fin de clasificarlas en diferentes categorías. Para llevar a cabo esta tarea, se ha requerido el uso de un *dataset* formado por disposiciones legales del *Boletín Oficial de Aragón*. Asimismo, se ha utilizado Support Vector Machine (SVM) como herramienta de clasificación.

### Clasificación

#### *Corpus*

La iniciativa *Open Data* busca facilitar la disponibilidad de datos e información, especialmente aquellos que proceden de la Administración Pública. Dicha información debe publicarse de manera abierta y reutilizable, para poder ser aprovechada y generar servicios en torno a ella. De forma paralela, se satisface el objetivo de mejorar la transparencia de las Administraciones Públicas. En este sentido, en España, el *Gobierno de Aragón* creó el portal *Aragón Open Data*[2], con el fin de crear un catálogo de datos abiertos publicados no solamente por el *Gobierno de Aragón*, sino también por el resto de instituciones públicas aragonesas. Los datos publicados en este portal se encuentran en formatos fácilmente reutilizables y legibles por aplicaciones informáticas (XML y JSON, principalmente).

Los boletines oficiales constituyen una fuente normativa primaria publicada por diferentes administraciones con el fin de difundir disposiciones legales aprobadas por las mismas (Merlo Vega 2010, 100).

Entre los datos publicados por el *Gobierno de Aragón* en *Aragón Open Data*, se encuentra el *Boletín Oficial de Aragón* (BOA). El BOA se pública diariamente y está constituido como un servicio público de acceso universal[3].

[2] Gobierno de Aragón. *Boletín Oficial de Aragón*. 18 de mayo de 2019. http://www.boa.aragon.es/

[3] Gobierno de Aragón. *Aragón Open Data*. 5 de febrero de 2019. https://opendata.aragon.es/informacion/open-data

Utilizando la plataforma de datos abiertos del *Gobierno de Aragón*, hemos construido un *dataset* con 1700 disposiciones legales extraídas del BOA. En la *Tabla 1* se muestran las diferentes categorías y el número de disposiciones legales descargadas en cada una de ellas.

Las disposiciones que componen el *dataset* se encuentran clasificadas dentro de las catorce categorías que se reflejan en la *Tabla 1*. Dichas categorías han sido proporcionadas por la plataforma *Aragón Open Data* y constituyen el objeto de estudio de nuestro trabajo, puesto que nos permiten analizar la bondad de las métricas de ganancia de información a la hora de clasificar diferentes disposiciones legales en estas categorías.

Tabla 1: Estructura del dataset del *Boletín Oficial de Aragón*

| *Categoría* | *Disposiciones legales* | *%* |
|---|---|---|
| Anuncios de licitaciones públicas | 200 | 11,76 |
| Ayudas y subvenciones | 200 | 11,76 |
| Becas | 77 | 4,53 |
| Cartas de servicios | 84 | 4,94 |
| Concursos de personal público | 200 | 11,76 |
| Convenios | 100 | 5,88 |
| Cursos | 100 | 5,88 |
| Declaraciones de impacto ambiental | 200 | 11,76 |
| Nombramientos y ceses de Altos Cargos | 100 | 5,88 |
| Normas básicas | 100 | 5,88 |
| Oposiciones | 100 | 5,88 |
| Premios | 100 | 5,88 |
| Registro de fundaciones | 39 | 2,29 |
| Sentencias | 100 | 5,88 |
| Total | 1700 | |

*Fuente: Elaboración propia.*

***Support Vector Machine (SVM)***

SVM es un método de aprendizaje supervisado utilizado tanto en regresión como en clasificación. Fue propuesto por Corinna Cortes y Vladimir Vapnik para resolver problemas de Aprendizaje Automático (Cortes y Vapnik 1995). SVM trata de establecer una separación entre datos seleccionando instancias relevantes de los mismos. Entre sus ventajas destaca su versatilidad a la hora de tratar con vectores que posean una alta dimensionalidad, circunstancia que se da habitualmente al trabajar con texto (Becker et al. 2011).

***Método***

Una vez construido el *dataset* del BOA, se ha estudiado el comportamiento de la clasificación del *dataset* para diferentes tamaños de la muestra de entrenamiento (25%, 50% y 90%). Para asignar instancias entre la muestra de entrenamiento y la de test, se realiza una selección aleatoria. Una vez concretada la división del *dataset*, se aplica una etapa de preprocesado a todas las muestras. Dicha etapa tiene las siguientes fases:

1. Se convierte todo el texto a minúsculas.
2. Se filtran las palabras poco relevantes aplicando una lista de palabras prohibidas proporcionada por la suite de bibliotecas *Natural Language Toolkit* (NLTK) para el lenguaje de programación *Python*[4].
3. Se elimina cualquier elemento que no constituya propiamente una palabra (comas, signos de puntuación, indicadores de monedas, paréntesis, signos de exclamación y de interrogación, etc) y todos los caracteres numéricos.
4. Sobre las palabras resultantes, se aplica el algoritmo *Snowball Stemmer* (Porter 2001), igualmente proporcionado por NLTK.

Tras esta primera etapa, se calcula a continuación la *matriz de frecuencias término-documento*. Sobre la cual aplicaremos las diferentes métricas de ganancia de información. Del mismo modo, se calcula la *matriz término-documento TF-IDF*. Los resultados de clasificación de las métricas de ganancia se compararán con los obtenidos utilizando TF-IDF, que es una métrica muy utilizada por diferentes autores, con buenos resultados a la hora de clasificar texto (Ahlgren 2009). Finalmente, se aplica el clasificador SVM sobre ambas matrices.

## Análisis de resultados

Se han realizado cinco pruebas diferentes por cada tamaño de la muestra, obteniendo la media y la desviación típica de la precisión, exhaustividad, exactitud y el valor–F. Se pretende analizar la bondad de las métricas de ganancia de información que han sido propuestas, así como comparar su rendimiento con el obtenido por TF-IDF.

En primer lugar, podemos comprobar los resultados obtenidos utilizando el 25% de las instancias como muestras de entrenamiento:

Tabla 2: Resultados obtenidos (muestra de entrenamiento = 25%)

[4] NLTK Project. 2019. "Natural Language Toolkit," acceso el 18 de mayo de 2019, https://www.nltk.org/.

| *Métrica* | *Corpus* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Exactitud* | | *Precisión* | | *Exhaustividad* | | *Valor-F* | |
| | A | D | A | D | A | D | A | D |
| *TF-IDF (Referencia)* | 0,837 | 0,001 | 0,988 | 0,007 | 0,896 | 0,025 | 0,939 | 0,014 |
| *Kullback – Leibler* | **0,874** | 0,012 | 0,967 | **0,001** | **0,925** | **0,020** | **0,945** | **0,009** |
| *Kullback – Leibler (Simétrica)* | **0,875** | 0,009 | 0,970 | 0,024 | **0,917** | **0,023** | **0,942** | 0,014 |
| *Ganacia de Información Selectiva* | 0,797 | 0,021 | 0,906 | 0,019 | 0,888 | **0,020** | 0,894 | 0,016 |
| *Ganancia compuesta* | 0,795 | 0,018 | 0,914 | 0,112 | 0,873 | **0,007** | 0,893 | **0,009** |
| *Ganancia Compuesta asimétrica* | 0,799 | 0,017 | 0,915 | **0,001** | 0,886 | **0,007** | 0,900 | **0,007** |

*Leyenda: A = Media de 5 ejecuciones. D = Desviación típica de 5 ejecuciones. En negrita = Valores que han superado a la métrica de referencia (TF-IDF).*

*Fuente: Elaboración propia.*

Kullback-Leibler, tanto en su variante asimétrica como en su variante simétrica, mejoran el valor-F obtenido por TF-IDF, superando a esta métrica en exhaustividad y exactitud. Todas las métricas obtienen un valor-F en torno al 90%.

A continuación, se presentan los resultados conseguidos utilizando el 50% de las instancias del *dataset* como muestra de entrenamiento:

Tabla 3: Resultados obtenidos (muestra de entrenamiento = 50%)

| *Métrica* | *Corpus* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Exactitud* | | *Precisión* | | *Exhaustividad* | | *Valor-F* | |
| | A | D | A | D | A | D | A | D |
| *TF-IDF (Referencia)* | 0,902 | 0,009 | 0,991 | 0,007 | 0,912 | 0,022 | 0,958 | 0,026 |
| *Kullback – Leibler* | **0,914** | **0,006** | 0,977 | 0,017 | **0,940** | **0,014** | **0,964** | **0,017** |
| *Kullback – Leibler (Simétrica)* | **0,913** | **0,006** | 0,978 | **0,008** | **0,936** | **0,011** | 0,956 | **0,009** |
| *Ganacia de Información Selectiva* | 0,856 | **0,001** | 0,917 | 0,026 | 0,906 | **0,017** | 0,917 | **0,018** |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Ganancia compuesta* | 0,848 | 0,010 | 0,921 | 0,026 | 0,896 | **0,014** | 0,909 | **0,010** |
| *Ganancia compuesta asimétrica* | 0,848 | 0,009 | 0,922 | **0,003** | 0,901 | **0,018** | 0,911 | **0,012** |

*Leyenda: A = Media de 5 ejecuciones. D = Desviación típica de 5 ejecuciones. En negrita = Valores que han superado a la métrica de referencia (TF–IDF).*

*Fuente: Elaboración propia.*

En términos generales, se puede ver cómo se consiguen resultados muy a la par de los logrados cuando la muestra de entrenamiento es del 25%. Nuevamente, Kullback-Leibler vuelve a obtener mejores resultados que TF-IDF en términos generales. En general, los resultados obtenidos son muy similares entre TF-IDF y las dos métricas basadas en Kullback-Leibler. El resto de métricas basadas en la ganancia de información obtienen resultados por encima de 0.9 con respecto al valor de valor-F.

Finalmente, presentamos los resultados obtenidos utilizando una muestra de entrenamiento del 90%:

Tabla 4: Resultados obtenidos (muestra de entrenamiento = 90%)

| *Métrica* | *Corpus* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Exactitud* | | *Precisión* | | *Exhaustividad* | | *Valor-F* | |
| | A | D | A | D | A | D | A | D |
| *TF-IDF (Referencia)* | 0,927 | 0,021 | 0,994 | 0,013 | 0,947 | 0,052 | 0,969 | 0,034 |
| *Kullback – Leibler* | **0,928** | **0,019** | 0,977 | 0,022 | 0,932 | **0,040** | 0,953 | **0,029** |
| *Kullback – Leibler (Simétrica)* | 0,926 | **0,014** | 0,978 | 0,023 | 0,943 | 0,052 | 0,959 | **0,027** |
| *Ganancia de Información Selectiva* | 0,864 | 0,025 | 0,924 | 0,037 | 0,939 | 0,062 | 0,928 | **0,025** |
| *Ganancia compuesta* | 0,869 | **0,013** | 0,920 | 0,036 | 0,922 | **0,040** | 0,918 | **0,027** |
| *Ganancia compuesta asimétrica* | 0,871 | **0,019** | 0,923 | 0,041 | 0,937 | 0,050 | 0,930 | 0,035 |

*Leyenda: A = Media de 5 ejecuciones. D = Desviación típica de 5 ejecuciones. En negrita = Valores que han superado a la métrica de referencia (TF-IDF).*

*Fuente: Elaboración propia.*

Aquí, ninguna métrica consigue superar a TF-IDF a excepción de Kullback-Leibler en cuanto a exhaustividad. Todas las métricas, no obstante, obtienen valores F superiores a 0.9.

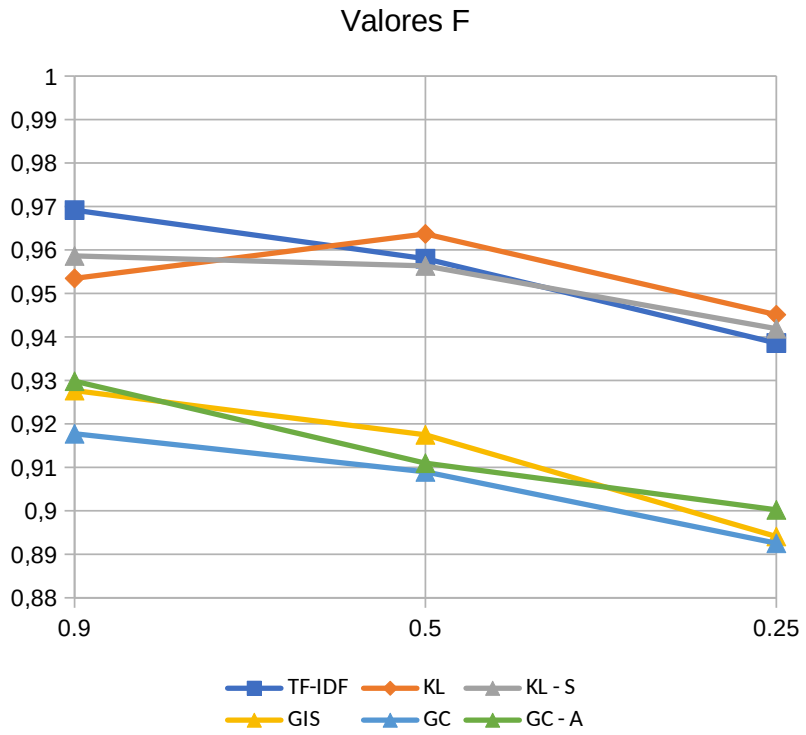



*Figura1: Valores F en función del tamaño de la muestra de entrenamiento.*

*KL = Kullback Leibler. KL-S = Kullback-Leibler (Simétrica). GIS = Ganancia de Información Selectiva. GC = Ganancia Compuesta. GC – A = Ganancia Compuesta (Asimétrica).*

*Fuente: Elaboración propia.*

Los resultados obtenidos muestran que todas las métricas de ganancia de información pueden ser útiles para clasificar el *dataset* que hemos explorado. Es relevante destacar como, utilizando una muestra de entrenamiento de tamaño reducido, ya se tiene información suficiente

para realizar una clasificación de calidad. Estos resultados se encuentran favorecidos, en parte, por el dominio concreto del lenguaje que se utiliza en este tipo de documentos.

Kullback-Leibler, en esta situación, logra superar a TF-IDF como métrica para realizar clasificar el *dataset*. Es remarcable el hecho de que Kullback–Leibler asimétrico rinde ligeramente por encima de Kullback-Leibler simétrico. Esta tendencia se repite en *la Ganancia Compuesta de Información*, que presenta mejores resultados en su versión no simétrica.

Como era de esperar, todas las métricas mejoran cuando se incrementa el tamaño de la muestra utilizada como de entrenamiento, siendo el porcentaje de mejora muy similar en todos los casos.

En cuanto a la desviación típica, todas las métricas presentan una dispersión muy baja y con valores similares. Aquí es reseñable el comportamiento de todas las métricas de ganancia de información, así como de Kullback-Leibler, puesto que presentan una menor dispersión que TF-IDF en casi todos los escenarios probados.

## Conclusiones

Estamos ante un problema de clasificación de textos de carácter jurídico, que se suelen caracterizar por moverse en un dominio restringido del castellano: tanto en estructura como en vocabulario. Estos factores contribuyen a facilitar enormemente la tarea de su clasificación mediante algoritmos de Aprendizaje Automático. Por el lado contrario, son textos, además, de extensión relativamente corta, lo que normalmente puede suponer un hándicap a la hora de extraer términos que sean suficientemente relevantes como para identificar un documento diferenciándolo de todos los demás.

Las métricas aquí estudiadas, basadas en la ganancia de información, se han comportado de manera muy eficiente constituyendo una alternativa a la clasificación utilizando TF-IDF. La *matriz término-documento* calculada mediante TF-IDF requiere unos requisitos de computación más elevados que un simple cálculo de frecuencias, que puede ser suficiente para construir una *matriz término-documento* que pueda ser analizada por medidas de ganancia de información. Especialmente, Kullback-Leibler ha mostrado un porcentaje de clasificación muy sólido especialmente cuando se han trabajado con pocas muestras.

El problema planteado en este trabajo también tiene un componente adicional que lo hace interesante, y es que parte de una clasificación temática de disposiciones legales en diferentes temáticas. Dicha clasificación no es habitual encontrarla en los portales de acceso a Boletines Oficiales de las administraciones públicas. Lo que podría permitir que, utilizando el *dataset* de Boletín Oficial de Aragón como base de conocimiento, se podría llegar a clasificar disposiciones legales de otros boletines en torno a una serie de temáticas previamente definidas, facilitando su búsqueda y clasificación. Cabe destacar que no todos los Boletines Oficiales permiten buscar disposiciones legales por temática, y los que lo permiten, no presentan las mismas temáticas. Entrenar un *dataset* de disposiciones legales como el que hemos construido para este trabajo permitiría clasificar otros boletines oficiales utilizando el mismo patrón temático.

El análisis de texto nos abre una puerta de gran importancia hacia la evolución de la Administración en su proceso de transformación digital. En primer lugar, facilita la creación y mantenimiento de archivos y repositorios de documentos cuyo mantenimiento podría simplificarse enormemente. De igual modo, la posibilidad de clasificación masiva de documentos favorece la facilidad de acceso a la información y, por ende, implica una mejora de la transparencia.

Se hace necesario, por ello, continuar trabajando con iniciativas *Open Data*, accediendo a repositorios digitales Institucionales y analizando la necesidad y la posibilidad de aplicar técnicas de análisis de texto, para así conocer qué información puede ser clasificada y gestionada eficientemente mediante las diferentes técnicas que ya conocemos en la actualidad.

## Agradecimientos

Los autores quieren agradecer al *Gobierno de Aragón* por su iniciativa *Aragón Open Data*, que nos ha permitido extraer la información necesaria para este trabajo.

Esta investigación ha sido subvencionada por el Ministerio de Economía, Industria y Competitividad (MICINN), del Gobierno de España, con la referencia TIN2017–85542–P, y cofinanciado con los fondos europeos FEDER.

## REFERENCIAS

Aggarwal, Charu C. 2018. *Machine Learning for Text*. Cham: Springer International Publishing.

Ahlgren, Per y Cristian Colliander. 2009. “Document–Document Similarity Approaches and Science Mapping: Experimental Comparison of Five Approaches.” *Journal of Informetrics* 3 (1): 49–63. doi:10.1016/j.joi.2008.11.003.

Ao, Sio-Iong, Oscar Castillo y Xu Huang. 2011. *Intelligent Control and Computer Engineering*. Dordrecht: Springer.

Artandi, Susan. 1965. “Automatic Indexing: A State-Of-The-Art Report.” *American Documentation* 16 (4): 334. doi:10.1002/asi.5090160409.

Balakrishnan, Vimala y Lloyd-Yemoh Ethel. 2014. “Stemming and Lemmatization: A Comparison of Retrieval Performances.” *Lecture Notes on Software Engineering* 2 (3): 262–67. doi:10.7763/lnse.2014.v2.134.

Becker, Natalia, Grischa Toedt, Peter Lichter y Axel Benner. 2011. “Elastic SCAD as a Novel Penalization Method for SVM Classification Tasks in High-Dimensional Data.” *BMC Bioinformatics* 12 (1). doi:10.1186/1471-2105-12-138.

Bigi, Brigitte. 2003. “Using Kullback-Leibler Distance for Text Categorization.” *Proceedings of the 25Th European Conference on IR Research*, 305–319. doi:10.1007/3-540-36618-0_22.

Bondarchuk, Dmitry y Galina Timofeeva. 2015. “Vector Space Model Based on Semantic Relatedness.” *AIP Conference Proceedings* 1690: 20005. doi:10.1063/1.4936683.

Bukatin, Michael., Ralph Kopperman, Steve Matthews, y Homeira Pajoohesh. 2009. “Partial Metric Spaces.” *American Mathematical Monthly* 116 (8): 708–18. doi:10.4169/193009709x460831.

Coppin, Ben. 2004. *Artificial intelligence illuminated*. Boston: Jones and Bartlett Publishers.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” *Machine Learning* 20 (3): 273–97. doi:10.1007/bf00994018.

García, José A., Joaquín Fdez-Valdivia, Xose R. Fdez-Vidal, y Rosa Rodriguez-Sánchez. 2001. “Information Theoretic Measure for Visual Target Distinctness.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (4): 362–83. doi:10.1109/34.917572.

Nguyen, Phuong, Paolo Tomeo, Tommaso di Noia y Eugenio di Sciascio. 2015. “An Evaluation of Simrank and Personalized Pagerank to Build a Recommender System for The Web of Data.” *Proceedings of The 24Th International Conference On World Wide Web - WWW '15 Companion*. doi:10.1145/2740908.2742141.

Hofmann, Thomas. 2017. “Probabilistic Latent Semantic Indexing”. *ACM SIGIR Forum* 51 (2): 211–18. doi:10.1145/3130348.3130370.

Jain, Abhishek, Aman Jain, Nihal Chauhan, Vikrant Singh y Narina Thakur. 2017. “Information Retrieval Using Cosine and Jaccard Similarity Measures in Vector Space Model.” *International Journal of Computer Applications* 164 (6): 28–30. doi:10.5120/ijca2017913699.

Lin, Yung-Shen., Jung-Yi Jiang y Shie-Jue Lee. 2014. “A Similarity Measure for Text Classification and Clustering.” *IEEE Transactions On Knowledge And Data Engineering* 26 (7): 1575–90. doi:10.1109/tkde.2013.19.

Liu, Ling, y Tamer Özsu. 2009. *Encyclopedia of Database Systems*. New York: Springer.

Merlo Vega, José Antonio. 2010. *Información y Referencia en Entornos Digitales*. Murcia: Ediciones de la Universidad de Murcia.

Mikolov, Tomas., Kai Chen, Greg Corrado y Jeffrey Dean. 2013, "Efficient Estimation of Word Representations in Vector Space". International Conference on Learning Representations. https://dblp.org/rec/bib/journals/corr/abs-1301-3781

Pinto, David, José-Miguel Benedí, and Paolo Rosso. 2007. "Clustering Narrow-Domain Short Texts by Using The Kullback-Leibler Distance." *Computational Linguistics and Intelligent Text Processing* 611–22. doi:10.1007/978-3-540-70939-8_54.

Pratama, Bayu Yudha y Riyanarto Sarno. 2015. "Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM." *2015 International Conference on Data and Software Engineering (Icodse)*. doi:10.1109/icodse.2015.7436992.

Porter, Martin F. 2001. "Snowball: A Language for Stemming Algorithms" *Tartarus*, acceso el 8 de febrero de 2019. http://snowball.tartarus.org/texts/introduction.html.

Sabbah, Thabit, Ali Selamat, Mohd Hafiz Selamat, Fawaz Al-Anzi, Enrique Herrera Viedma, Ondrej Krejcar y Hamido Fujita. 2017. "Modified Frequency-Based Term Weighting Schemes for Text Classification." *Applied Soft Computing* 58: 193–206. doi:10.1016/j.asoc.2017.04.069.

Srividhya, Vasudevan, y R. Anitha. 2010. "Evaluating Preprocessing Techniques in Text Categorization." *International Journal of Computer Science and Application Issue*, 49–51. http://sinhgad.edu/ijcsa-2012/pdfpapers/1_11.pdf.

Shafiabady, Niusha, Lam Hong Lee, Rajprasad Kumar Rajkumar, Vish P. Kallimani, Nik Ahmad Akram y Dino Isa. 2016. "Using Unsupervised Clustering Approach to Train the Support Vector Machine for Text Classification." *Neurocomputing* 211: 4–10. doi:10.1016/j.neucom.2015.10.137.

Weinberg, Bella Hass. 1974. "Bibliographic Coupling: A Review." *Information Storage and Retrieval* 10 (5–6): 189–196. doi:10.1016/0020-0271(74)90058-8.

## SOBRE LOS AUTORES

***Jorge Chamorro-Padial**:* Doctorando, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Granada, España.

***Rosa Rodríguez-Sánchez**:* Profesora Titular de Universidad, Computer Vision Group (CVG), Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Granada, España.

# Capítulo 6

# Aplicaciones prácticas de esta Tesis Doctoral

## 6.1. Más allá de los trabajos académicos

Como fruto del trabajo de investigación de esta tesis, se han desarrollado aplicaciones y herramientas que han sido publicadas en repositorios y en servidores web con dos objetivos: Validar los resultados de nuestros trabajos y ponerlas a disposición de la Comunidad Académica y de todo aquel a quien pudiera resultarles de utilidad. En este capítulo, se explora la colección de programas y herramientas que han sido desarrolladas.

La mayoría de esta producción, de carácter práctico, son aplicaciones web que están siendo alojadas en servidores del *Grupo de Visión por Computador* de la Universidad de Granada.

### 6.1.1. Quasi-Species Peer review

*Quasi-Species Peer review* [1] es una aplicación web que permite a los usuarios que se registren simular el proceso de revisión de artículos científicos a partir del análisis del título, abstract y palabras clave (de autor y KeyWords Plus).

La web está dividida en dos secciones (Ver Figura 6.1): una sección de entrenamiento, donde se presentan al usuario artículos aleatorios procedentes de (Chamorro-Padial y Rodríguez-Sánchez, 2020c) (Para más información, se puede consultar la Subsección 6.2.1). El usuario debe, con esta información, emitir un juicio sobre la revista científica más apropiada para el artículo en cuestión, eligiendo entre una revista de bajo impacto, una revista de un impacto medio, y una revista de alto impacto. En todo momento, además de las respuestas proporcionadas por el usuario, se mide el tiempo de respuesta.

[1] https://blackcat.ugr.es/quasispecies (Accedida el 8 de Abril del 2023).

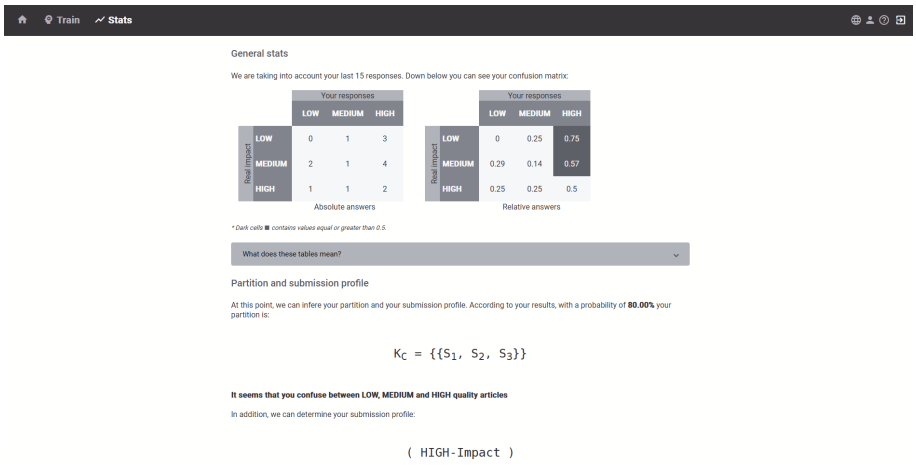
Tras completar un mínimo de quince evaluaciones, el usuario puede acceder a la sección de estadísticas, donde es posible comprobar su *partición* y su *perfil de envío*.

La aplicación web es multilingüe, admitiendo inglés y español como idiomas a mostrar al usuario, si bien tiene soporte para añadir fácilmente nuevos idiomas, sin requerir un conocimiento técnico profundo. Para evitar datos sesgados o generados automáticamente, existen pruebas para bloquear a scripts automatizados, *bots*, que puedan encontrar la web y atacarla.

Esta web nos sirve de apoyo para el artículo *The Relevance of Title, Abstract, and Keywords for scientific paper quality and potential impact.*

La aplicación está desarrollada en Angular (interfaz web) y en Python, con Django.


(a) Sección de entrenamiento.




(b) Sección de estadística.

Figura 6.1: Captura de pantalla para la secciones de entrenamiento y de estadística de la aplicación *Quasi-Species Peer Review.*

### 6.1.2. SenSpePeer (SSP)

*SenSpePeer (SSP)* [2] es una aplicación web (Ver Figura 6.2) que permite determinar el grado de *especificidad* y *sensitividad* de un proceso de revisión por pares, de acuerdo con nuestro trabajo: *What is the sensitivity and specificity of the peer review process?* La aplicación está pensada para editores, revisores y personas involucradas en el proceso de publicación.

En esta aplicación, se plantean cuatro pasos diferentes para obtener los valores deseados:

1. Se introduce el valor $q$, recordemos que este valor representa la probabilidad *a priori* de que el manuscrito sea aceptado en una revista determinada según el modelo de probabilidad bayesiana.
2. Se define el valor de utilidad $U_1$ que el editor obtiene al aceptar un manuscrito que **cumple** con los estándares de la revista.
3. Se define el valor de utilidad $U_0$ que el editor obtiene al aceptar un manuscrito que **no cumple** con los estándares de la revista.
4. Finalmente, definimos el valor $\lambda$, que establece el coste de procesar una porción de información en el informe de los revisores.

La aplicación está desarrollada enteramente en Angular.

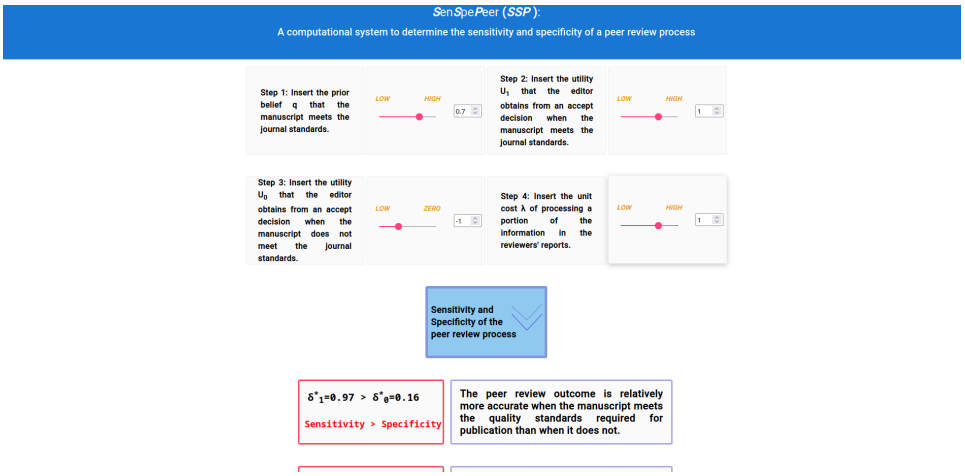



Figura 6.2: Ejemplo de uso de *SenSpePeer*.

## 6.2. Datasets

Conjuntamente con los artículos y manuscritos en los que hemos estado trabajando durante el desarrollo de esta tesis, hemos generado datos uti-

[2] https://blackcat.ugr.es/senspepeer/ (Accedida el 8 de Abril del 2023).

lizados en nuestros análisis experimentales. La mayor parte de estos datos contienen información sobre artículos (generalmente: título, abstract y descripción) o palabras clave.

Con el fin de aportar transparencia a la investigación realizada así como ayudar a otros investigadores, hemos publicado estos datos. A continuación, se hace una descripción de cada uno de los datasets publicados.

### 6.2.1. Computer Science Articles & Journals, 2019

Este dataset contiene información sobre el *título, el abstract y palabras clave* de, aproximadamente, 22.000 artículos científicos indexados en *Web of Science* [3] en la categoría *Computer Science, Artificial Intelligence*, durante el año 2019.

El dataset ha sido utilizado para elaborar el artículo *The Relevance of Title, Abstract, and Keywords for scientific paper quality and potential impact* y se encuentra publicado en el repositorio Kaggle (Chamorro-Padial y Rodríguez-Sánchez, 2020c).

### 6.2.2. akkp69000

En este dataset, publicamos una lista de palabras clave de autor (AK) y de *KeyWord Plus*©(KP) de un total de 69.000 artículos descargados desde *Web of Science*. Los datos están divididos en tres colecciones diferentes:

- **Raw**: Datos en bruto, tal cual fueron obtenidos desde WoS. Están distribuídos en diferentes ficheros en formato csv, los cuales tienen dos columnas: DE (AK) e ID (KP).
- **filtered**: No todos los artículos descargados tienen información sobre las AKs o las KPs. Por ello, esta colección contiene los datos en bruto, pero hemos eliminado artículos donde falte un tipo de palabra clave.
- **pre_processed**: Las palabras clave han sido sometidas a un proceso de limpieza y tratamiento de datos: eliminando caracteres especiales, convirtiendo todos los caracteres a minúscula y aplicando un algoritmo básico de *stemming* (Porter, 1980).

Este dataset fue creado para trabajar con un artículo donde se analizaban las diferencias entre las redes de autores basadas en AKs y aquellas basadas en KPs. Este trabajo, finalmente, no fue publicado pero preferimos permitir que la comunidad pudiera hacer uso de los datos descargados. El dataset se encuentra publicado en dos repositorios: Kaggle (Chamorro-Padial y Rodríguez-Sánchez, 2020a) y Mendeley Data (Chamorro-Padial y Rodríguez-Sánchez, 2020b).

[3] https://www.webofscience.com/wos/ (Accedida el 8 de Abril del 2023).

### 6.2.3. Author Keywords - KeywordsPlus

Este dataset contiene diferentes colecciones datos que utilizamos para la elaboración del diseño experimental del artículo *Corner Centrality of Nodes in Multilayer Networks: A Case Study in the Network Analysis of Keywords*.

El dataset tiene dos partes bien diferenciadas. En una de ellas, en un directorio interno están registradas 48.115 AKs y 24.040 KPs. El resto de directorios contienen dos redes multicapa, de dos capas cada una de ellas. La primera red multicapa contiene palabras clave de autor de artículos de Ciencias de la Computación mientras que la segunda contiene relaciones entre AKs y KPs, En ambas redes, cuando existe un enlace entre palabras clave, significa que ambas han aparecido conjuntamente en un mismo artículo científico.

Toda la información ha sido descargada de la base de datos *Web of Science*. El dataset se encuentra publicado en Kaggle (Rodríguez-Sánchez y Padial, 2022).

# Capítulo 7

# Trabajos futuros

## 7.1. Uniendo líneas de investigación: Cuasiespecies y Palabras clave

Durante todo el desarrollo de esta tesis doctoral hemos hablado de dos líneas de investigación que, si bien forman parte del análisis de los patrones de comportamiento de editores, revisores y autores científicos, han discurrido de una manera independiente y diferenciada. Como ya se ha mencionado anteriormente: una línea ha supuesto continuar el trabajo realizado por el grupo de investigación en el que me he incorporado: el Grupo de Visión por Computador de la Universidad de de Granada, mientras que la otra línea ha supuesto la apertura de un nuevo frente y de una aportación original a la ciencia en un campo poco estudiado.

Sin embargo, no son campos excluyentes entre sí y creemos que lo aprendido por lo investigado en este grupo puede ser perfectamente aplicado al estudio de las palabras clave en la Comunidad Científica. Por ello, se ha planteado la elaboración de un artículo que aplica el modelo de cuasiespecies a la elección de palabras clave realizada por los autores.

La intención de esta sección es ilustrar un trabajo futuro, en el que aún se está trabajando, pero que supondrá la unión sinérgica de ambas líneas de investigación. Si bien aún queda trabajo por delante para que este artículo vea la luz, he considerado oportuno exponer el borrador con las primeras ideas de cara a terminar de dar sentido a la tesis doctoral así como generar un posible debate de ideas en la defensa de la tesis.

En este futuro trabajo, se proponen tres modelos diferentes a estudiar:

1. Modelo *Survival*.
2. Modelo *Attention*.
3. Modelo *Survival/Attention*.

### 7.1.1. Ecuaciones comunes a los tres modelos

El beneficio de una determinada palabra clave, $k$, se mide en términos de Attention, $A$, de Survival, $S$ o de Attention-Survival, $AS$. Sea $C(k)$ el vecindario de $k$, definimos $S(k)$ como la propensión de $k$ a sobrevivir [1]:

$$S(k) = \frac{1}{|C(k))|}$$

Por su parte $A(k)$ hace referencia al nivel de interés de la Comunidad en $k$ y, $AS(k)$ es el beneficio obtenido por $k$ en términos de Attention y Survival:

$$AS(k) = \alpha \cdot S(k) + (1 - \alpha) \cdot A(k)$$

Cuando un autor elige una palabra clave, obtiene una recompensa en términos de $S$ obtenida, otra recompensa en términos de $A$, y una tercera recompensa expresada en términos de $AS$. También existe un coste de oportunidad a asumir por elegir una palabra clave determinada en vez de cualquier otra. Este coste de oportunidad se traduce en una penalización en términos de $S$, de $A$ y de $AS$.

El coste de elegir una palabra clave, $k$, se expresa de la siguiente manera:

$$C_k = 1 - S(k)$$

$$C_k = 1 - A(k)$$

$$C_k = 1 - AS(k)$$

### 7.1.2. Modelo Survival

Una palabra clave se puede clasificar en base a su posición en términos de Survival, $S$, en las siguientes categorías:

- $S_L \longrightarrow$ La palabra clave se encuentra en el tercer tercil en términos de Survival.
- $S_M \longrightarrow$ La palabra clave se encuentra en el segundo tercil en términos de Survival.
- $S_H \longrightarrow$ La palabra clave se encuentra en el primer tercil en términos de Survival.

[1] Entendemos *supervivencia* como la facultad de una palabra clave de pertenecer a un artículo escogido por un lector durante su proceso de búsqueda de información.

El conjunto de categorías de una palabra clave en el Modelo Survival se expresa de la siguiente forma: $K_S = \{S_L, S_M, S_H\}$.

Un autor puede tener diferentes perfiles según su capacidad para distinguir el Survival de las palabras clave. Podemos representar esta capacidad mediante una aplicación biyectiva, $f : D \to P$. donde $D_{Sn}$ representa el perfil de distinción de un autor y $P$ su partición.

$$\begin{aligned} D_{S1} &\longrightarrow \{\{S_L\}, \{S_M\}, \{S_H\}\} \\ D_{S2} &\longrightarrow \{\{S_L, S_M\}, \{S_H\}\} \\ D_{S3} &\longrightarrow \{\{S_L, S_H\}, \{S_M\}\} \\ D_{S4} &\longrightarrow \{\{S_M, S_H\}, \{S_L\}\} \\ D_{S5} &\longrightarrow \{\{S_L, S_m, S_H\}\} \end{aligned}$$

Para cada perfil; un autor tendrá, con una probabilidad determinada, un comportamiento que denominamos perfil de respuesta. Con cierta probabilidad, un autor con perfil de distinción $D_{S5}$ elegirá una palabra clave de tipo $S_L$, $S_M$ o bien, $S_H$. Por ejemplo, para $D_{S5}$ existen tres perfiles de respuesta diferentes:

$$D_{S5} = \{\{S_L\}, \{S_M\}, \{S_H\}\}$$

Para el caso de $D_{S2}$ tendríamos nueve perfiles de respuesta:

$$\begin{aligned} D_{S2} = \{ & \\ & \{\{S_L\}, \{S_L\}\}, \\ & \{\{S_L\}, \{S_H\}\}, \\ & \{\{S_M\}, \{S_L\}\}, \\ & \{\{S_M\}, \{S_M\}\}, \\ & \{\{S_M\}, \{S_H\}\}, \\ & \{\{S_H\}, \{S_L\}\}, \\ & \{\{S_H\}, \{S_M\}\}, \\ & \{\{S_H\}, \{S_H\}\} \\ & \} \end{aligned} \tag{7.1}$$

Siendo similar para $D_{S3}$ y $D_{S4}$. Mientras que para $D_{S1}$ tendríamos 27 perfiles.

Para cada perfil de distinción, hay un perfil de respuesta óptimo que denotaremos $C_{min}(k)$ y que hace referencia al mínimo coste posible que puede pagar un autor al elegir una palabra clave. Hay que tener en cuenta que, según el campo de estudio o la estructura ontológica de la palabra clave escogida, tanto el coste mínimo como el máximo pueden variar.

Denominamos $\pi(k)$ a la función de recompensa obtenida por un autor al escoger una palabra clave:

$$\pi(k) = \begin{cases} 1 & \text{si} \qquad C(k) = C_{min}(k) \\ 0 & \text{en otro caso} \end{cases}$$

Al escoger una palabra clave, esta se enmarcará en un perfil de respuesta determinado con una frecuencia determinada. Siendo $f$ la frecuencia de $S_L$, $g$ la frecuencia de $S_M$ y $1 - f - g$ la frecuencia de $S_H$.

### 7.1.3. Modelo Attention

Una palabra clave se puede clasificar en base a su posición en términos de Attention, $A$, en las siguientes categorías:

- $A_L \longrightarrow$ La palabra clave se encuentra en el tercer tercil en términos de Attention.
- $A_M \longrightarrow$ La palabra clave se encuentra en el segundo tercil en términos de Attention.
- $A_H \longrightarrow$ La palabra clave se encuentra en el primer tercil en términos de Attention.

El conjunto de categorías de una palabra clave en el Modelo Attention se expresa de la siguiente forma: $K_A = \{A_L, A_M, A_H\}$.

Un autor puede tener diferentes perfiles según su capacidad para distinguir el nivel de Attention de las palabras clave. Podemos representar esta capacidad mediante una aplicación biyectiva, $f : D \to P$. donde $D$ es el dominio de los perfiles de distinción de un autor y $P$ su partición.

$$
\begin{aligned}
D_{A1} &\longrightarrow \{\{A_L\}, \{A_M\}, \{A_H\}\} \\
D_{A2} &\longrightarrow \{\{A_L, A_M\}, \{A_H\}\} \\
D_{A3} &\longrightarrow \{\{A_L, A_H\}, \{A_M\}\} \\
D_{A4} &\longrightarrow \{\{A_M, A_H\}, \{A_L\}\} \\
D_{A5} &\longrightarrow \{\{A_L, A_m, AS_H\}\}
\end{aligned}
$$

Para cada perfil; un autor tendrá, con una probabilidad determinada, un comportamiento que denominamos perfil de respuesta. Con cierta probabilidad, un autor con perfil de distinción $D_{A5}$ elegirá una palabra clave de tipo $AL$, $AM$ o bien, $AH$. Por ejemplo, para $D_{A5}$ existen tres perfiles de respuesta diferentes:

$$D_{A5} = \{\{A_L\}, \{A_M\}, \{A_H\}\}$$

Para el caso de $D_{A2}$ tendríamos nueve perfiles de respuesta:

$$
\begin{aligned}
D_{A2} = \{ & \\
& \{\{A_L\}, \{A_L\}\}, \\
& \{\{A_L\}, \{A_H\}\}, \\
& \{\{A_M\}, \{A_L\}\}, \\
& \{\{A_M\}, \{A_M\}\}, \\
& \{\{A_M\}, \{A_H\}\}, \\
& \{\{A_H\}, \{A_L\}\}, \\
& \{\{A_H\}, \{A_M\}\}, \\
& \{\{A_H\}, \{A_H\}\} \\
& \}
\end{aligned}
\tag{7.2}
$$

Siendo similar para $D_{A3}$ y $D_{A4}$. Mientras que para $D_{A1}$ tendríamos 27 perfiles.

Para cada perfil de distinción, hay un perfil de respuesta óptimo que denotaremos $C_{min}(k)$ y que hace referencia al mínimo coste posible que puede pagar un autor al elegir una palabra clave. Hay que tener en cuenta que, según el campo de estudio o la estructura ontológica de la palabra clave escogida, tanto el coste mínimo como el máximo pueden variar.

Denominamos $\pi(k)$ a la función de recompensa obtenida por un autor al escoger una palabra clave:

$$
\pi(k) = \begin{cases} 1 & \text{si} \qquad C(k) = C_{min}(k) \\ 0 & \text{en otro caso} \end{cases}
$$

Al escoger una palabra clave, esta se enmarcará en un perfil de respuesta determinado con una frecuencia determinada. Siendo $f$ la frecuencia de $A_L$, $g$ la frecuencia de $A_M$ y $1 - f - g$ la frecuencia de $A_H$.

### 7.1.4. Modelo Attention/Survival

Una palabra clave se puede clasificar en base a su posición en términos de Attention y Survival, $AS$, en las siguientes categorías:

- $AS_L \longrightarrow$ La palabra clave se encuentra en el tercer tercil en términos de puntuación AS.
- $AS_M \longrightarrow$ La palabra clave se encuentra en el segundo tercil en términos de puntuación AS.
- $AS_H \longrightarrow$ La palabra clave se encuentra en el primer tercil en términos de puntuación AS.

El conjunto de categorías de una palabra clave en el Modelo Attention/-Survival se expresa de la siguiente forma: $K_{AS} = \{AS_L, AS_M, AS_H\}$.

Un autor puede tener diferentes perfiles según su capacidad para distinguir el nivel de Attention/Survival de las palabras clave. Podemos representar esta capacidad mediante una aplicación biyectiva, $f : D \to P$. donde $D$ es el dominio de los perfiles de distinción de un autor y $P$ su partición.

$$\begin{aligned} D_1AS &\longrightarrow \{\{AS_L\},\{AS_M\},\{AS_H\}\} \\ D_2AS &\longrightarrow \{\{AS_L, AS_M\},\{AS_H\}\} \\ D_3AS &\longrightarrow \{\{AS_L, AS_H\},\{AS_M\}\} \\ D_4AS &\longrightarrow \{\{AS_M, AS_H\},\{AS_L\}\} \\ D_5AS &\longrightarrow \{\{AS_L, AS_M, AS_H\}\} \end{aligned}$$

Para cada perfil; un autor tendrá, con una probabilidad determinada, un comportamiento que denominamos perfil de respuesta. Con cierta probabilidad, un autor con perfil de distinción $D_{A5}$ elegirá una palabra clave de tipo $AS_L$, $AS_M$ o bien, $AS_H$. Por ejemplo, para $D_5AS$ existen tres perfiles de respuesta diferentes:

$$D_5AS = \{\{AS_L\},\{AS_M\},\{AS_H\}\}$$

Para el caso de $D_2AS$ tendríamos nueve perfiles de respuesta:

$$\begin{aligned} D_2AS = \{ & \\ & \{\{AS_L\},\{AS_L\}\}, \\ & \{\{AS_L\},\{AS_H\}\}, \\ & \{\{AS_M\},\{AS_L\}\}, \\ & \{\{AS_M\},\{AS_M\}\}, \\ & \{\{AS_M\},\{AS_H\}\}, \\ & \{\{AS_H\},\{AS_L\}\}, \\ & \{\{AS_H\},\{AS_M\}\}, \\ & \{\{AS_H\},\{AS_H\}\}\} \end{aligned} \tag{7.3}$$

Siendo similar para $D_3AS$ y $D_4AS$. Mientras que para $D_{1AS}$ tendríamos 27 perfiles.

Para cada perfil de distinción, hay un perfil de respuesta óptimo que denotaremos $C_{min}(k)$ y que hace referencia al mínimo coste posible que puede pagar un autor al elegir una palabra clave. Hay que tener en cuenta que, según el campo de estudio o la estructura ontológica de la palabra clave escogida, tanto el coste mínimo como el máximo pueden variar.

Denominamos $\pi(k)$ a la función de recompensa obtenida por un autor al escoger una palabra clave:

$$\pi(k) = \begin{cases} 1 & \text{si} \qquad C(k) = C_{min}(k) \\ 0 & \text{en otro caso} \end{cases}$$

Al escoger una palabra clave, esta se enmarcará en un perfil de respuesta determinado con una frecuencia determinada. Siendo $f$ la frecuencia de $AS_L$, $g$ la frecuencia de $AS_M$ y $1 - f - g$ la frecuencia de $AS_H$.

### 7.1.5. Diseño experimental

Para probar el modelo se plantea la realización de un experimento participado por personas que cumplan los siguientes requisitos:

- Estudios finalizados en el ámbito de las Ciencias de la Computación.
- Máster en Ciencias de la Computación o experiencia laboral en este área.
- Nivel de inglés nativo o equivalente.

Los participantes leerán el título y el abstract de un artículo y podrán elegir entre un listado de palabras clave cerrado (Ver akkp69000). Obteniendo así el perfil de cada usuario. De manera similar a como se realizó en el experimento del artículo Sección 3.4.

# Capítulo 8

# Análisis y Discusión de resultados obtenidos

En esta Sección pretendemos señalar los principales resultados obtenidos y la aportación al conocimiento realizada durante el desarrollo de la presente tesis doctoral. Dada la estructura temática de la misma, se hace necesario dividir la discusión sobre resultados en dos áreas fundamentales: aportaciones al proceso de revisión por pares y aportaciones al estudio sobre palabras clave.

## 8.1. Revisión por pares

En el contexto de esta tesis doctoral, un total de cinco artículos han sido publicados analizando procesos de revisión por pares. Estos artículos suponen una continuidad a la investigación desarrollada por el Grupo de Visión por Computador de la Universidad de Granada y pretenden arrojar luz sobre diferentes aspecto de este proceso de garantía de la calidad de las aportaciones a las ciencia.

Hay que señalar, en primer lugar, que el proceso de revisión por pares no está exento de problemas y dificultades. Este hecho ya lo hemos señalado tanto en la presente memoria de tesis como en diferentes artículos, pero es necesario tenerlo presente a la hora de contextualizar nuestros trabajos.

La revisión por pares debe lidiar con numerosos sesgos que ponen en peligro su papel de garante de calidad del conocimiento científico. Uno de los principales retos a los que nos enfrentamos es la capacidad de una persona de identificar correctamente la calidad de un trabajo. Los autores, especialmente los más inexpertos que no conocen bien un área de la ciencia determinada, pueden tomar decisiones erróneas cuando juzgan la calidad de un artículo, lo que les puede conducir a resultados no deseados a la hora de realizar análisis del Estado del Arte de una determinada cuestión o a la hora de identificar

la calidad de sus propios artículos. Esto les puede llevar a tomar decisiones desafortunadas a la hora de elegir la revista donde sus trabajos deben ser publicados, lo que puede generar frustración y a invertir más tiempo del esperado en conseguir la publicación de un artículo.

Pero el proceso de revisión de pares no solo depende de la percepción de la calidad de un autor, sino que también se ve afectado por la capacidad de editores y revisores a la hora de juzgar la calidad de un trabajo. Lo que puede llevar a los editores a comportarse como asesinos o como fanáticos, tal y como explicábamos en el artículo *An evolutionary explanation of assassins and zealots in peer review*. En este punto del análisis, puede ser recomendable la lectura de (Huang, 2013), que analiza la influencia del *Efecto Dunning-Kruger* en el proceso de revisión, el cual puede afectar a todas las partes involucradas en la revisión por pares.

En segundo lugar, debemos mencionar que detenerse a mencionar estos problemas no supondría, por sí mismo, una contribución al conocimiento, ya que las debilidades del proceso de revisión por pares son sobradamente conocidas. Nuestros trabajos pretenden arrojar información sobre los diferentes roles que se adoptan en una revisión y las consecuencias resultado de la falta de conocimiento y los sesgos de los autores, así como proponer estrategias para atenuar estas consecuencias. Por ejemplo, a tenor de los resultados observados en *An evolutionary explanation of assassins and zealots in peer review* podemos deducir que la distribución de documentos a revisar en base a su calidad es un factor que maximiza los resultados de un proceso de revisión: Si un revisor es capaz de identificar correctamente la calidad de un artículo, un reparto equilibrado maximiza resultados. Si el revisor no es capaz de identificar correctamente la calidad de los artículos, obtendremos mejores resultados cuando tenemos una colección de manuscritos desequilibrada.

Los procesos de revisión pares dependen en cierta medida de la sensibilidad y de la especificidad, conceptos propuestos en *What is the sensitivity and specificity of the peer review process?* y que son fruto de las ideas preconcebidas que tienen los editores acerca de los manuscritos que les son enviados a las revistas bajo su responsabilidad. El incremento de la especificidad permite a los editores confiar en las revisiones favorables.

Hasta ahora, hemos hablado de sesgos a la hora de identificar la calidad de un artículo por parte de una persona, pero no nos habíamos adentrado en esta cuestión hasta la hora de publicar *The Relevance of Title, Abstract, and Keywords for scientific paper quality and potential impact* donde analizamos más en detalle esta cuestión, advirtiendo de cómo la lectura del título, abstract y palabras clave de un artículo puede suponer la cantidad mínima de información necesaria para que un investigador juzgue la calidad de un trabajo. En este artículo, donde hemos contado con 106 participantes que han colaborado en un experimento, hemos podido determinar esta informa-

ción y estimar las particiones y perfiles de envío de las personas que han colaborado en el experimento. Siendo un artículo clave que permite afianzar los conceptos manejados en el resto de trabajos del Capítulo 3.

En todo momento, hablamos de costes: El coste de enviar un artículo a una revista equivocada es más alto que el que asumimos al enviarlo a la revista correcta. El coste que paga un editor que no confía en el proceso de revisión es más elevado que el que debe pagar cuando sí confía en él. Esta es otra dimensión que habla de la importancia de los temas tratados en la tesis. Detrás de cada decisión tomada en un proceso de revisión por pares, no solamente hay un impacto en la calidad del conocimiento científico, sino también un impacto económico. Quizás para trabajos futuros, sería interesante cuantificar el impacto que asumen los grupos de investigación debido a los artículos que les son rechazados o a las revisiones *problemáticas*. Actualmente, una búsqueda bibliográfica sencilla parece determinar que no hay demasiado escrito a este respecto. Si bien podemos encontrar estudios que analizan el impacto económico de enviar manuscritos a revistas *depredadoras* (Sureda-Negre et al., 2022), hecho que, en muchos casos, obedece a la incapacidad de un autor de identificar correctamente la calidad de una revista.

Finalmente, solo queda resaltar la versatilidad del modelo de cuasi especie y su potencial de ser aplicado para el estudio de diferentes áreas y fenómenos. En esta tesis, hemos hecho uso de este modelo para el estudio de la revisión por pares. A efectos de trabajo futuro, estamos analizando la posibilidad de utilizarlo para el estudio de las palabras clave, como queda patente en el Capítulo 7.

## 8.2. Palabras clave

La otra gran línea de investigación desarrollada en esta tesis se corresponde al estudio de las palabras clave en la ciencia y, concretamente, su rol dentro de los manuscritos y los artículos científicos.

Actualmente, las palabras clave son una herramienta de categorización de información que no está exenta de problemas y sesgos que la apartan de su objetivo de facilitar la recuperación de la información. Además, a veces las palabras clave se utilizan de una manera inadecuada, utilizando aquellas que garantizan la maximización del impacto del artículo frente a la categoría real del mismo, como han estudiado los autores del artículo (González et al., 2018) dentro del campo de las Ciencias del Deporte.

En nuestro caso, nos centramos en el análisis de las redes que generan las palabras clave, ya que éstas también pueden ser utilizadas para entender la estructura de la ciencia. Las palabras clave co-ocurren en artículos, en autores, en corpus y en áreas de la ciencia y su estudio nos permite extraer conclusiones. Las palabras clave también están sujetas a fluctuaciones en base

al interés y el foco de la comunidad científica: unas palabras clave emergen ante desarrollos tecnológicos o eventos sociales, políticos o culturales. Otras palabras clave caen en el olvido. Dentro de esta dinámica, el autor debe elegir correctamente las palabras clave que definen su artículo si quiere que éste pueda maximizar las posibilidades de ser leído y de recibir las tan ansiadas citas por parte de otros trabajos científicos. En este contexto, elegir palabras muy genéricas o muy populares no siempre suponen una apuesta segura, puesto que el autor no es el único que ha pensado en categorizar sus trabajos con estas palabras y, probablemente, su artículo tenga que competir con otros trabajos para tener un puesto privilegiado en los buscadores académicos. De igual modo, una palabra extremadamente específica puede reducir las posibilidades de que un artículo sea localizado.

Es en este complicado equilibrio de fuerzas donde hemos centrado nuestro trabajo sobre la visibilidad de las palabras clave: En nuestro artículo, estudiamos estas circunstancias y proponemos una métrica y una metodología de selección de palabras clave. Consideramos que la métrica AT-IDF es una alternativa a TF-IDF que tiene en cuenta la demanda social actual de una palabra clave. En estos momentos, las llamadas métricas alternativas (*altmetrics*) cobran una relevancia cada vez mayor y deben ser tenidas en cuenta en la ecuación (Luc et al., 2021). AT-IDF puede ayudar a los autores a otorgar a diferentes palabras clave una puntuación en base a la demanda social de un determinado concepto. Por otro lado, nuestra métrica permite categorizar no solo documentos de texto, sino cualquier entidad sujeta a categorización como podría ser un archivo de vídeo o una imagen.

Otra perspectiva sobre las palabras clave se puede hacer desde el punto de vista de la recuperación de información. En el año 2020, en el contexto de la pandemia de COVID-19, propusimos un método para aprovechar la información obtenida de, no solamente las palabras clave que un usuario introduce en un buscador o base de datos, sino también de aquellas que son intrínsecas a un documento, que son invariables del contexto de la búsqueda. De esta forma, conseguimos localizar documentos que tienen información sobre temas muy específicos, esto era algo reclamado en aquel momento por diferentes organismos e instituciones públicas como la Casa Blanca. Sin embargo, tres años después de realizar la propuesta de este método, nuestro artículo *Finding answers to COVID-19 specific questions: An Information Retrieval System based on latent keywords and adapted TF-IDF* ha recibido citas por parte de otros trabajos, demostrando así que puede ser aplicable a otros contextos y relevante más allá del momento concreto en el que se propuso.

Finalmente, otro aspecto importante que relaciona las palabras clave con el comportamiento de los autores son las palabras clave de autor, así como las Keywords Plus. Utilizando un análisis de redes multicapa se consigue identificar los nodos centrales, es decir, las palabras clave más destacadas

y que sirven de punto de conexión entre el resto de la red y con respecto a otras redes de palabras clave. Si bien existen otros algoritmos que ya abordan esta cuestión, tales como APABI y PageRank versatility, nuestro algoritmo consigue resultados similares, pero reduciendo los requisitos de computación.

## 8.3. Comentarios finales

No es posible finalizar esta sección de discusión sin mencionar la tercera categoría de trabajos enmarcados en esta tesis, un artículo publicado en una revista interdisciplinar enfocada en las Ciencias Sociales y diferentes aplicaciones y datasets. Estos trabajos son también importantes y necesarios para sostener el marco teórico o experimental sobre el que se asientan los artículos publicados en las otras dos categorías así como, en el caso de los datasets, fomentar la replicabilidad de los datos y proporcionar a la comunidad científica de información que puede ser útil para hipotéticas investigaciones.

Quedan, además, una serie de trabajos sobre palabras clave que no han sido publicados ni han visto la luz. Se corresponden, generalmente, a estudios sobre redes generadas por palabras clave de autor y su comparación con las KeyWords Plus, que supusieron los primeros intentos de investigar en este campo pero que finalmente decidimos no publicar, al menos de momento. Entre los principales motivos de esta decisión se encuentran, entre otros, la dificultad de encontrar una revista que se pudiera adaptar al foco de esta investigación o que la realización de las pruebas pedidas por los revisores y cuyo desarrollo hubiera cambiado completamente el rumbo de esta tesis. En su lugar, el desarrollo de estos trabajos iniciales dieron lugar a que surgieran nuevas preguntas de investigación más factibles y que han desembocado en los trabajos finalmente realizados durante estos años.

# Capítulo 9

# Conclusiones

Con la realización de esta tesis, me marqué diferentes objetivos más allá de la formación como investigador y la colaboración con un grupo de Investigación. Considero la informática como una herramienta al servicio del Conocimiento, la Tecnología y la Ciencia y no, únicamente, como un *fin*. En este sentido, durante estos años he intentado colocar las ciencias de la computación al servicio, especialmente, de las ciencias sociales; concretamente de la sociología. Considero que las ciencias sociales pueden beneficiarse de las ciencias de la computación enormemente, mucho más de lo que ocurre a día de hoy. Por fortuna, la investigación interdisciplinar se encuentra en auge, y cada vez hay más voces que ponen foco en este asunto (en este sentido, recomiendo la lectura de sobre los retos que aún se han de afrontar en el campo de la ciencia social computacional (Lazer et al., 2020)).

Los modelos que se utilizan en los artículos, tienen la ventaja de ser sumamente generalizables y de poder ser aplicados a diferentes ámbitos. Posiblemente, Manfred Eigen y Peter Schuster (Eigen y Schuster, 1977) nunca pudieron imaginar que su modelo de cuasi especie pudiera ser aplicado en el campo de la teoría de juegos por Friederike Mengel (quien, por cierto, no había nacido cuando fue propuesto el modelo de cuasi especie) casi tres décadas después (Mengel, 2012). Seguramente, también sería difícil de imaginar que este mismo modelo pudiera terminar sirviendo para estudiar el proceso de revisión por pares y, probablemente, el proceso de selección de palabras claves de autor, siempre y cuando que nuestro trabajo presentado en el Capítulo 7 sea finalizado y aceptado en un proceso de revisión por pares.

En el mundo de la programación, tratamos de crear componentes que sean reutilizables por otras personas y aplicables en diferentes contextos. Tenemos estos ejemplos tanto en los patrones de diseño, a un nivel más abstracto (Gamma et al., 1994) como en la mera existencia de GitHub o de proyectos como Maven [1] o Pypi[2], entre otros. Con esta misma filosofía, los trabajos

[1] https://maven.apache.org/ (Accedida el 9 de Abril del 2023).
[2] https://pypi.org/ (Accedida el 9 de Abril del 2023).

aquí realizados contienen metodologías y modelos que están pensados para favorecer su reutilización y ser aplicados en otros ámbitos de la ciencia. Las revistas donde hemos publicado, por lo general, son todas de corte multidisciplinario; lo que se corresponde, de manera cercana, con mi visión sobre cómo debe ser la ciencia y cómo debe trabajar la comunidad científica. La gran ventaja del conocimiento multidisciplinar es que, sin tener conocimientos de Biología y sin ser un experto en el modelo de cuasi especie, he podido aprovechar su potencial y del conocimiento de otras áreas de la ciencia del mismo modo que, sin ser experto en compiladores, puedo beneficiarme de los avances que Python ha realizado a su intérprete por personas como el ingeniero de compilación Pablo Galindo Salgado (Galindo-Salgado, 2022), que ha jugado un papel clave en la construcción del software necesario para llevar a cabo las aplicaciones y experimentos de esta tesis.

Fruto de esta tesis, se han producido un total de diez trabajos científicos: nueve de ellos ya publicados, y un tercero que está aún por publicar, a falta de realizar las correspondientes pruebas experimentales. Además, se han producido tres *datasets* y se ha asistido a un Congreso.

La tesis se ha desarrollado dentro del Grupo de Visión por Computador de la Universidad de Granada, aunque puntualmente también se ha redactado un artículo en colaboración con un investigador de la Universidad Nacional de Educación a Distancia (UNED).

Los resultados hasta la fecha (2023) arrojan la cifra de 13 citas y un índice h = 2.

Con este trabajo, espero haber aportado un granito de arena al conocimiento científico, especialmente en las dos áreas más relevantes de esta tesis: revisión por pares y palabras clave.

# Bibliografía


Aczel, B., Szaszi, B. y Holcombe, A. O. A billion-dollar donation: estimating the cost of researchers'time spent on peer review. *Research Integrity and Peer Review*, vol. 6, página 14, 2021. ISSN 2058-8615.

Adhikari, G. R. Demystifying the peer review process of manuscripts submitted to journals. *Bulletin of Nepal Geological Society*, vol. 38, páginas 131–136, 2021.

Agryzkov, T., Curado, M., Pedroche, F., Tortosa, L. y Vicent, J. Extending the adapted pagerank algorithm centrality to multiplex networks with data using the pagerank two-layer approach. *Symmetry*, vol. 11, página 284, 2019. ISSN 2073-8994.

Aizawa, A. An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, vol. 39(1), páginas 45–65, 2003. ISSN 03064573.

Arora, A. y Arora, A. Synthetic patient data in health care: a widening legal loophole. *The Lancet*, página 1601, 2022. ISSN 01406736.

Berman, L. S., Klusse, T. M. y López, L. P. Revisión por pares: evidencias y desafíos. *Revista chilena de pediatría*, vol. 88, páginas 577–581, 2017. ISSN 0370-4106.

BioMed Central. Peer review process. `https://www.biomedcentral.com/getpublished/peer-review-process`, 2023. Accedida el 19/03/2023.

Björk, B.-C. y Solomon, D. Pricing principles used by scholarly open access publishers. *Learned Publishing*, vol. 25, páginas 132–137, 2012. ISSN 09531513.

Breiger, R. L. y Pattison, P. E. Cumulated social roles: The duality of persons and their algebras. *Social Networks*, vol. 8, páginas 215–256, 1986. ISSN 03788733.

BURNHAM, J. C. The evolution of editorial peer review. *JAMA: The Journal of the American Medical Association*, vol. 263, página 1323, 1990. ISSN 0098-7484.

CARLIN, B. I. Strategic price complexity in retail financial markets. *Journal of Financial Economics*, vol. 91, páginas 278–287, 2009. ISSN 0304405X.

CASTRO, L. R. y CASTRO, S. M. Wavelets y sus aplicaciones. *I Congreso Argentino de Ciencias de la Computación*, páginas 195–204, 1995.

CHAMORRO PADIAL, J. Ingeniería de servidores desde la perspectiva de un estudiante. *Enseñanza y Aprendizaje de Ingeniería de Computadores*, 2014.

CHAMORRO-PADIAL, J., RODRIGO-GINÉS, F.-J. y RODRÍGUEZ-SÁNCHEZ, R. Finding answers to covid-19-specific questions: An information retrieval system based on latent keywords and adapted tf-idf. *Journal of Information Science*, página 016555152211109, 2022. ISSN 0165-5515.

CHAMORRO-PADIAL, J. y RODRÍGUEZ-SÁNCHEZ, R. Clasificación de texto. utilizando métricas de ganancia de información para categorizar disposiciones legales. *Revista Internacional de Tecnología, Conocimiento y Sociedad*, vol. 7, páginas 37–48, 2019. ISSN 2474-588X.

CHAMORRO-PADIAL, J. y RODRÍGUEZ-SÁNCHEZ, R. akkp69000. `https://www.kaggle.com/ds/856525`, 2020a. Accedido el 19/03/2023.

CHAMORRO-PADIAL, J. y RODRÍGUEZ-SÁNCHEZ, R. akkp69000. `https://data.mendeley.com/datasets/rjhh6cm55z`, 2020b. Accedido el 19/03/2023.

CHAMORRO-PADIAL, J. y RODRÍGUEZ-SÁNCHEZ, R. Computer science articles & journals, 2019. `https://www.kaggle.com/ds/1268595`, 2020c. Accedido el 19/03/2023.

CHAMORRO-PADIAL, J. y RODRÍGUEZ-SÁNCHEZ, R. Text categorisation through dimensionality reduction using wavelet transform. *Journal of Information & Knowledge Management*, vol. 19, página 2050039, 2020d. ISSN 0219-6492.

CHAMORRO-PADIAL, J. y RODRÍGUEZ-SÁNCHEZ, R. Attention –survival score: A metric to choose better keywords and improve visibility of information. *Algorithms*, vol. 16(4), 2023a. ISSN 1999-4893.

CHAMORRO-PADIAL, J. y RODRÍGUEZ-SÁNCHEZ, R. The relevance of title, abstract, and keywords for scientific paper quality and potential impact. *Multimedia Tools and Applications*, 2023b. ISSN 1380-7501.

CHAMORRO-PADIAL, J., RODRIGUEZ-SÁNCHEZ, R., FDEZ-VALDIVIA, J. y GARCÍA, J. A. An evolutionary explanation of assassins and zealots in peer review. *Scientometrics*, vol. 120, páginas 1373–1385, 2019. ISSN 0138-9130.

COBEY, K. D., LALU, M. M., SKIDMORE, B., AHMADZAI, N., GRUDNIEWICZ, A. y MOHER, D. What is a predatory journal? A scoping review. *F1000Research*, vol. 7, página 1001, 2018. ISSN 2046-1402.

DEBNATH, L. *The Wavelet Transform and Its Basic Properties*, páginas 361–402. Birkhäuser Boston, 2002.

DOMENICO, M. D., SOLÉ-RIBALTA, A., OMODEI, E., GÓMEZ, S. y ARENAS, A. Ranking in interconnected multilayer networks reveals versatile nodes. *Nature Communications*, vol. 6, página 6868, 2015. ISSN 2041-1723.

DUQUE, E. Análisis de contenido mediante análisis de palabras clave: La representación de los participantes en los discursos de Esperanza Aguirre. *Mediaciones Sociales*, vol. 0(13), páginas 39–73, 2015.

EIGEN, M. y SCHUSTER, P. A principle of natural self-organization. *Naturwissenschaften*, vol. 64, páginas 541–565, 1977. ISSN 0028-1042.

EMILE, S. H., HAMID, H. K. S., ATICI, S. D., KOSKER, D. N., PAPA, M. V., ELFEKI, H., TAN, C. Y., EL-HUSSUNA, A. y WEXNER, S. D. Types, limitations, and possible alternatives of peer review based on the literature and surgeonsópinions via twitter: a narrative review. *Science Editing*, vol. 9, páginas 3–14, 2022. ISSN 2288-8063.

FEYERABEND, P. *Against Method: Outline of an Anarchistic Theory of Knowledge*. New Left Books, 1975. ISBN 0902308912.

FLAHERTY, C. The peer-review crisis. *Inside Higher Education*, 2022.

GALINDO-SALGADO, P. Python 3.11.1rc1 is now available. `https://web.archive.org/web/20221005111712/https://discuss.python.org/t/python-3-11-1rc1-is-now-available/18068`, 2022. Accedida el 19/03/2023.

GAMMA, E., HELM, R., JOHNSON, R. y VLISSIDES, J. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1994.

GARCIA, J., FDEZ-VALDIVIA, J., FDEZ-VIDAL, X. y RODRIGUEZ-SANCHEZ, R. Information theoretic measure for visual target distinctness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, páginas 362–383, 2001. ISSN 01628828.

GARCÍA, J. A., CHAMORRO-PADIAL, J., RODRIGUEZ-SÁNCHEZ, R. y FDEZ-VALDIVIA, J. What is the sensitivity and specificity of the peer review process? *Accountability in Research*, páginas 1–22, 2022. ISSN 0898-9621.

GARCÍA, J. A., RODRIGUEZ-SÁNCHEZ, R., FDEZ-VALDIVIA, J. y CHAMORRO-PADIAL, J. The author's ignorance on the publication fees is a source of power for publishers. *Scientometrics*, vol. 121, páginas 1435–1445, 2019. ISSN 0138-9130.

GARCIA-COSTA, D., FORTE, A., LÒPEZ-IÑESTA, E., SQUAZZONI, F. y GRIMALDO, F. Does peer review improve the statistical content of manuscripts? A study on 27 467 submissions to four journals. *Royal Society Open Science*, vol. 9(9), página 210681, 2022. ISSN 2054-5703.

GARFIELD, E. Keywords plus-isi's breakthrough retrieval method. 1. expanding your searching power on current-contents on diskette. *Current contents*, vol. 32, páginas 5–9, 1990.

GODDARD, C. y WIERZBICKA, A. *Meaning and Universal Grammar: Theory and empirical findings*. John Benjamins Publishing Company, 2002. ISBN 9789027281876.

GONZÁLEZ, L. M., GARCÍA-MASSÓ, X., NEZ, A. P.-I., PESET, F. y DEVÍS-DEVÍS, J. An author keyword analysis for mapping sport sciences. *PLoS ONE*, vol. 13, páginas 1–22, 2018. ISSN 19326203.

HAAR, A. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, vol. 69, páginas 331–371, 1910. ISSN 0025-5831.

HUANG, S. When peers are not peers and don't know it: The dunning-kruger effect and self-fulfilling prophecy in peer-review. *BioEssays*, vol. 35, páginas 414–416, 2013. ISSN 02659247.

IMRAN, H. y SHARAN, A. *Genetic Algorithm Based Model for Effective Document Retrieval*, páginas 191–201. Springer Netherlands, Dordrecht, 2011. ISBN 978-94-007-0286-8.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. Documentation – Methods for examining documents, determining their subjects, and selecting indexing terms. Standard, International Organization for Standardization, Geneva, CH, 1985.

JUBB, M. Heading for the open road: Costs and benefits of transitions in scholarly communications. 2008.

KRONICK, D. A. Peer review in 18th-century scientific journalism. *JAMA: The Journal of the American Medical Association*, vol. 263, página 1321, 1990. ISSN 0098-7484.

Labbé, C. y Labbé, D. Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? *Scientometrics*, vol. 94(1), páginas 379–396, 2013. ISSN 0138-9130, 1588-2861.

Lakatos, I. Understanding Toulmin. En *Mathematics, Science and Epistemology* (editado por J. Worrall y G. Currie), páginas 224–244. Cambridge University Press, 1 edición, 1978. ISBN 978-0-521-21769-9 978-0-521-28030-3 978-0-511-62492-6.

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A. y Wagner, C. Computational social science: Obstacles and opportunities. *Science*, vol. 369, páginas 1060–1062, 2020. ISSN 0036-8075.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. y Liu, H. Feature selection: A data perspective. *ACM Computing Surveys*, vol. 50, páginas 1–45, 2018. ISSN 0360-0300.

Luc, J. G., Archer, M. A., Arora, R. C., Bender, E. M., Blitz, A., Cooke, D. T., Hlci, T. N., Kidane, B., Ouzounian, M., Varghese, T. K. y Antonoff, M. B. Does tweeting improve citations? one-year results from the tssmn prospective randomized trial. *The Annals of Thoracic Surgery*, vol. 111, páginas 296–300, 2021. ISSN 0003-4975.

Mengel, F. On the evolution of coarse categories. *Journal of Theoretical Biology*, vol. 307, páginas 117–124, 2012. ISSN 00225193.

Minini, A. La primitiva semantica su Google. `https://web.archive.org/web/20160126115726/https://www.andreaminini.com/semantica/la-primitiva-semantica-su-google`, 2013. Accedida el 22/04/2023.

Moustafa, K. No to paid peer review. *The Lancet*, página 160, 2022. ISSN 01406736.

Munévar, G. Consenso y evolución en ciencia. *Praxis Filosófica*, 2005. ISSN 0120-4688.

Newton, D. P. Quality and peer review of research: An adjudicating role for editors. *Accountability in Research*, vol. 17, páginas 130–145, 2010. ISSN 0898-9621.

Osorio-Sánchez, A. Algoritmo para detección de vibraciones anormales en maquinarias utilizando la transformada wavelet. 2006. Tesis profesional.

Overington, M. A. The Scientific Community as Audience: Toward a Rhetorical Analysis of Science. *Philosophy and Rhetoric*, vol. 10(3), páginas 143–164, 1977.

PALACIOS LÓPEZ, F. y CHAMORRO PADIAL, J. Aplicacíón de las materias de ingeniería de computadores en la mejora de los algoritmos meméticos y metaheurísticas en general. *Enseñanza y Aprendizaje de Ingeniería de Computadores*, 2015.

PEARLMAN, W. A. y SAID, A. Image wavelet coding systems: Part ii of set partition coding and image wavelet coding systems. *Foundations and Trends®in Signal Processing*, vol. 2, páginas 181–246, 2007. ISSN 1932-8346.

POLANYI, M. *Personal knowledge: Towards a Post-Critical Philosophy*. Routledge and Kegan Paul Ltd, 1962. ISBN 020375039X.

POPPER, K. R. *Conjeturas y Refutaciones*. Ediciones PAIDOS, 1962. ISBN 8475091466.

PORTER, M. An algorithm for suffix stripping. *Program*, vol. 14, páginas 130–137, 1980. ISSN 0033-0337.

PRINCETON UNIVERSITY. About wordnet. `https://wordnet.princeton.edu/`, 2010. Accedida el 19/03/2023.

REAL ACADEMIA ESPAÑOLA. Palabra clave. `https://dle.rae.es/palabra#Btac3m5`, 2021. Accedida el 19/03/2023.

REALE, G. y ANTISERI, D. *Historia del pensamiento filosófico y científico*, vol. 3, capítulo XXXVII. Herder, 1988. ISBN 8415415918.

RODRÍGUEZ-SÁNCHEZ, R. y PADIAL, J. C. Author keywords - keywordsplus. `https://www.kaggle.com/datasets/jorgechamorropadial/author-keywords-keywordsplus`, 2022. Accedido el 22/03/2023.

RODRÍGUEZ-SÁNCHEZ, R. M. y CHAMORRO-PADIAL, J. Corner centrality of nodes in multilayer networks: A case study in the network analysis of keywords. *Algorithms*, vol. 15, página 336, 2022. ISSN 1999-4893.

ROOBAERT, D., KARAKOULAS, G. y CHAWLA, N. V. *Information Gain, Correlation and Support Vector Machines*, páginas 463–470. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-35488-8.

SALATINO, A. A., THANAPALASINGAM, T., MANNOCCI, A., OSBORNE, F. y MOTTA, E. The computer science ontology: A large-scale taxonomy of research areas. En *The Semantic Web – ISWC 2018* (editado por D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee y E. Simperl), páginas 187–205. Springer International Publishing, Cham, 2018. ISBN 978-3-030-00668-6.

SÁNCHEZ-SAUS LASERNA, M. Keyword analysis in communication for development and social change: The case of #comunicambio on Twitter. *Cultura, Lenguaje y Representación*, vol. 19, páginas 119–139, 2018. ISSN 16977750.

SARICA, S. y LUO, J. Stopwords in technical language processing. *PLOS ONE*, vol. 16(8), página e0254937, 2021. ISSN 1932-6203.

SCOTT, M. y TRIBBLE, C. *Textual Patterns: Key words and corpus analysis in language education*. John Benjamins Publishing Company, 2006. ISBN 9789027293633.

SIEGELMAN, S. S. Assassins and zealots: variations in peer review. special report. *Radiology*, vol. 178, páginas 637–642, 1991. ISSN 0033-8419.

SPIER, R. The history of the peer-review process. *Trends in Biotechnology*, vol. 20, páginas 357–358, 2002. ISSN 01677799.

STRADER, C. R. Author-assigned keywords versus library of congress subject headings. *Library Resources & Technical Services*, vol. 53, páginas 243–250, 2011. ISSN 2159-9610.

STRIBLING, J., KROHN, M. y AGUAYO, D. Scigen - an automatic cs paper generator. `https://pdos.csail.mit.edu/archive/scigen/`, 2005. Accedida el 19/03/2023.

SUJATHA, P. y DEVI, R. An overview of digital watermarking with a performance analysis of wavelet families for image compression. *Indian Journal of Science and Technology*, vol. 8, 2015. ISSN 0974-5645.

SUREDA-NEGRE, J., CALVO-SASTRE, A. y COMAS-FORGAS, R. Predatory journals and publishers: Characteristics and impact of academic spam to researchers in educational sciences. *Learned Publishing*, vol. 35(4), páginas 441–447, 2022.

TRUNK, G. V. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, páginas 306–307, 1979. ISSN 0162-8828.

VINES, T. y MUDDIT, A. What's wrong with paying for peer review? *The Scholarly Kitchen*, 2021.

WHITTAKER, J. Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science*, vol. 19(3), páginas 473–496, 1989.

WIERZBICKA, A. *Semantics: Primes and Universals*. Oxford University Press, 1996. ISBN 9780198700029.

WIERZBICKA, A. Cross-cultural communication and miscommunication: The role of cultural keywords. *Intercultural Pragmatics*, vol. 7, páginas 1–23, 2010. ISSN 1612295X.

WILLIAMS, R. *Keywords: A vocabulary of Culture and Society*. Oxford University Press, 1985. ISBN 9780195204698.

XEXÉO, G., DE SOUZA, J., CASTRO, P. F. y PINHEIRO, W. A. Using wavelets to classify documents. En *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, páginas 272–278. 2008.

ZHANG, J. y LUO, Y. Degree centrality, betweenness centrality, and closeness centrality in social network. En *Proceedings of the 2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017)*, páginas 300–303. Atlantis Press, 2017/03. ISBN 978-94-6252-324-1. ISSN 1951-6851.

ZIMAN, J. M. *Public Knowledge: An Essay Concerning the Social Dimension of Science*. Cambridge University Press, 1974. ISBN 9780521095198.

ZUO, Y., LI, C., LIN, H. y WU, J. Topic modeling of short texts: A pseudo-document view with word embedding enhancement. *IEEE Transactions on Knowledge and Data Engineering*, páginas 1–1, 2021. ISSN 1041-4347.