UNIVERSIDAD DE GRANADA

UNIVERSIDAD
DE GRANADA

PROGRAMA DE DOCTORADO EN TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN

Departamento de Ciencias de la Computación e Inteligencia Artificial

---

Análisis multidimensional de datos textuales en redes sociales

---

Tesis Doctoral

Karel Gutiérrez Batista

DIRECTORES:

Dra. María Amparo Vila Miranda y Dra. María José Martín Bautista

Granada, Marzo de 2018

Editor: Universidad de Granada. Tesis Doctorales
Autor: Karel Gutiérrez Batista
ISBN: 978-84-1306-125-2
URI: http://hdl.handle.net/10481/54952

La memoria titulada "Análisis multidimensional de datos textuales en redes sociales", que presenta D. Karel Gutiérrez Batista para optar al grado de Doctor, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección de las Doctoras María Amparo Vila Miranda y María José Martín Bautista.

Granada, Marzo de 2018

El doctorando	Los directores
Karel Gutiérrez Batista	María Amparo Vila Miranda María José Martín Bautista

Dedicatoria

Por fin ha llegado el momento de dedicar este trabajo a las personas más importantes de mi vida. Al hacerlo, pienso en todos los momentos buenos y malos que he tenido que pasar, para lograr este ansiado objetivo, el cual constituye mi mayor éxito profesional dentro de una vida dedicada completamente al estudio.

A mi fiel compañera Nuria, quien con su delicadeza y amor incondicional ha sido mi mejor apoyo en todos estos años de duro trabajo lejos de mi familia y amistades.

A nuestra Kale que representa el primero de nuestros hijos y nuestra razón de vivir.

A mis padres (Melba, Carlos y Alfredo) y hermanos (Karina y Carlitos), por su amor y entrega constante en estos años, los cuales han representado un ejemplo a seguir durante toda mi vida.

A todos mis abuelos, los que están y los que no, por ser motivo de inspiración para toda la familia.

Al resto de toda mi hermosa y unida familia y la familia de Nuria quienes me han apoyado todo el tiempo.

A mis suegros Lourdes y Ramón, por ser personas extremadamente nobles y abnegadas.

A mis amigos de la vida: Toni, Pica, Lázaro, Sergio, Mario, Rietni y sus esposas y familiares.

Agradecimientos

"Cualquier actividad se vuelve creativa si el autor se preocupa de hacerlo bien o de hacerlo mejor"

John Updike

Este momento es un sueño que he esperado durante años, el cual considero el más grande de mi vida en el plano profesional, pero sin la ayuda de las personas más cercanas nunca hubiera sido posible, es por eso que quiero agradecer a todas las personas que me han ayudado a conseguirlo.

Antes que nada quiero dar la gracias a la Junta de Andalucía, la cual mediante los proyectos P11-TIC-7460 y P10-TIC-6109 ha respaldado la investigación presentada en esta memoria.

Quiero agradecer de manera muy especial a mis directoras de tesis María Amparo Vila Miranda y María José Martín Bautista, quienes más que directora de tesis, han sido un ejemplo a seguir durante todo el proceso. Les agradezco por su ayuda incondicional y por su apoyo total en todas las oportunidades que he tenido. Sin duda alguna, trabajar a su lado ha sido un privilegio para mí, y ha cambiado mi vida para siempre.

De forma especial quiero agradecer a María José Martín Bautista, pues además de ser una excelente profesional, sin conocerme me dio la oportunidad de formar parte de su proyecto de investigación y crecer como investigador. Gracias de verdad por TODO.

Agradecer también de forma especial a Jesús Roque Campaña Gómez por su ayuda incondicional en los momentos más difíciles, quien es para mí una persona excepcional y considero que es un profesional como la copa de un pino.

A Sandro Martínez Folgoso y Elizabet Tejeda Ávila por permitirme esta gran oportunidad, por tenerme la suficiente confianza en momentos que resultaba difícil confiar. Además por todo el apoyo incondicional y los acertados consejos que siempre tenían para mí.

A mis compañeros del Departamento de Computación de la Universidad de Camagüey, que han dado un extra para que yo pudiera tener esta oportunidad. De forma especial a Julio, Lisset y Yanaima.

Agradecer también a todas los miembros del Departamento de Ciencias de la Computación e Inteligencia Artificial, a los miembros del grupo de Bases de Datos

y Sistemas de Información Inteligentes (IDBIS) y a mis compañeros del CITIC, en especial a Manu, Byron y Carlos.

También me gustaría agradecer a personas maravillosas e incondicionales que he conocido durante mi estancia en España: Wandel, Antonio, Madiama, Puri y Natalie. La verdad es que sin ellos todo este tiempo hubiese sido mucho más difícil. Gracias por sus consejos, apoyo y principalmente por sacarme siempre una sonrisa en momentos de añoranza. ¡¡¡GRACIAS!!!.

Por último y muy importante, agradecer a José Luis Villena, David, Pepi, Juan, María y al resto de amigos y familiares. Todos personas maravillosas, nobles y con una bondad sin límites. A todos ustedes muchísimas gracias por todo.

Índice general

1. Introducción **1**

1.1. Planteamiento del problema 2
1.2. Marco de trabajo 4
1.3. Objetivos 5
1.4. Aportaciones 6
1.5. Contenidos de la memoria 7

2. Antecedentes **9**

2.1. Detección de tópicos 9
2.2. Análisis multidimensional en redes sociales 11
2.3. Propuesta 14

3. Detección automática de tópicos **15**

3.1. Propuesta metodológica 15
3.1.1. Ideas básicas 15
3.1.2. Definiciones 16
3.1.3. Descripción de la metodología 17
3.1.4. Preprocesamiento sintáctico 22
3.1.5. Preprocesamiento semántico 23
3.1.6. Agrupamiento jerárquico 27
3.1.7. Etiquetado de los grupos 32
3.1.8. Evaluación de la metodología 34
3.2. Experimentos 35
3.2.1. Conjunto de datos 35
3.2.2. Evaluación 37
3.2.3. Resultados y discusión 37
3.3. Conclusiones 43

4. Influencia de los términos con orientación sentimental en la detección de contextos: Análisis en redes sociales **45**

4.1. Marco de trabajo 45

4.2. Filtro para detectar y eliminar los términos con orientación sentimental 47

4.2.1. SentiWordNet 3.0 49

4.2.2. SenticNet 3 51

4.2.3. WordNet Affect 52

4.2.4. Análisis gradual de la polaridad de los términos de sentimientos en la detección de contextos 52

4.3. Experimentos 54

4.3.1. Conjunto de datos 54

4.3.2. Evaluación 55

4.3.3. Resultados y discusión 55

4.4. Conclusiones 83

5. Construcción de una dimensión contextual para el análisis multidimensional de textos de redes sociales **85**

5.1. Estructura de la dimensión contextual 86

5.1.1. Jerarquía de contextos 88

5.1.2. Jerarquía de consulta 89

Conjunto-AP 91

Estructura-AP 92

5.1.3. Descripción del sistema OLAP 93

5.2. Metodología para crear e integrar la dimensión contextual en un modelo multidimensional 96

5.2.1. Módulo para crear la jerarquía de consulta (jerarquía de dominio) 99

Obtención de la estructura-AP global 100

Obtención del atributo-AP 100

5.2.2. Módulo de integración 101

5.3. Experimentos 107

5.3.1. Conjunto de datos 107

5.3.2. Evaluación de la detección automática de contextos 108

5.3.3. Ejemplo práctico 111

5.3.4. Evaluación del rendimiento de las consultas 113
5.4. Conclusiones 115

6. Explotación de la dimensión contextual mediante el servidor OLAP "Wonder OLAP Server 3.0" 117
6.1. Descripción del sistema 117
6.1.1. Arquitectura general 118
6.1.2. Arquitectura de Wonder 3.0 119
6.2. Definición de un cubo contextual en Wonder OLAP Server 3.0 121
6.2.1. Creación del cubo de datos con dimensión-AP 121
Cubo OLAP para el caso de Twitter 122
Cubo OLAP para el caso de Dreamcatchers 124
6.2.2. Creación del cubo de datos con dimensión contextual 126
Cubo OLAP para el caso de Twitter 126
Cubo OLAP para el caso de Dreamcatchers 130
6.3. Análisis multidimensional de Twitter 132
6.3.1. Descripción de los datos a utilizar 132
6.3.2. Ejemplos de consultas 132
6.4. Análisis multidimensional de Dreamcatchers 137
6.4.1. Descripción de los datos a utilizar 137
6.4.2. Ejemplos de consultas 137
6.5. Conclusiones 142

7. Conclusiones y trabajos futuros 143
7.1. Conclusiones 143
7.2. Trabajos futuros 146

A. Sistema para la construcción de la dimensión contextual 149
A.1. Descripción general del sistema 149
A.2. Requisitos funcionales del sistema 149
A.3. Arquitectura del sistema 150
A.4. Modelo de datos 151

B. Herramientas utilizadas 153
B.1. Stanford Part-of-Speech Tagger 153
B.2. Stanford Named Entity Recognition Tagger 153

B.3. Multilingual Central Repository 3.0 153
B.3.1. WordNet Domains 154
B.4. BabelNet 154

Índice de figuras

3.1. Metodología para la detección de contextos 19
3.2. Modelos para la detección de contextos 22
3.3. Etiquetas de los dos primeros niveles de WordNet Domains 24
3.4. Creación de la matriz de pesos Etiqueta-Documento 29
3.5. Ejemplo de gráfica del Coeficiente de Silueta 32
3.6. Coeficiente de Silueta para el conjunto Reuters (inglés) 38
3.7. Coeficiente de Silueta para el conjunto Semcor (inglés) 38
3.8. Coeficiente de Silueta para el conjunto Semcor (español) 39
3.9. Coeficiente de Silueta para el conjunto EFE (español) 39
3.10. Coeficiente de Silueta para el conjunto Twitter (inglés) 40
3.11. Coeficiente de Silueta para el conjunto Dreamcatchers (español) 40
3.12. Gráfica que relaciona el Coeficiente de Silueta y los algoritmos de agrupamiento jerárquico 41
3.13. Gráfica que relaciona el Coeficiente de Silueta y el método utilizado para la detección de contextos 42
3.14. Gráfica que relaciona el Coeficiente de Silueta y el idioma para el conjunto Semcor 43

4.1. Marco de trabajo para el análisis de la influencia de los términos con orientación sentimental en la detección de contextos 48
4.2. Histograma que muestra las polaridades positivas en SentiWordNet 3.0. 50
4.3. Histograma que muestra las polaridades negativas en SentiWordNet 3.0. 50
4.4. Histograma de polaridades en SenticNet 3. 51
4.5. Nube de etiquetas para el Conjunto 4 **a)** No se aplica el filtro de sentimientos **b)** Se aplica el filtro de sentimientos (con el recurso SentiWordNet 3.0 y eliminando todos los términos de sentimientos) 57

4.6. Nube de etiquetas para el Conjunto 8 **a)** No se aplica el filtro de sentimientos **b)** Se aplica el filtro de sentimientos (con el recurso SentiWordNet 3.0 y eliminando todos los términos de sentimientos) 58
4.7. Boxplot entre el Coeficiente de Silueta y la cantidad de grupos 60
4.8. Boxplot entre el Coeficiente de Silueta y los recursos utilizados 63
4.9. Boxplot entre el Coeficiente de Silueta y los algoritmos de agrupamiento 64
4.10. Boxplot entre el Coeficiente de Silueta y las redes sociales utilizadas . . 64
4.11. Valores del Coeficiente de Silueta con respecto al método de selección de los conjuntos de datos (inducido o aleatorio) (a) Valores de silueta con respecto al método de selección cuando los términos con orientación sentimental no son descartados (b) Valores de silueta con respecto al método de selección cuando los términos con orientación sentimental son descartados 65
4.12. Coeficiente de Silueta para el Conjunto 1 utilizando el recurso SentiWordNet 66
4.13. Coeficiente de Silueta para el Conjunto 2 utilizando el recurso SentiWordNet 66
4.14. Coeficiente de Silueta para el Conjunto 3 utilizando el recurso SentiWordNet 67
4.15. Coeficiente de Silueta para el Conjunto 4 utilizando el recurso SentiWordNet 67
4.16. Coeficiente de Silueta para el Conjunto 5 utilizando el recurso SentiWordNet 68
4.17. Coeficiente de Silueta para el Conjunto 6 utilizando el recurso SentiWordNet 68
4.18. Coeficiente de Silueta para el Conjunto 7 utilizando el recurso SentiWordNet 69
4.19. Coeficiente de Silueta para el Conjunto 8 utilizando el recurso SentiWordNet 69
4.20. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 1) . . . 70

4.21. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 2) . . . 70
4.22. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 3) . . . 71
4.23. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 4) . . . 71
4.24. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 5) . . . 72
4.25. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 6) . . . 72
4.26. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 7) . . . 73
4.27. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 8) . . . 73
4.28. Coeficiente de Silueta para el Conjunto 1 utilizando el recurso SenticNet 74
4.29. Coeficiente de Silueta para el Conjunto 2 utilizando el recurso SenticNet 74
4.30. Coeficiente de Silueta para el Conjunto 3 utilizando el recurso SenticNet 75
4.31. Coeficiente de Silueta para el Conjunto 4 utilizando el recurso SenticNet 75
4.32. Coeficiente de Silueta para el Conjunto 5 utilizando el recurso SenticNet 76
4.33. Coeficiente de Silueta para el Conjunto 6 utilizando el recurso SenticNet 76
4.34. Coeficiente de Silueta para el Conjunto 7 utilizando el recurso SenticNet 77
4.35. Coeficiente de Silueta para el Conjunto 8 utilizando el recurso SenticNet 77
4.36. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 1) 78

4.37. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 2) 78
4.38. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 3) 79
4.39. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 4) 79
4.40. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 5) 80
4.41. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 6) 80
4.42. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 7) 81
4.43. Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 8) 81

5.1. Componentes de la dimensión contextual 88
5.2. Jerarquía de contextos para un conjunto de textos de Twitter utilizando el método de enlace completo (Complete Link) 90
5.3. Ejemplo conjunto-AP para el contexto *Computer Science* 92
5.4. Ejemplo de estructura-AP para el contexto *Computer Science* 92
5.5. Ejemplo de esquema del modelo multidimensional implementado por Wonder 3.0 94
5.6. Ejemplo de uso de las operaciones roll-up y drill-down mediante la jerarquía de consulta 95
5.7. Ejemplo de una dimensión contextual para un conjunto de datos de Twitter 97

5.8. Metodología para la creación e integración de la dimensión contextual en el modelo multidimensional 99
5.9. Interfaz principal para la creación de un cubo contextual 103
5.10. Interfaz principal para la creación de un cubo contextual (sin utilizar la metodología propuesta) 104
5.11. Página para seleccionar las dimensiones, jerarquías y niveles del cubo OLAP 105
5.12. Página que permite realizar las operaciones roll-up y drill-down mediante la jerarquía de consulta 106
5.13. Interfaz del gráfico resultante al lanzar la consulta 107
5.14. Valores del coeficiente de silueta para ambas redes sociales con 5000 documentos 110
5.15. Valores del coeficiente de silueta para ambas redes sociales con 10000 documentos 110
5.16. Valores del coeficiente de silueta para ambas redes sociales con 20000 documentos 110
5.17. Valores del coeficiente de silueta para ambas redes sociales con 30000 documentos 111
5.18. Ejemplo de consulta para el Conjunto 4 de Twitter (contexto = *COMPUTER SCIENCE* y nivel = 80 grupos) 112
5.19. Ejemplo de consulta para el Conjunto 4 de Twitter utilizando la jerarquía de consulta (contexto = *COMPUTER SCIENCE* y nivel = 80 grupos) 112
5.20. Ejemplo de consulta para el Conjunto 4 de Dreamcatchers (contexto = *ANATOMY* y nivel = 100 grupos) 113
5.21. Ejemplo de consulta para el Conjunto 4 de Dreamcatchers utilizando la Jerarquía de consulta (contexto = *ANATOMY* y nivel = 100 grupos) . 114

6.1. Arquitectura general del sistema 118
6.2. Arquitectura de Wonder 120
6.3. Definición de las dimensiones clásica 122
6.4. Definición de las dimensiones-AP 123
6.5. Vista que permite ver las propiedades del cubo creado 123
6.6. Definición de las dimensiones clásica 124

6.7. Definición de las dimensiones-AP 125
6.8. Vista que permite ver las propiedades del cubo creado 125
6.9. Selección de los atributos del cubo contextual 127
6.10. Nube de etiquetas con los contextos presentes en los datos seleccionados 128
6.11. Definición de las dimensiones contextual 129
6.12. Vista preliminar del cubo creado 129
6.13. Nube de etiquetas con los contextos presentes en los datos seleccionados 130
6.14. Definición de las dimensiones contextuales 131
6.15. Vista que permite ver las propiedades del cubo creado 131
6.16. Vista que permite construir la consulta para una dimensión clásica . . 133
6.17. Vista que permite construir la consulta para una dimensión-AP 134
6.18. Vista que muestra la jerarquía de consulta 135
6.19. Gráfico resultante de la consulta realizada sobre el cubo con dimensión-AP 135
6.20. Vista que muestra la jerarquía de consulta 136
6.21. Gráfico resultante de la consulta realizada sobre el cubo contextual . . 136
6.22. Vista que permite construir la consulta para una dimensión clásica . . 138
6.23. Vista que permite construir la consulta para una dimensión-AP 139
6.24. Vista que muestra la jerarquía de consulta 140
6.25. Gráfico resultante de la consulta realizada sobre el cubo con dimensión-AP 140
6.26. Vista que muestra la jerarquía de consulta 141
6.27. Gráfico resultante de la consulta realizada sobre el cubo contextual . . 142

A.1. Arquitectura del sistema web para la construcción automática de la dimensión contextual. 151
A.2. Modelo del sistema web para la construcción automática de la dimensión contextual. 152

Índice de tablas

3.1. Ejemplo de desambiguación para la palabra BANCO (*BANK*) 27
3.2. Documentos de ejemplos 30
3.3. Matriz de pesos Etiqueta-Documento 31
3.4. Matriz de distancia 31
3.5. Descripción de los conjuntos de datos 36

4.1. Descripción de los conjuntos de datos 56
4.2. Términos más frecuentes para los Conjuntos 4 y 8 cuando no se aplica el filtro para descartar los términos con orientación sentimental 59
4.3. Cantidad de términos descartados con orientación sentimental por recurso para cada conjunto de datos 59
4.4. Coeficiente de Silueta para los conjuntos del 1-4 61
4.5. Coeficiente de Silueta para los conjuntos del 5-8 62
4.6. Coeficiente de Silueta y cantidad de documentos que permanecen para los conjuntos del 1-4 82
4.7. Coeficiente de Silueta y cantidad de documentos que permanecen para los conjuntos del 5-8 83

5.1. Descripción de los conjuntos de datos utilizados 109
5.2. Tiempo de ejecución de las consultas 114

Capítulo 1

Introducción

La popularidad y uso vertiginoso de las redes sociales en los últimos diez años, ha llevado a que decenas de millones de usuarios generen diariamente gigantescas cantidades de datos textuales [Guille et al., 2013]. Este hecho ha agravado considerablemente la brecha que existe entre el crecimiento de los datos heterogéneos, semiestructurados y no estructurados, y las capacidades de procesamiento y análisis automático de forma masiva de la mayoría de las tecnologías y sistemas actuales que permitirían explotarlos adecuadamente. Se une a esto, el reto de la integración de dicha información textual con datos tradicionales, y de esta forma permitir a los analistas obtener provecho de este nuevo recurso.

Los datos textuales constituyen la inmensa mayoría de los datos que se acumulan en internet y las organizaciones, provenientes de documentos de trabajo, correos electrónicos, sitios web de opinión, encuestas, opiniones en redes sociales, etc. Sin embargo, el gran cúmulo y falta de estructura de los textos, hace que sea prácticamente imposible su procesamiento automático de forma masiva. En este entorno, resulta particularmente útil, detectar automáticamente los principales tópicos que son abordados y que constituyen información relevante, para presentarla a los usuarios en forma legible.

Por este motivo, el desarrollo de herramientas que combinen tecnologías avanzadas para facilitar su análisis, se hace cada vez más necesario. Si estas soluciones se insertan dentro de sistemas de ayuda a la toma de decisiones, las prestaciones de estos últimos aumentan considerablemente.

Como se mencionó anteriormente, la falta de estructura de los textos dificulta su análisis automático, y si tenemos en cuenta que dichos datos provienen de redes sociales donde cualquier usuario puede realizar comentarios y expresar su opinión

sobre cualquier tema, es de esperar que los textos estén sujetos a errores de redacción, ortográficos, etc. Por estos motivos es necesario realizar un mejor preprocesamiento sintáctico y semántico de los textos para facilitar su procesamiento y análisis de forma automática.

El procesamiento de datos masivos implica resumir y agrupar, y para ello las tecnologías Data Warehousing (DW) [Galhardas et al., 2001] y Online Analytical Processing (OLAP) [Codd et al., 1993] se presentan como las más adecuadas. Estas tecnologías basan su éxito en las ventajas de la integración, el almacenamiento y operaciones del modelo multidimensional. De esta forma, permiten el desarrollo de agregaciones a través de dimensiones convencionales y no convencionales sobre datos heterogéneos. Para el caso concreto de los datos textuales, primeramente deben sufrir algún tipo de transformación para llevarlos a una forma más estructurada que facilite su análisis.

Para poder aplicar de forma satisfactoria las tecnologías DW y OLAP en el análisis de información textual provista por las redes sociales, resulta útil detectar previamente los principales contextos presentes en los textos, y para cada contexto, los tópicos más relevantes. Esto permitiría a los analistas segmentar los datos textuales por contextos, para luego tratarlos aprovechando las características y capacidades proporcionadas por el análisis multidimensional.

1.1. Planteamiento del problema

Problema a resolver

Los sistemas DW y OLAP como ya se ha expuesto, abren una puerta al mundo del análisis multidimensional de los datos, el cual resulta muy útil en tareas relacionadas con la toma de decisiones. Este análisis aporta grandes ventajas cuando los datos a analizar, constituyen datos clásicos o tradiciones, no siendo así para el caso de los datos textuales, ya que no cuentan con valores discretos, sino que se caracterizan por su falta de estructura y homogeneidad.

Supongamos que se cuenta con un sistema para analizar datos de una red social. En este caso, son manejados con facilidad todos los datos tradicionales que, por lo general, son datos del usuario como sexo, edad, ciudad de nacimiento, etc; pero no sucede así con los datos textuales. Los campos tales, como tweets, comentarios, post,

etc. (dependiendo de la red social) son cualitativamente valiosos y al estar formados por textos se hace difícil su procesamiento automático.

No son pocos los trabajos orientados al uso de las tecnologías DW y OLAP para el estudio de los datos textuales, especialmente por la invaluable información que aportan sobre un determinado producto, servicio, etc. La gran mayoría de estos trabajos, realizan tareas tales como Recuperación de Información, Análisis de Sentimientos, Sistemas de Recomendación, etc., sin tener en cuenta previamente el contexto al cual pertenecen dichos datos textuales.

En [Martin-Bautista et al., 2013], los autores presentan una propuesta en la cual logran procesar automáticamente el conocimiento asociado a textos libres y utilizar dicho conocimiento en un sistema multidimensional. Sin embargo, los autores no tienen en cuenta segmentar previamente los datos textuales por los principales contextos tratados, lo cual facilitaría el posterior análisis de los textos vinculados con datos clásicos. Además se debe tener en cuenta que al estar agrupados los textos por categorías, los tiempos de respuesta de las consultas de los usuarios disminuiría considerablemente, siempre y cuando el análisis se haga por un contexto determinado. El presente trabajo está dedicado a estudiar esta dificultad, asociada a los sistemas DW y OLAP, con el fin de proponer una solución.

Para el caso concreto de redes sociales, podemos encontrar trabajos que vinculan las tecnologías DW y OLAP con redes sociales. Tal es el caso en [Zhao et al., 2011a] donde se introduce un nuevo modelo de Data Warehousing con soporte a consultas OLAP en redes multidimensionales. Mientras en [Park et al., 2013] se propone un nuevo modelo para el análisis de grandes volúmenes de tráfico almacenados durante un largo período de tiempo. Estos trabajos están indudablemente orientados al estudio de la estructura de determinado sistema, los cuales no resultan de mucho interés para el presente trabajo. En particular, nos interesan más aquellos trabajos vinculados con el análisis y estudio de los datos textuales almacenados en redes sociales.

Tal es el caso en [Bringay et al., 2011], donde Bringay et al. definen un modelo de data warehouse para analizar grandes volúmenes de tweets a partir de medidas relevantes propuestas en el contexto de descubrimiento de conocimiento. Se proponen dos modelos para manipular los cubos: utilizando una jerarquía predefinida (la jerarquía médica Medical Subject Headings) y mediante el uso de la jerarquía presente en una dimensión tradicional (Ej. dimensión fecha, lugar, etc.). Nuestra propuesta,

en cambio, permite construir una jerarquía de contextos independiente de la diversidad de dominios presentes en los texto, por lo tanto, no es necesario contar con una jerarquía de dominio específico.

El tratamiento y análisis de atributos textuales procedentes de redes sociales en los entornos DW y OLAP conjuntamente con cualquier otro tipo de dato, constituye el objetivo principal del presente trabajo. Para darle solución a este problema de investigación, partimos de la formulación de la siguiente hipótesis.

Hipótesis

Se puede mejorar el análisis multidimensional de datos textuales de redes sociales mediante la construcción de una dimensión contextual con técnicas de minería de datos. Dicha dimensión encierra la semántica asociada de los datos textuales y permite segmentar los textos por contextos. Al integrar la dimensión contextual en un modelo multidimensional, facilita el análisis de datos textuales conjuntamente y de la misma manera que las dimensiones no textuales.

Una vez planteada nuestra hipótesis, en la próxima sección se detalla el marco de trabajo en el que se desarrolla la solución al problema planteado.

1.2. Marco de trabajo

Este trabajo se enmarca dentro de los temas detección de tópicos, DW y OLAP. Concretamente en su utilización en datos de redes sociales. La detección de tópicos consiste en detectar los principales tópicos a partir de un conjunto de textos, para luego vincular cada documento con una categoría o etiqueta representativa del tópico al que pertenece dicho texto.

En la literatura podemos encontrar una gran variedad de técnicas empleadas en la detección de tópicos, entre las que se encuentran algoritmos supervisados, semi-supervisados y no supervisados, técnicas basadas en modelos probabilísticos, así como las basadas en la incorporación de información adicional por mediante el uso de ontologías de dominio o recursos léxicos similares. Estas técnicas, en la mayoría de los casos, no están ligadas a procesos data warehousing, a pesar de que comparten un objetivo general común: la extracción de conocimiento. Aunque durante el presente capítulo hemos utilizado indistintamente *"Detección de Tópicos"* y *"Detección de Contextos"* para referirnos a esta temática, en el Capítulo 3 se pueden encontrar

las definiciones de los términos *"Tópicos"* y *"Contextos"* que permiten establecer las bases de nuestra propuesta.

Por otro lado, DW y OLAP constituyen tecnologías que mantienen una estrecha relación. Ambas surgen con el fin de lograr un mejor aprovechamiento de grandes volúmenes de datos. Existen productos DW que proveen servicios OLAP y utilizando herramientas OLAP, los usuarios pueden acceder al DW mejorando así la comprensión del negocio para la toma de decisiones.

En esta tesis se han logrado combinar las tres temáticas de la siguiente manera:

Primero mediante el uso de técnicas de minería de datos, específicamente algoritmos de agrupamiento jerárquico y con el uso de recursos léxicos, se construye una dimensión contextual. Esta dimensión, presenta una estructura jerárquica donde cada nodo de la jerarquía representa un contexto al que están relacionados un conjuntos de documentos. Además para cada contexto y nivel de la dimensión, se cuenta con una jerarquía de consulta mediante la que se pueden realizar consultas por los principales tópicos presentes en este contexto. Todo el proceso anterior se realiza de forma automática.

Contando con esta información, podemos crear un modelo multidimensional que brinde soporte a la dimensión contextual obtenida. Este modelo, además de soportar un nuevo tipo de dimensión, implementa las operaciones OLAP clásicas para este tipo de dimensiones, de tal forma que se puedan realizar análisis detallados mediante dicha dimensión relacionada con las dimensiones clásicas.

Por último se implementan un conjunto de funcionalidades, las cuales son incorporadas en la herramienta Wonder OLAP Server 3.0, en adelante Wonder 3.0, mediante la cual se pondrá en práctica nuestra propuesta. Este sistema nos va a permitir analizar los datos textuales de las redes sociales Twitter y Dreamcatchers junto con datos estructurados, demostrando la utilidad de la dimensión contextual y el buen funcionamiento de Wonder.

1.3. Objetivos

Como se comentó anteriormente, el objetivo fundamental de este trabajo consiste en analizar los datos textuales de redes sociales mediante un modelo multidimensional. Para facilitar este análisis, primeramente los textos son segmentados o agrupados en contextos. Luego, una vez ordenados por categorías, los decisores pueden

realizar análisis a un gran nivel de detalle con el objetivo de poder trazar estrategias válidas en las empresas.

Todo lo anterior se concreta mediante una serie de objetivos específicos o tareas a realizar que se exponen a continuación:

Realizar un estudio bibliográfico para conocer el estado del arte sobre detección de tópicos, DW y OLAP. Con ello pretendemos situarnos en el estado actual del problema a resolver, así como estudiar y seleccionar las técnicas a utilizar para la solución planteada.

Detectar automáticamente los principales tópicos abordados en datos textuales con el fin de segmentar los textos por tópicos y de esta forma facilitar su posterior análisis.

Analizar la influencia que tienen los términos de sentimientos en la detección automática de tópicos en redes sociales, para determinar el punto de equilibrio entre la cantidad de términos de sentimientos descartados y la cantidad de documentos que permanecen luego de preprocesar sintáctica y semánticamente los textos.

Construir e integrar una dimensión contextual en un modelo multidimensional, facilitando el análisis de los textos de redes sociales mediante esta dimensión en conjunto con dimensiones clásicas.

Realizar la explotación de la dimensión contextual mediante el uso de un servidor OLAP con datos reales de dos redes sociales Twitter y Dreamcatchers, materializando así, los resultados teóricos previamente obtenidos. Por último, para corroborar la calidad de los resultados y las ventajas que introduce el uso de la dimensión contextual en la realización de consultas sobre atributos textuales de redes sociales un en entorno OLAP, se tendrán en cuenta consultas propuestas por expertos, que demuestran la utilidad práctica de la dimensión contextual propuesta y que confirman la veracidad de la información obtenida.

Un resumen del cumplimiento de estos objetivos específicos se puede ver en el Capítulo 7 de esta memoria.

1.4. Aportaciones

Desde el punto de vista teórico este trabajo realiza las siguientes aportaciones:

1. Se ha desarrollado una metodología para la detección automática de tópicos a partir de datos textuales. Dicha metodología permite determinar los tópicos

a los que pertenecen cada documento, facilitando así su posterior análisis en conjunto con datos estructurados.

2. Se ha realizado un estudio sobre la influencia que tienen los términos de sentimientos en la detección automática de tópicos en textos de redes sociales. Dicho estudio permite establecer un consenso entre la cantidad de términos de sentimientos descartados y el número de documentos que permanecen luego de aplicar el filtro para los términos de sentimientos.

3. Se ha obtenido un nuevo modelo multidimensional el cual soporta entre sus dimensiones, la dimensión contextual. Dicha dimensión permite analizar los datos textuales de redes sociales junto con dimensiones tradicionales por los diferentes contextos, y para cada contexto por los principales tópicos abordados.

Desde el punto de vista práctico y tecnológico se han desarrollado las siguientes aportaciones:

1. Se han implementado las funcionalidades necesarias que permiten:
   - detectar los principales tópicos abordados en datos textuales,
   - aplicar el filtro para detectar y descartar los términos de sentimientos, y
   - finalmente crear la dimensión contextual.

2. Se han implementado las requisitos funcionales necesarios que permiten que la herramienta OLAP utilizada Wonder OLAP Server 3.0 brinde soporte para las dimensiones contextuales.

1.5. Contenidos de la memoria

La presente memoria está compuesta por siete capítulos, los cuales cumplen los objetivos antes expuestos. Luego de esta introducción, se valoran los principales trabajos relacionados con el tema que nos ocupa presentes en la literatura. En el Capítulo 2 **Antecedentes**, se discutirá sobre el entorno que caracteriza el problema abordado y las diferentes soluciones que aparecen en la literatura.

En el Capítulo 3 **Detección automática de tópicos**, nos centramos en el proceso de detección de tópicos en datos textuales. Se introducen las definiciones necesarias

que facilitan el buen entendimiento de la metodología para la detección automática de tópicos. También se explican detalladamente las etapas y procesos presentes en dicha metodología. Además se exponen los resultados obtenidos al poner en práctica la propuesta.

En el Capítulo 4 **Influencia de los términos con orientación sentimental en la detección de contextos: Análisis en redes sociales**, se presenta un nuevo enfoque con el fin de mejorar la metodología propuesta en el capítulo anterior para la detección automática de contextos en textos de redes sociales. Para ello se incluye un filtro que permite detectar y descartar los términos de sentimientos. Finalmente se realiza un análisis de los resultados obtenidos al poner en práctica el nuevo enfoque.

En el Capítulo 5 **Construcción de una dimensión contextual para el análisis multidimensional de textos de redes sociales**, se presentan los definiciones necesarias que constituyen la base teórica para la construcción de la dimensión contextual a partir de datos textuales de redes sociales y su posterior integración en un modelo multidimensional.

En el Capítulo 6 **Explotación de la dimensión contextual mediante el servidor OLAP "Wonder OLAP Server 3.0"**, se lleva a cabo de forma práctica la inclusión de una dimensión contextual en un modelo multidimensional. Para ello se utiliza el servidor OLAP Wonder 3.0. Se realizan consultas para los conjuntos de datos empleados con vista demostrar la viavilidad del sistema.

Por último, en el Capítulo 7 **Conclusiones y trabajos futuros**, se encuentran las conclusiones obtenidas con la realización de este trabajo y las líneas de investigación a seguir en el futuro.

Capítulo 2

Antecedentes

En el presente capítulo se realiza un estudio del arte de las principales tendencias relacionadas con la detección de tópicos y el análisis multidimensional en redes sociales. Con ello pretendemos conocer el estado actual del problema que se desea resolver y estudiar las técnicas y soluciones propuestas en los estudios más recientes.

Como hemos mencionado en el capítulo anterior, nuestra propuesta combina una gran variedad de técnicas y tecnologías. Trabajaremos con técnicas de minería de datos, utilizaremos recursos léxicos (diccionarios electrónicos, ontologías de dominio y recursos relacionados con el análisis de sentimientos), y haremos uso de tecnologías relacionadas con el análisis multidimensional de datos textuales. Por lo que resulta conveniente analizar los trabajos más relacionados con el presente trabajo.

Haremos especial énfasis en aquellos trabajos relacionados con la detección de tópicos mediante algoritmos no supervisados y el análisis multidimensional de datos textuales en redes sociales. Con ello daremos una visión general de las distintas líneas de investigación existentes. Finalmente formularemos nuestra propuesta para resolver el problema propuesto.

2.1. Detección de tópicos

La detección de tópicos a partir de grandes volúmenes de textos, ha sido un tema ampliamente analizado en la literatura desde varios puntos vistas. Entre ellos resalta el uso de métodos como algoritmos de clasificación, Latent Dirichlet Allocation (LDA), algoritmos de agrupamiento, entre otros. Para el caso de los algoritmos de clasificación es necesario contar con un conjunto de datos de entrenamiento que permita entrenar el clasificador, mientras tanto LDA y los algoritmos de agrupamiento no resulta necesario contar con un corpus previamente clasificado.

Los algoritmos de agrupamiento jerárquicos han adquirido una vital importancia en tareas relacionadas con la categorización de registros de diferentes tipos, incluyendo datos textuales [RaghavaRao et al., 2012, Deshmukh et al., 2013]. En [Voorhees, 1986, Willett, 1988] podemos observar un estudio sobre los algoritmos de agrupamiento jerárquico aglomerativo tradicionales, especialmente con datos textuales.

Son muchos los trabajos que podemos encontrar relacionados con la detección de tópicos mediante el uso de algoritmos de agrupamiento jerárquico supervisados y semi-supervisados, no así para los no supervisados. Tales son los casos propuestos en [Chung-Hong, 2012, Skarmeta et al., 2000a, Zheng and Li, 2011], donde los autores proponen enfoques basados en el uso de información experta, y de esta forma mejorar los resultados en la detección de los principales tópicos. Skarmeta el al., presenta un estudio sobre el uso de un semi-supervised Agglomerative Hierarchical Clustering (ssAHC) algorithm, el cual asigna los textos a categorías predefinidas[Skarmeta et al., 2000b].

En redes sociales, como el que nos ocupa, la detección de tópicos ha sido extensamente utilizada para el análisis de datos textuales. Muchas han sido los soluciones que han aparecido para el análisis textual en redes sociales, tales como el análisis de sentimientos [Lin and He, 2009], el filtrado de contenidos [Duan and Zeng, 2013, Martinez-Romo and Araujo, 2013], la modelación de los intereses del usuario [Pennacchiotti and Gurumurthy, 2011], así como el seguimiento de eventos de interés [Chung-Hong, 2012, Wu et al., 2011]. En [Zhao et al., 2011b] se realiza una comparación entre el contenido de los textos de Twitter con un medio de comunicación tradicional, el New York Times. Para ello se utiliza el modelado de tópicos sin supervisión utilizando el modelo Twitter-LDA, para descubrir dichos tópicos en mensajes cortos.

Por nuestra parte, en este trabajo se considera la detección automática de tópicos en datos textuales de redes sociales sin ningún conocimiento previo, por lo que no se necesita de un experto. Por esta razón, es de nuestro interés el grupo de trabajos previos relacionados con el uso de métodos de agrupamiento jerárquicos en la detección de tópicos y los que realizan esta tarea sobre datos textuales en el contexto de medios sociales.

Gao et al. presenta en [Gao et al., 2013] un nuevo algoritmo de detección de tópicos en noticias publicadas en Internet sobre grandes desastres, basado en Group

Average Hierarchical Clustering (GAHC). La idea central de dicho algoritmo consiste en dividir los grandes datos en grupos más pequeños, y luego agrupar jerárquicamente dichos grupos, para generar los tópicos finales. Una herramienta práctica para ayudar a periodistas y lectores de noticias a buscar temas de interés periodístico de flujos de mensajes sin sentirse abrumados es presentada en [Martin et al., 2013]. En este caso se presenta una variante dependiente del tiempo del enfoque tf-idf clásico, y se agrupan frases en ráfagas que aparecen a menudo en los mismos mensajes, con el fin de identificar los tópicos emergentes en una misma ventana de tiempo. Los experimentos se realizaron con datos de Twitter relacionados con el deporte y la política. En [Xiaohui et al., 2013] se construye un grafo conceptual con los conceptos como nodos y se conectan por las aristas con los nodos que comparten los mismos términos temáticos. Al realizar la agrupación jerárquica en este grafo conceptual, las curvas de comportamiento de conceptos altamente correlacionados se agrupan como tópicos.

Se debe mencionar que estos trabajos están orientados mayormente a la detección de tópicos o eventos en redes sociales, donde los datos pertenecen a un dominio específico y durante un intervalo de tiempo dado. Además como es el caso en [Xiaohui et al., 2013], se conoce el número de tópicos para cada conjunto de datos utilizado en la experimentación. Por el contrario, el enfoque propuesto en el presente trabajo permite determinar los contextos sobre datos cuyo dominio es desconocido (genérico), independiente del idioma ya que la base de conocimiento utilizada brinda soporte para varios idiomas y es aplicable a textos de cualquier red social.

2.2. Análisis multidimensional en redes sociales

Como se mencionó anteriormente, la idea principal de nuestra investigación, consiste en mejorar el análisis multidimensional de datos textuales de redes sociales conjuntamente con datos clásicos. Para dar cumplimiento a lo planteado anteriormente, se propone la construcción e integración de forma automática de una Dimensión Contextual en un modelo multidimensional, y de esta forma facilitar a usuarios, organizaciones e investigadores el análisis de los datos en redes sociales por los principales contextos y tópicos detectados. Por este motivo se ha realizado un estudio de los principales trabajos relacionados con las temáticas vinculadas a la detección de

tópicos a partir de los textos en redes sociales, y al análisis multidimensional sobre datos heterogéneos, que integren tanto datos textuales como convencionales en redes sociales.

Data Warehouses y OLAP constituyen tecnologías de vital importancia en sistemas relacionados con la toma de decisiones y han demostrado su competitividad y ventajas en varios tipos de aplicaciones. Entre las principales ventajas de los sistemas OLAP, se encuentra la facilidad para realizar resúmenes y agrupaciones de grandes conjuntos datos de forma multidimensional, lo cual lo hace una herramienta de gran utilidad en aplicaciones de diversos dominios.

En la actualidad, con el auge y crecimiento de las redes sociales, organizaciones, grupos de investigación, etc., han dedicado tiempo y recursos al estudio de los datos almacenados en estas redes. Una de las alternativas que han surgido como respuesta al análisis de los grandes volúmenes de datos de las redes sociales, es la integración de dichos datos en un modelo multidimensional. Como resultado de esta integración se han establecido dos líneas de investigación: *Social network analysis* y *Social media analysis*. El presente trabajo está enmarcado en la segunda línea.

Social Network Analysis (SNA) es una estrategia para el estudio de las estructuras sociales mediante el uso de teorías de redes y gráficos [Otte and Rousseau, 2002]. Mientras Social Media Analysis (SMA) es el proceso de obtener información a partir de las conversaciones que se encuentran en formato digital, y dicha información puede ser utilizada en en procesos relacionados con la toma de decisiones, marketing, atención al cliente, ventas, etc.

En la literatura se puede encontrar un sinnúmero de trabajos relacionados con el SNA, tal es el caso en [Zhao et al., 2011a] donde se introduce un nuevo modelo de Data Warehousing con soporte a consultas OLAP on large multidimensional networks. Mientras Park et al. en [Park et al., 2013] propone un nuevo modelo para el análisis de grandes volúmenes de tráfico en almacenados durante un período de tiempo largo. Estos trabajos están indudablemente orientados al estudio de la estructura social de determinado sistema. En particular, nos interesa más aquellos trabajos vinculados con el análisis y estudio de los datos almacenados en redes sociales.

Tal es el caso en [Bringay et al., 2011], donde Bringay et al. definen un modelo de data warehouses para analizar grandes volúmenes de tweets a partir de medidas relevantes en el contexto de descubrimiento de conocimiento. También se propone

cómo extraer el contexto de un concepto en una jerarquía, para ello se utiliza la jerarquía médica MeSH (Medical Subject Headings). En nuestro trabajo al igual que en [Bringay et al., 2011], para determinar el contexto de un término se utiliza un recurso léxico con estructura jerárquica, pero se debe resaltar que nuestro enfoque es más genérico, al tratarse de una jerarquía donde están presentes las principales áreas del conocimiento.

Otro trabajo que tiene en cuenta los textos presentes en los tweets, es el de Liu et al., [Liu et al., 2013]. En éste, los autores presentan un *"text cube"* para estudiar los diferentes tipos de human, social and cultural behavior (HSCB) presentes en el Twitter stream. *"Text cube"* permite organizar los datos(Twitter text) en múltiples dimensiones y jerarquías para lograr consultar y visualizar los datos de forma eficiente. El principal enfoque es realizar análisis de sentimientos y la visualización de los datos. Como se ha mencionado, el objetivo principal de nuestro trabajo es permitir el análisis multidimensional de los datos textuales en redes sociales por un contexto determinado. Por tal motivo nuestro primer paso consiste en detectar los principales contextos, para luego analizar los datos a partir de las dimensiones extraídas de los textos pertenecientes a un contexto, mejorando así la calidad de la información.

En [Zhang et al., 2009], se propone un modelo de datos *"Topic Cube"*, el cual combina OLAP con un modelo probabilístico de tópicos, permitiendo realizar OLAP en las dimensiones textuales de una base de datos de textos multidimensional. *"Topic Cube"* extiende del cubo de datos OLAP tradicional, para brindar soporte a una Jerarquía de Tópicos y almacena medidas probabilísticas de los textos, aprendidas mediante un modelo probabilístico de tópicos. La principal diferencia con nuestro enfoque se basa en en que la Jerarquía de Tópicos debe ser especificada por un analista, mientras que la nueva Dimensión Contextual aquí propuesta se construye automáticamente.

En la gran mayoría de trabajos orientados a integrar los textos de redes sociales en un modelo multidimensional, se basan fundamentalmente en la extracción de nuevas dimensiones(contextos/tópicos, entidades, sentimientos, etc.) o medidas de los textos. Tal es el caso en [Pérez et al., 2008], donde se propone un nuevo framework para integrar opiniones de redes sociales con un data warehouse corporativo.

En [Moya et al., 2011] se propone integrar el resultado de realizar análisis de sentimientos en opiniones de redes sociales con un data warehouse, y así poder realizar OLAP sobre dichos datos. Rehman et al., intentan extender OLAP, para permitir el análisis multidimensional de los datos de redes sociales, integrando métodos de minería de textos y opiniones con un sistema data warehouse y utilizando varias técnicas relacionadas con el descubrimiento de conocimientos de datos semiestructurados y no estructurados de las redes sociales [Rehman et al., 2013].

Según nuestro conocimiento, es completamente novedosa la idea de primeramente detectar los principales contextos presentes en los textos e integrarlos en un modelo multidimensional, y una vez hecho esto, extraer nuevas dimensiones y medidas de los textos por contextos, para de esta forma brindar información de mejor calidad a los analistas.

2.3. Propuesta

Como se puede observar, la inmensa mayoría de los trabajos mencionados están orientados principalmente a la detección de tópicos en datos textuales donde cada texto ha sido previamente asignado a una temática específica. En la mayoría de los casos, los datos pertenecen a un dominio específico, durante un intervalo de tiempo dado y, finalmente, cada texto analizado es asignado a una temática.

A diferencia de los estudios mencionados anteriormente, nuestra investigación considera la detección automática de tópicos en datos textuales de redes sociales sin ningún conocimiento previo. El enfoque aquí propuesto, permite detectar los principales tópicos basado en el uso de una ontología e independientemente del idioma en que estén redactados los textos, ya que la base de conocimiento utilizada brinda soporte para varios idiomas. Se debe mencionar que nos hemos centrado en el enriquecimiento semánticamente de los textos mediante información adicional con la ayuda de un recurso léxico (Multilingual Central Repository 3.0), para así mejorar los resultados del proceso de detección de tópicos. En los experimentos no se han utilizado conjuntos de datos previamente etiquetados, ya que no es posible establecer una comparación entre las etiquetas de la jerarquía utilizada (WordNet Domains) y las etiquetas de los textos analizados.

Capítulo 3

Detección automática de tópicos

En el presente capítulo proponemos una metodología para detectar automáticamente los principales tópicos presentes en datos textuales, basado en el uso de una ontología de dominio, utilizando técnicas de minería de datos y una base de conocimiento multilingüe. Se debe tener en cuenta que dichos tópicos son considerados generales, ya que nuestro principal objetivo consiste en tener los textos previamente organizados, para luego ser utilizados como una dimensión en sistemas multidimensionales, y de esta forma facilitar el análisis de los textos integrados con datos estructurados.

3.1. Propuesta metodológica

3.1.1. Ideas básicas

De forma general, el proceso de detección de tópicos permite asignar a un texto específico una o varias etiquetas que describan los principales temas presentes en los textos. La detección de tópicos ha sido abordada desde varios enfoques. Dichos enfoques podríamos clasificarlos como supervisados y no supervisados. Los primeros asignan etiquetas a los textos a partir del uso de algoritmos de clasificación, donde se cuenta con un conjunto de textos de entrenamientos los cuales han sido etiquetados por expertos previamente. Por otra parte, los no supervisados tratan de asignar etiquetas a los textos sin contar con información previa, tal es el caso de los Algoritmos de Agrupamiento Jerárquico (este enfoque será el utilizado en nuestro estudio). Se debe mencionar que en la actualidad con el fin de mejorar los resultados en el proceso de detección automática de tópicos, los enfoques antes mencionados, están siendo combinados con otros acercamientos. Entre los que se encuentran:

1. los que tienen en cuenta características sintácticas y semánticas de los textos, como pueden ser la categoría gramatical de los términos, así como las diferentes entidades presentes en los documentos (persona, tiempo, localización, organización, etc.), y

2. los que utilizan recursos léxicos tales como base de conocimientos y ontologías con el objetivo de enriquecer semánticamente los textos y de esta forma incrementar la precisión en la detección de tópicos.

3.1.2. Definiciones

Antes de entrar en detalles relacionados con la metodología para la detección de tópicos, realizaremos un par de definiciones las cuales facilitarán la comprensión de nuestra propuesta. En la literatura podemos encontrar muchos trabajos relacionados con la *Detección de Tópicos*, sobre todo vinculados a la temática conocida como *Topic Detection and Tracking (TDT)*. Aunque en Ciencias de la Computación la frase *"Detección de Tópicos"* ha sido acuñada, el término "tópico", es empleado indistintamente por los autores para referirse a *eventos*, *conceptos*, *contextos*, etc. [Allan et al., 1998a, Allan et al., 1998b, Young-Woo and Sycara, 2004, Chung-Hong, 2012, Xiaohui et al., 2013, Gao et al., 2013, Martin et al., 2013].

Sin embargo, ningún autor propone una definición formal de dicho término. Para un mejor entendimiento del presente capítulo, en la presente sección se incluyen las definiciones de los términos *tópico* y *contexto* que mejor se ajustan para nuestros propósitos.

Consideremos $W = \{w_1, w_2, ..., w_n\}$ un vocabulario cualquiera (en nuestro caso sería el vocabulario asociado a un conjunto de textos), definimos:

Definición 1 ***Tópico***

Un tópico es un subconjunto de términos del vocabulario $T \subset W$, los cuales tienen una semántica asociada de acuerdo con algún recurso léxico (WordNet, Wikipedia, etc.).

Ejemplo 1 *internet = {LAN, WIFI, WEBSITE}, programming = {JAVA, PHP, WEBSITE}, laptop = {ACER, TOSHIBA, WIFI}*

Definición 2 ***Contexto***

Un contexto es un conjunto de tópicos relacionados semánticamente $C = \{T_1, T_2, ..., T_n\}$.

Obviamente todo contexto puede verse como un tópico más amplio, es decir, como el elemento más general en la jerarquía de tópicos.

Ejemplo 2 *Computer Science = {internet, programming, laptop}*

En otras palabras, un contexto puede ser considerado como una temática general, compuesto por un conjunto de tópicos estrechamente relacionados (ej. los tópicos *internet*, *programming*, *laptop*, etc. formarían parte del contexto *Computer Science*).

3.1.3. Descripción de la metodología

Como se mencionó en la Sección 3.1.1, el uso de Algoritmos de Agrupamiento Jerárquico, nos permite obtener grupos que representan los diferentes contextos tratados en los textos sin información previa. Además nos ofrece la posibilidad de obtener una jerarquía de contextos, la cual puede ser muy útil para el análisis de los datos en diferentes niveles de abstracción. De esta forma será posible analizar datos textuales integrados con datos convencionales, particularmente en procesos de minería de datos y data warehousing, permitiendo el desarrollo de herramientas robustas para el análisis de datos heterogéneos y dando la posibilidad a los encargados de tomar decisiones, de realizar análisis con un gran nivel de detalle de los datos textuales mezclados con datos clásicos.

Es posible mejorar los resultados de los algoritmos de agrupamiento si a todo lo mencionado anteriormente, le agregamos el uso de recursos léxicos con el fin de enriquecer sintáctica y semánticamente los documentos analizados.

Nuestra propuesta se basa en el uso de recursos léxicos tales como Multilingual Central Repository 3.0 (MCR 3.0) [Agirre et al., 2012], BabelNet [Navigli and Ponzetto, 2012] y de las herramientas Stanford Part-of-Speech (POS) Tagger [Toutanova et al., 2003] y Stanford Named Entity Recognition (NER) [Finkel et al., 2005], combinados con Algoritmos de Agrupamiento Jerárquico. Por último, los grupos de contextos obtenidos son etiquetado. En dicha etapa se seleccionan las etiquetas descriptivas de cada grupo. Las etiquetas de dichos grupos pertenecen al conjunto de etiquetas definidas en el recurso semántico WordNet Domains [Magnini and Cavaglia, 2000] incluido en la base de conocimiento MCR 3.0.

Para facilitar el procesamiento y análisis de los textos, y debido principalmente a su falta de estructura, es necesario realizar un buen preprocesamiento sintáctico y semántico. Cada término presente en el texto analizado es procesado sintáctica

y semánticamente, desambiguado y asociado a unas de las etiquetas de WordNet Domains. Estas etiquetas sustituyen a los términos originales de los textos y de esta forma se reduce la dimensionalidad del problema, ya que sin importar el número de términos diferentes, el máximo número de términos en los textos luego de sustituir los términos originales, será el número de etiquetas presente en WordNet Domains.

La principal ventaja de nuestra propuesta es que permite analizar los datos textuales de forma automática sin tener información previa de los contextos abordados en los textos. Además, los textos analizados pueden estar en diferentes idiomas, siempre y cuando la base de conocimiento utilizada MCR 3.0 brinde soporte para dicho idioma.

Para un conjunto de datos textuales, nuestra propuesta será capaz de detectar los contextos tratados en los textos como se demuestra en el estudio experimental. La Figura 3.1 muestra las cuatro etapas o fases de la metodología para la detección de contextos. A continuación se comentan brevemente cada uno de los pasos que posteriormente se explicarán en detalle.

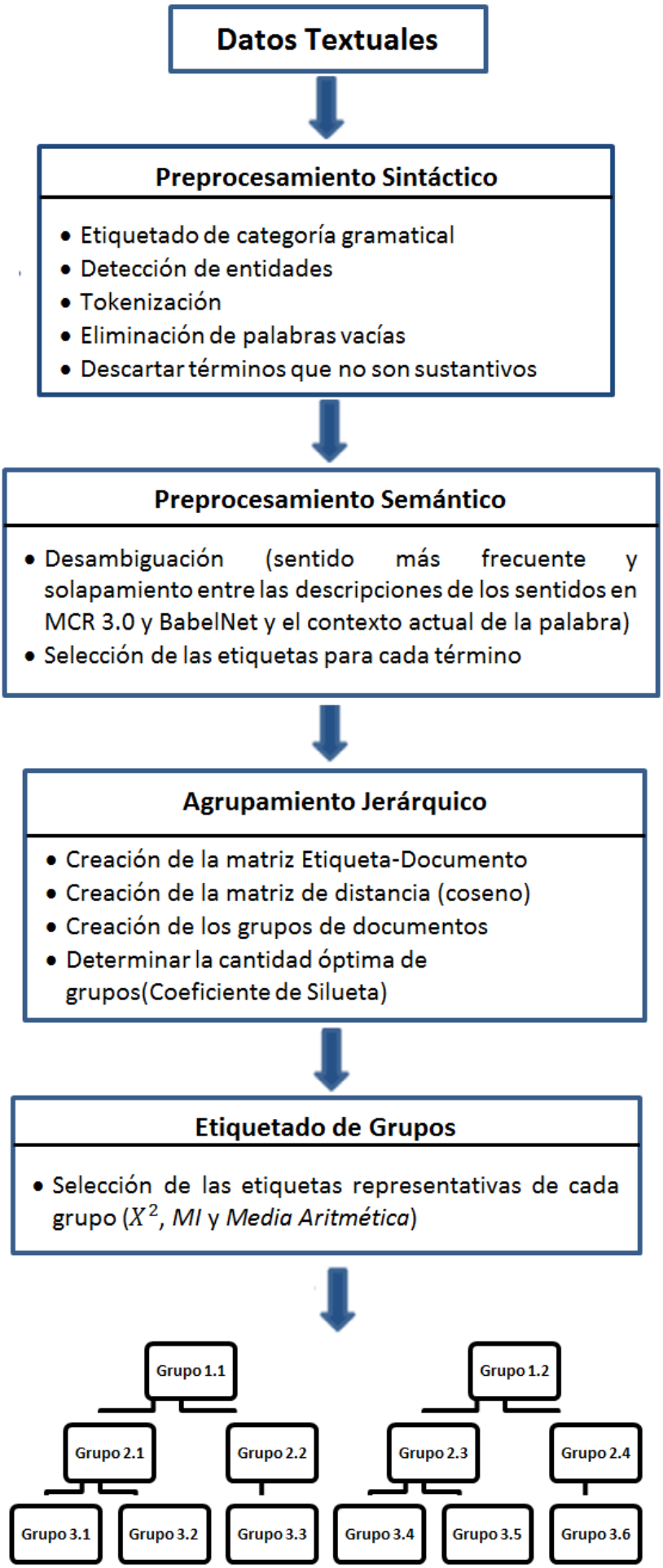


FIGURA 3.1: Metodología para la detección de contextos

1. **Preprocesamiento sintáctico:** La Sección 3.1.4 describe cada uno de los pasos de esta etapa (etiquetado de categoría gramatical, detección de entidades, tokenización, eliminación de palabras vacías, y descartar las palabras que no son identificadas como sustantivo). Una vez aplicados los filtros anteriores, los textos están listos para ser preprocesados semánticamente, ya que sólo permanecen aquellos términos que aportan información relevante para la detección de contextos.

2. **Preprocesamiento semántico:** La Sección 3.1.5 describe el método de desambiguación empleado para determinar el sentido correcto para cada término (se ha utilizado una heurística que considera el solapamiento entre la descripción de los sentidos en MCR 3.0 y BabelNet y el contexto actual donde ha sido empleado el término). Luego de la desambiguación, se realiza la selección de las etiquetas para cada término, las cuales van a sustituir a los términos originales de los textos.

3. **Agrupamiento jerárquico:** La Sección 3.1.6 describe el proceso de creación de la matriz de pesos Etiqueta-Documento y a partir de dicha matriz se crea una matriz de distancia, para luego aplicar los algoritmos de agrupamiento. Finalmente utilizando como medida el Coeficiente de Silueta se determina el número de grupos para el cual los algoritmos de agrupamiento utilizados brindan mejores resultados.

4. **Etiquetado de grupos:** La Sección 3.1.7 describe el proceso relacionado con el etiquetado de los grupos. Se exponen los métodos principales para la selección de las etiquetas representativas de cada grupo, así como los utilizados en el presente trabajo.

Para un mejor entendimiento, en la Figura 3.2 se muestran los pasos necesarios en los que son transformados los datos textuales a lo largo del proceso para la detección de contextos. En primer lugar, los textos se representan mediante el "Vector Space Model" (VSM) [Salton and McGill, 1983], donde D_t representan las t-*th* intervenciones o textos y T_n representa la presencia de los términos o palabras n-*th* en cada texto, como se muestra en la Figura 3.2 (M_1). Seguidamente son preprocesados sintácticamente y representados mediante VSM, donde el número de intervenciones $r \leq t$ y los términos $m \leq n$, ya que tanto la cantidad de documentos como el número de términos pueden variar con respecto a la etapa anterior Figura 3.2 (M_2).

Durante este proceso pueden ser descartados todos los términos de una intervención, y de esta forma dicha intervención no será considerada en las siguientes etapas de la metodología, formando parte del contexto considerado como *Vacío* (grupo en el cual estarán los textos que no brindan información relevante para la detección de contextos). Luego del proceso de desambiguación y de selección de las etiquetas provenientes de la taxonomía WordNet Domains, los textos son representados mediante una matriz Etiqueta-Documento (la columna L_z representa la presencia de la etiqueta z-*th* en cada documento D), y los términos son sustituidos por sus correspondientes etiquetas Figura 3.2 (M_3).

De igual forma esta matriz puede tener dimensiones diferentes a la anterior ($c \leq r$ y $z \leq m$), ya que al sustituir los términos por las etiquetas, éstas pueden ser de los dos primeros niveles de WordNet Domains, los cuales han sido descartados por tratarse de dominios genéricos. Este proceso es explicado detalladamente en la Sección 3.1.5.

Una vez creada la matriz Etiqueta-Documento, a partir de ésta se construye la matriz de distancia entre documentos. Para ello se ha utilizado la distancia del coseno entre los documentos, ya que dicha medida proporciona el grado de similitud entre dos textos. Mientras más cerca a 1 sea el valor de la distancia del coseno, más semejantes serán los documentos 3.2 (M_4).

Finalmente se aplican los métodos de agrupamiento jerárquico a partir de la matriz de distancia anterior, obteniendo como resultado final una jerarquía semántica de contextos, en la cual cada nivel de la jerarquía se representa mediante una matriz que relaciona las etiquetas z (las cuales provienen de WordNet Domains) con los grupos k Figura 3.2 (M_5).

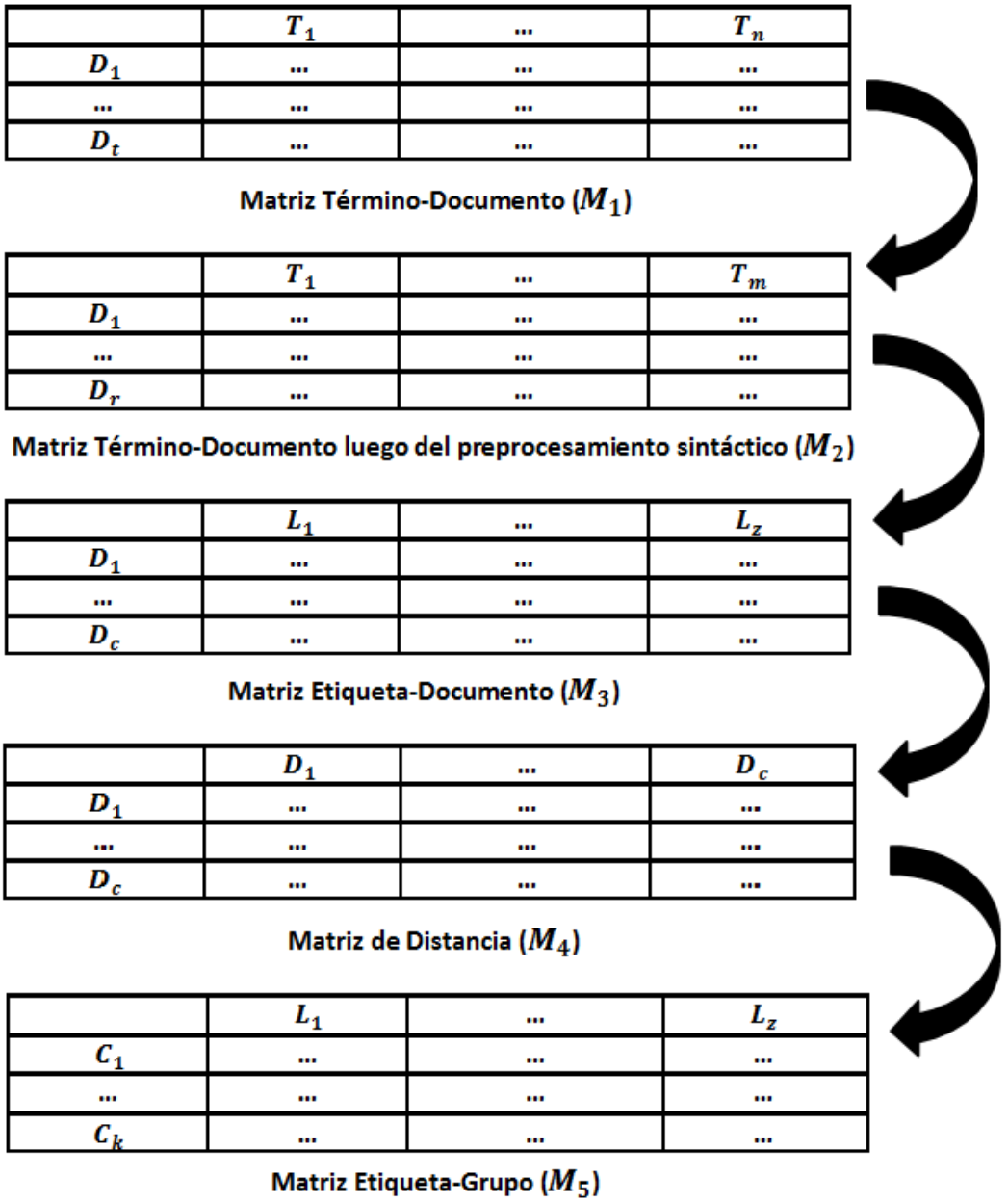


FIGURA 3.2: Modelos para la detección de contextos

3.1.4. Preprocesamiento sintáctico

Esta primera etapa consiste en una limpieza sintáctica donde se aplican filtros a los datos textuales para facilitar su procesamiento automático. Los textos originales son representados mediante una matriz que relaciona documentos y términos, como fue visto en la Figura 3.2 (M_1).

El primer paso consiste en ejecutar los procesos de etiquetado de categoría gramatical (**part-of-speech tagger**) y de reconocimiento de entidades, dichos procesos se realizan con las herramientas Stanford POS [Toutanova et al., 2003] y Stanford

NER [Finkel et al., 2005] respectivamente. Ambas herramientas han sido desarrolladas por *The Stanford Natural Language Processing Group* de la Universidad de Stanford, implementadas en Java y bajo la licencia GNU General Public License. Para más detalles sobre estas herramientas ver Apéndices B.1 y B.2

A continuación se aplica el filtro de tokenización, mediante el cual se eliminan los elementos de puntuación que complican el procesamiento automático de los textos. Seguidamente se aplican los filtros necesarios para eliminar aquellos términos que pertenecen al conjunto de palabras vacías (stop words), los que no son identificados como sustantivos por el etiquetador gramatical, los que son identificados como sustantivos propios por el identificador de entidades, así como aquellos que no se encuentren en la base de conocimiento externa MCR 3.0, ya que todos ellos no aportan información útil para la detección de los contextos.

Una vez que se han aplicado los filtros anteriores los textos se encuentran listos para ser preprocesados semánticamente, ya que sólo se mantienen los términos que en realidad aportan un valor significativo para la detección de los contextos Figura 3.2 (M_2). Se debe resaltar que los términos identificados como organizaciones por la herramienta Stanford NER han sido etiquetados como *ORGANIZATION* para el idioma Inglés y *ORG* para el Español, por lo que durante el proceso de sustitución de los términos por las etiquetas correspondientes, estos términos no se procesarán ya que su etiqueta es conocida. Además, durante la aplicación de los filtros un texto puede quedarse sin términos, y por esta razón será vinculado al grupo definido como *Vacío* como se mencionó anteriormente.

Se debe mencionar que aunque en muchos trabajos relacionados con la detección de contextos durante el preprocesamiento sintáctico se aplica el filtro de lematización (stemming), en el presente hemos considerado no aplicarlo, ya que dicho filtro consiste en truncar los términos hasta su raíz. Dado que en fases posteriores necesitamos determinar el sentido de los términos mediante los diccionarios MCR 3.0 y BabelNet, este filtro lejos de facilitar el preprocesamiento semántico, lo haría más engorroso.

3.1.5. Preprocesamiento semántico

El objetivo del preprocesamiento semántico presente en la metodología, es el de homogeneizar la representación sintáctica de los conceptos presentes en el texto.

Lo que se hará es sustituir cada término por sus etiquetas correspondientes en la taxonomía WordNet Domains de la base de conocimientos MCR 3.0.

MCR 3.0 está basado en WordNet 3.0 e integra WordNets de cinco idiomas, entre ellos Inglés y Español los cuales son de interés para el presente trabajo, haciendo la metodología independiente del idioma. Además MCR 3.0 integra recursos léxicos como: WordNet Domains [Magnini and Cavaglia, 2000], una nueva versión de Base Concepts, Top Ontology [Álvez et al., 2008], y la ontología AdimenSUMO [Pease et al., 2002] ver Apéndice B.3 .

WordNet Domains[1] es un recurso léxico creado de forma semiautomática para dotar a WordNet con etiquetas de dominios [Magnini and Cavaglia, 2000]. Para mayor información sobre WordNet Domains ver Apéndice B.3.1.

Es válido mencionar que WordNet Domains consta de cuatro niveles, de los cuales se han excluido los dos primeros niveles por tratarse de dominios muy generales Figura 3.3. Por este motivo, es posible que un texto quede vacío al sustituir los términos por las etiquetas de estos niveles, y de igual forma que en la fase anterior, dichos textos formarán parte del grupo definido como *Vacío*.

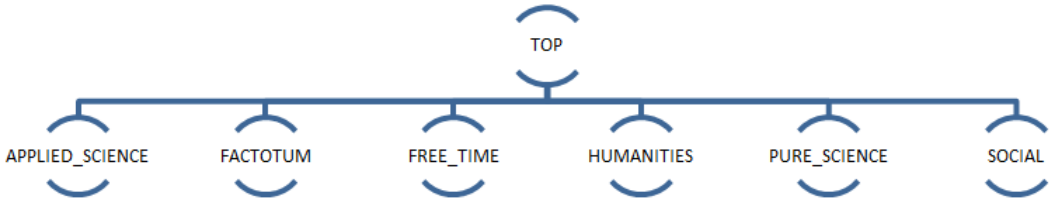


FIGURA 3.3: Etiquetas de los dos primeros niveles de WordNet Domains

Determinar el conjunto de etiquetas de un término constituye una tarea compleja, que a su vez incluye la realización de otras tareas que tienen cierto grado de dificultad. Para obtener las etiquetas de un término, es necesario conocer el significado de dicho término, ya que cada significado tiene asociado un conjunto de etiquetas de WordNet Domains, y para conocer el significado de un término se necesita su categoría gramatical.

[1]http://wndomains.fbk.eu/

Conociendo la categoría gramatical (dato con el que se cuenta luego de realizar el preprocesamiento sintáctico), se puede llevar a cabo la desambiguación para obtener su significado. La desambiguación es una parte de vital importancia en esta etapa, pues una buena desambiguación permite que se identifiquen correctamente las etiquetas del término.

Proceso de desambiguación

Un algoritmo trivial de desambiguación es seleccionar el sentido más frecuente, en el cual se toma como sentido el más habitual de la palabra. Muchos diccionarios electrónicas almacenan valores de uso de los sentidos de las palabras, permitiendo obtener el sentido de una palabra por la frecuencia de uso. Otro grupo de algoritmos realizan la desambiguación utilizando la descripción de los diferentes sentidos de un término en los diccionarios electrónicos. La idea de éstos, es seleccionar el sentido que mayor solapamiento tenga entre el contexto actual de la palabra y los diferentes sentidos de dicha palabra en el diccionario.

En la presente tesis, se ha seleccionado la segunda opción, específicamente se ha implementado el algoritmo de Lesk [Lesk, 1986]. Dicho algoritmo determina el sentido que tiene mayor solapamiento entre sus distintas descripciones en el diccionario utilizado y el contexto actual en el que ha sido utilizada la palabra. Como diccionario se han utilizado MCR 3.0 y BabelNet, ya que son dos de los recursos multilingüe más importantes. Si la descripción del sentido analizado no está presente en MCR 3.0, entonces BabelNet es usado para obtener la descripción correcta. BabelNet es una red semántica multilingüe, y fue creada automáticamente enlazando la mayor enciclopedia web multilingüe, Wikipedia, al recurso léxico electrónico más popular, WordNet [Navigli and Ponzetto, 2012] (Apéndice B.4).

Para determinar el sentido con mayor solapamiento, se tienen en cuenta los términos compartidos entre el contexto actual donde ha sido empleado el término en cuestión y las descripciones de los sentidos de dicho término en MCR 3.0 o BabelNet. El Algoritmo 1 muestra de forma programática el proceso de desambiguación utilizado. Primeramente se necesita el contexto en el que ha sido utilizado el término $context$ y la lista de sentidos $synsetList$ en MCR 3.0 correspondientes al término que está analizando. Luego para cada sentido se calcula el solapamiento entre $context$ y la descripción de cada $synset$ en MCR 3.0 o BabelNet

($getOverlapping(synset.description, context)$). Finalmente es seleccionado el sentido cuya descripción presente el mayor solapamiento con el contexto actual $selectedSynset$. En caso que dos o más sentidos presenten el máximo valor de solapamiento, es seleccionado el sentido con mayor frecuencia de uso según el diccionario MCR 3.0 (línea 8).

Algoritmo 1 Algoritmo de desambiguación

Entrada: Listado de sentidos $synsetList$ y contexto en el que ha sido utilizado el término en cuestión $context$.
Salida: Sentido con mayor solapamiento $selectedSynset$.

$selectedSynset \leftarrow EMPTY$
$max \leftarrow 0$
1: **para todo** $synset$ en $synsetList$ **hacer**
2: $overlap \leftarrow getOverlapping(synset.description, context)$
3: **si** $overlap >= max$ **entonces**
4: **si** $overlap > max$ **entonces**
5: $max \leftarrow overlap$
6: $selectedSynset \leftarrow synset$
7: **si no**
8: **si** $synset.freq > selectedSynset.freq$ **entonces**
9: $selectedSynset \leftarrow synset$
10: **fin si**
11: **fin si**
12: **fin si**
13: **fin para**
14: **devolver** $selectedSynset$

En la Tabla 3.1 se muestra el proceso de desambiguación para la palabra *BANK* en el texto *"Due to their importance in the financial system and influence on national economies, banks are highly regulated in most countries."* ("Debido a su importancia en el sistema financiero y a su influencia en las economías nacionales, los bancos están altamente regulados en la mayoría de los países."). La primera columna representa las diferentes descripciones de el término *BANK* en MCR 3.0, la segunda presenta los identificadores de los sentidos ordenados por su frecuencia de uso (donde el menor valor corresponde con al significado más usado), y finalmente, la tercera columna los resultados del solapamiento de la descripción de cada sentido de la palabra *BANK* en MCR 3.0 con el contexto actual donde ha sido utilizada la palabra. Para este ejemplo se selecciona el segundo significado, ya que su descripción es la que tiene mayor solapamiento. Si más de un significado tiene el mismo solapamiento, entonces se selecciona el sentido con mayor frecuencia de uso.

TABLA 3.1: Ejemplo de desambiguación para la palabra BANCO (*BANK*)

Descripción	Sentido	Solapamiento
sloping land (especially the slope beside a body of water)	1	4
a financial institution that accepts deposits and channels the money into lending activities	2	**7**
a supply or stock held in reserve for future use (especially in emergencies)	5	3
the funds held by a gambling house or the dealer in some gambling games	6	3
a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force	7	4
a container (usually with a slot in the top) for keeping money at home	8	5
a building in which the business of banking transacted	9	3
a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)	10	3

Como se mencionó anteriormente, el principal objetivo de esta etapa es homogeneizar la representación sintáctica de los conceptos presentes en los textos, y así mejorar la calidad de los algoritmos de agrupamiento jerárquico. Este problema queda completamente resuelto, puesto que al sustituir cada término por el conjunto de etiquetas correspondientes, el número máximo de etiquetas diferentes será la cantidad de nodos de la taxonomía WordNet Domains. Este proceso se explica detalladamente en la siguiente sección.

3.1.6. Agrupamiento jerárquico

Una vez homogeneizados los textos, se procede a realizar el agrupamiento jerárquico de los textos a partir de las etiquetas de WordNet Domains. Para ello es necesario crear la matriz de pesos Etiqueta-Documento como se puede apreciar en la Figura 3.2 (M_3).

Para cada documento D_i y los términos $T = \left\{t_j / M_{2_{ij}} > 0\right\}, \forall t_j \in T$, consideramos $\mathcal{L}_{t_j} \subset \{L_1, L_2, ..., L_z\}$ el conjunto de etiquetas por el cual puede ser sustituido el término t_j. Luego $\forall L_h \in \mathcal{L}_{t_j}$ el peso de la etiqueta L_h correspondiente al término t_j de se define de la siguiente forma:

$$w_i(t_j, L_h) = \frac{1}{n_{t_j} H_{D_i}} \tag{3.1}$$

Donde n_{t_j} es el cardinal del conjunto de etiquetas y H_{D_i} es el cardinal del conjunto de términos. Dado que una etiqueta puede sustituir a varios términos de un mismo documento, el peso total de la etiqueta L_n en el documento D_i estará dado por la Ecuación 3.2

$$M_3(i, h) = \sum_{j=1}^{p} w_i(t_j, L_h) \tag{3.2}$$

La Figura 3.4, muestra un ejemplo concreto del algoritmo utilizado para crear la matriz de pesos. Partiendo de un texto original, éste es analizado sintácticamente donde se aplican todos los filtros para eliminar todos aquellos términos que no aportan información para la detección de contextos. En este caso sólo permanecen los términos *FINANCING y HOTEL*. Luego estos términos son preprocesados semánticamente, determinando para cada término todos los posibles sentidos presentes en la base de conocimiento MCR 3.0, para posteriormente sustituir el término por la o las etiquetas del sentido o synset seleccionado en el proceso de desambiguación.

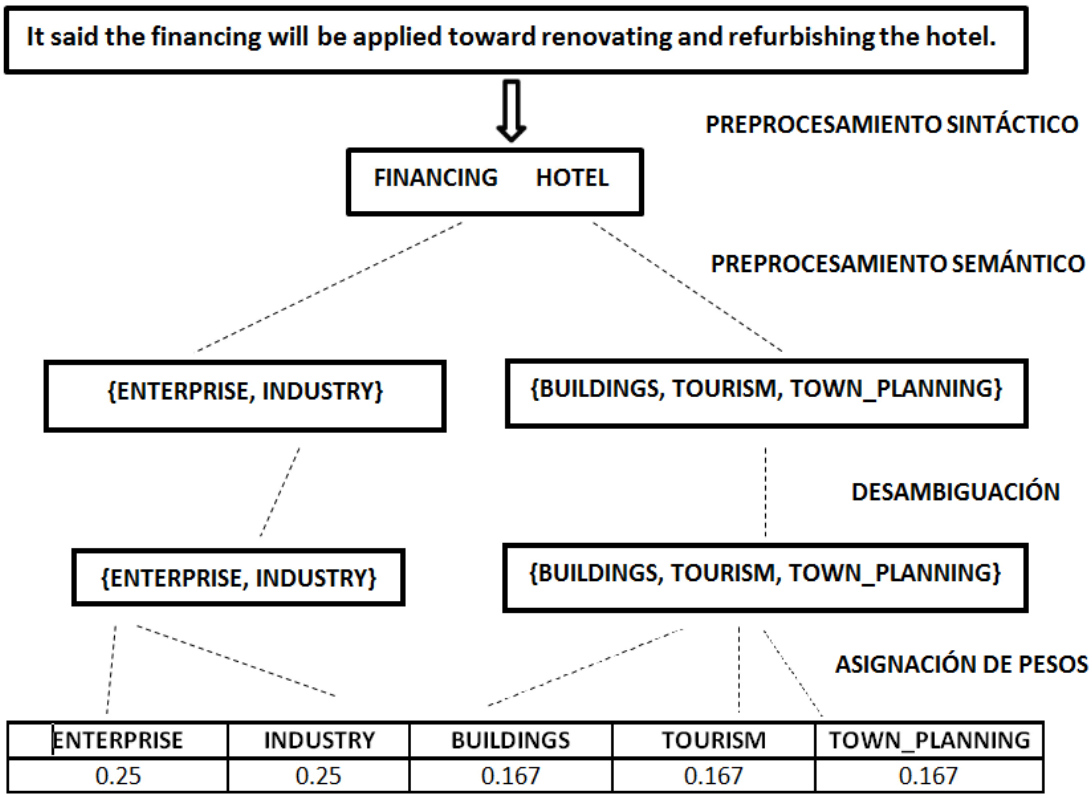


FIGURA 3.4: Creación de la matriz de pesos Etiqueta-Documento

En el ejemplo de la Figura 3.4, el término *FINANCING* presenta un sólo sentido en MCR 3.0, el cual ha sido etiquetado con *ENTERPRISE y INDUSTRY*. En este caso particular, como el término tiene un sentido, el proceso de desambiguación no es necesario. En caso de ser necesario, se utilizaría la heurística explicada en la Sección "Proceso de Desambiguación" para seleccionar el significado más apropiado, y posteriormente el término es sustituido por las etiquetas correspondientes al sentido seleccionado. Lo mismo pasa para el término *HOTEL* el cual tiene un sólo sentido, y ha sido etiquetado con *BUILDINGS, TOURISM y TOWN PLANNING*. Una vez que todos los términos han sido sustituidos por sus etiquetas correspondientes, se crea la matriz de pesos Etiqueta-Documento, la cual es utilizada como datos de entrada en los algoritmos de agrupamiento jerárquico. Para el caso específico anterior, el término *FINANCING* es sustituido por dos etiquetas por lo que $n_{t_j} = 2$ y $H_{D_i} = 2$, por lo que $w_i(FINANCING,$
$ENTERPRISE) = \frac{1}{2*2} = 0{,}25$ y $w_i(FINANCING, INDUSTRY) = 0{,}25$, mientras el término *HOTEL* es sustituido por tres etiquetas, por lo cual $n_{t_j} = 3$, $H_{D_i} = 2$, $w_i(HOTEL, BUILDINGS) = \frac{1}{3*2} = 0{,}167$, $w_i(HOTEL,$

$TOURISM) = 0{,}167$ y $w_i(HOTEL, TOWNPLANNING) = 0{,}167$.

Creada la matriz de pesos Etiqueta-Documento, a partir de ésta se construye la matriz de distancia Figura 3.2 (M_4). Dicha matriz contiene la distancia del coseno entre cada documento, y será utilizada por el algoritmo de agrupamiento para conocer en cada iteración cuáles grupos debe agrupar (los más semejantes). A continuación se presenta un ejemplo que detalla el proceso explicado anteriormente. Partiendo de un conjunto de documentos (Tabla 3.2), estos son preprocesados sintáctica y semánticamente. Luego de sustituir los términos por las etiquetas de WordNet Domains, se crea la matriz de pesos (Tabla 3.3) y finalmente se construye la matriz de distancia (Tabla 3.4). Se puede apreciar que para este ejemplo en concreto, tomando en cuenta los valores del coseno en la Tabla 3.4, los documentos más semejantes son D1, D2 y D4, los cuales según la matriz de pesos Etiqueta-Documento (Tabla 3.3) abordan principalmente temas relacionados con "ADMINISTRATION".

TABLA 3.2: Documentos de ejemplos

ID	**Documento**
D1	The City Purchasing Department, the jury said, "is lacking in experienced clerical personnel as a result of city personnel policies".
D2	It urged that the city "take steps to remedy" this problem.
D3	Henry L. Bowden was listed on the petition as the mayor's attorney.
D4	His political career goes back to his election to city council in 1923.
D5	He will be succeeded by Ivan Allen Jr., who became a candidate in the Sept. 13 primary after Mayor Hartsfield announced that he would not run for reelection.
D6	Vandiver likely will mention the 100 million highway bond issue approved earlier in the session as his first priority item.

TABLA 3.3: Matriz de pesos Etiqueta-Documento

Etiqueta-Documento	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
ADMINISTRATION	0.45	0.33	0.08	0.44	0.25	0
CHEMISTRY	0	0	0	0	0	0.08
ECONOMY	0.125	0	0	0	0	0
GEOGRAPHY	0.15	0.33	0	0.11	0	0
GRAMMAR	0.125	0	0	0	0	0
LAW	0.06	0	0.33	0	0	0.08
LITERATURE	0	0	0	0	0	0.08
PERSON	0	0	0.08	0	0.08	0
PHYSICS	0	0	0	0	0	0.08
POLITICS	0	0	0.08	0.17	0.58	0
RELIGION	0	0	0.33	0	0	0
SOCIOLOGY	0	0	0	0	0	0.33
THEOLOGY	0	0	0	0.17	0	0
TOWN PLANNING	0.08	0.33	0.08	0.11	0.08	0
TRANSPORT	0	0	0	0	0	0.33

TABLA 3.4: Matriz de distancia

Documentos	**D1**	**D2**	**D3**	**D4**	**D5**
D2	0.75907212				
D3	0.25232604	0.19245009			
D4	0.83152184	0.73029674	0.22838672		
D5	0.36004115	0.29814240	0.25819889	0.63958899	
D6	0.01992048	0.00000000	0.11111111	0.00000000	0.00000000

Selección del número de grupos

Al ejecutar el algoritmo de agrupamiento jerárquico se crea una jerarquía, la cual mediante niveles describe todo el proceso de agrupamiento (desde que cada texto es considerado un grupo, hasta que quede un único grupo que contenga todos los textos), esto específicamente para los algoritmos de agrupamiento jerárquico aglomerativos como es el caso que nos ocupa. Nuestro siguiente paso será seleccionar el número de grupos para el cual el algoritmo tiene un mejor comportamiento.

Para ello vamos a realizar cortes sobre la jerarquía obtenida teniendo en cuenta el número de grupos de las particiones correspondientes, y como medida de bondad del agrupamiento utilizaremos el Coeficiente de Silueta [Rousseeuw, 1987], ya que permite determinar de forma gráfica para qué cantidad de grupos el algoritmo tiene un mejor rendimiento. Esta medida será explicada con más detalles en la Sección 3.1.8. La Figura 3.5 muestra un ejemplo genérico de posibles valores del Coeficiente de Silueta, y se puede apreciar que el mayor valor para dicho coeficiente se obtiene para 25 grupos. En la Sección 3.2 se explica detalladamente el eso de esta medida. En el presente trabajo se analizará el comportamiento de tres algoritmos de agrupamiento jerárquico (Complete Link, Average, y Ward's Method), en todos los casos utilizando siempre como medida de similitud la distancia del coseno.

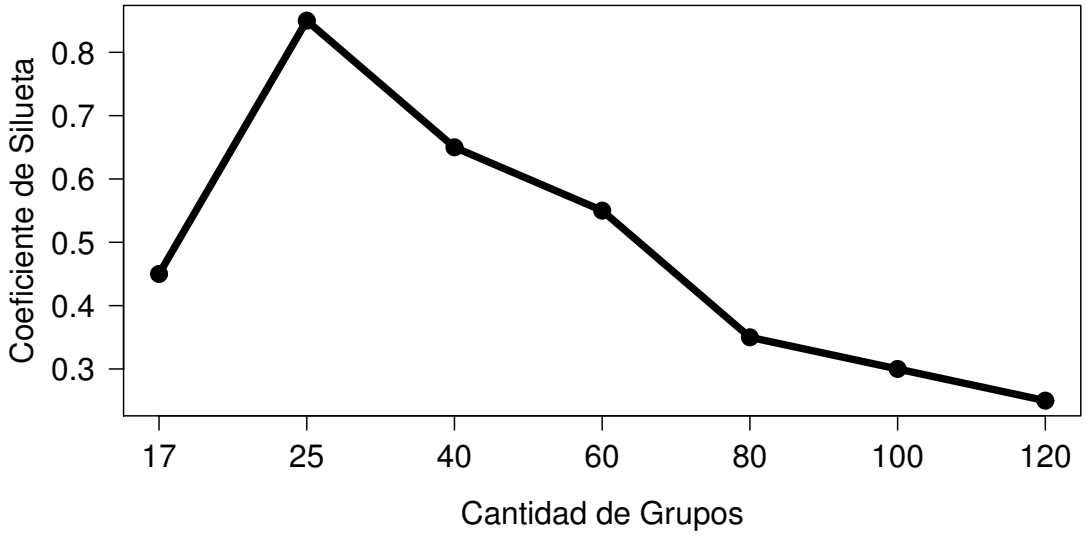


FIGURA 3.5: Ejemplo de gráfica del Coeficiente de Silueta

3.1.7. Etiquetado de los grupos

El etiquetado de grupos está estrechamente relacionado con el agrupamiento de documentos. Este proceso intenta seleccionar etiquetas descriptivas para los grupos obtenidos a través de un algoritmo de agrupamiento, ya sea de agrupamiento jerárquico o no jerárquico. Generalmente las etiquetas se obtienen al analizar los documentos pertenecientes a un grupo. Una buena etiqueta no sólo resume la idea central de los documentos agrupados, sino que los diferencia de los otros grupos. La selección de etiquetas de los grupos constituye una tarea de gran importancia, sobre todo

en aplicaciones relacionadas con el análisis de datos, donde el usuario final necesita conocer de qué trata determinado grupo [Manning et al., 2008].

Los métodos de etiquetado de grupos se clasifican en *Etiquetado interno de grupos* y *Etiquetado diferencial de grupos*. El primero selecciona las etiquetas de un grupo teniendo en cuenta únicamente el contenido del grupo que se está analizando, es decir, no se tiene en cuenta la información de los demás grupos. El etiquetado interno de grupos puede usar una variedad de métodos, tales como encontrar términos que ocurran frecuentemente en el centroide (Media Aritmética) o encontrar el documento que queda más cercano al centroide. Por otra parte, el segundo selecciona las etiquetas de un grupo mediante la comparación de los términos en un grupo con los términos que aparecen en otros grupos. Las técnicas utilizadas para la selección de características (feature selection) en recuperación de información también se pueden aplicar al etiquetado diferencial de grupos, las más utilizadas y con mejores resultados son Información Mutua (en inglés Mutual Information MI) y Chi-Square (X^2) [Manning et al., 2008].

En el presente trabajo se han utilizado tres métodos para determinar las etiquetas más relevantes de cada grupo de textos (Media Aritmética, MI y X^2). El motivo por el cual se han seleccionado técnicas pertenecientes a ambos métodos de etiquetado de grupos (*Cluster-internal labeling* y *Differential cluster labeling*) se debe a que resulta interesante analizar cuando los diferentes grupos obtenidos comparten una misma etiqueta, y cuando dichas etiquetas son excluyentes. A continuación las fórmulas utilizadas para calcular MI y X^2:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right) \tag{3.3}$$

donde $p(x,y)$ es la función de distribución de la probabilidad conjunta de las dos variables, $p1(x)$ es la función de distribución de X y $p2(y)$ es la función de distribución de Y. En el caso que nos ocupa, el etiquetado de grupos, la variable X está asociada a la pertenencia de un documento a un grupo, y Y está asociada a la presencia de un término en un documento seleccionado al azar. Ambas variables pueden tomar valores *{0,1}*.

$$X^2 = \sum_{a \in A} \sum_{b \in B} \frac{(O_{a,b} - E_{a,b})^2}{E_{a,b}} \tag{3.4}$$

donde $O_{a,b}$ representa la frecuencia observada de *a* y *b* a la vez y $E_{a,b}$ es la frecuencia con la cual se espera que ambos ocurran. Cuando es aplicado en el etiquetado de grupos, *A* es asociado al hecho de que un documento pertenece a un grupo y *B* es asociado a la presencia de un término en un documento. Ambas variables pueden tomar los valores *{0,1}*. Las etiquetas asociadas a un grupo de textos, pueden ser expresadas en forma de matriz, como se ve en la Figura 3.2 (M_4).

3.1.8. Evaluación de la metodología

Son muchos los trabajos relacionados con la detección de contextos mediante el uso de algoritmos de agrupamiento jerárquico. La gran mayoría de estos trabajos, emplean medidas de evaluación supervisadas (*purity* y *entropy*), al contar con conjuntos de datos etiquetados manualmente por expertos [Zhao and Karypis, 2002, Zhao and Karypis, 2004]. Estas medidas no pueden ser aplicadas en el presente trabajo, ya que los textos utilizados para la experimentación no se encuentran etiquetados.

Por el motivo explicado anteriormente, resulta muy difícil establecer un estudio comparativo entre la metodología propuesta y trabajos anteriores, debido principalmente a que no se cuenta con información previa (documentos etiquetados) para los conjuntos de datos utilizados en los experimentos. Se debe mencionar que aunque Reuters-21578 (conjunto de datos utilizado en el presente capítulo) se encuentra etiquetado, dichas etiquetas no sirven como punto de comparación, pues una vez agrupados los textos por las etiquetas de WordNet Domains (las cuales han sustituido los términos originales de los textos), los grupos serán anotados a partir de dichas etiquetas. Por tal motivo en el presente trabajo sólo se pueden utilizar medidas de evaluación no supervisadas para evaluar los algoritmos de agrupamiento jerárquico clásico sin aplicar la metodología y cuando es aplicada. Como resultado de lo anterior, se ha seleccionado como medida de evaluación el Coeficiente de Silueta Ecuación 3.5 [Rousseeuw, 1987], dado que incluye la calidad de la separación y la cohesión de los grupos.

$$S(i) = \frac{a(i) - b(i)}{max\left\{a(i), b(i)\right\}} \tag{3.5}$$

Donde *a(i)* representa el promedio de la distancia del elemento *i* a los demás elementos del mismo grupo (cohesión) y *b(i)* la distancia al centroide del grupo más

cercano (separación). Esta medida permite determinar el número de grupos para el cual los algoritmos de agrupamiento jerárquico tienen un mejor rendimiento mediante la identificación de un punto de inflexión. Los valores de este coeficiente están en el intervalo de [-1;1], siendo 1 el valor ideal que deben alcanzar los distintos algoritmos de agrupamiento jerárquico.

3.2. Experimentos

En esta sección realizaremos un conjunto de experimentos para demostrar la viabilidad de nuestra propuesta. Para ello hemos seleccionado seis conjuntos de datos (tres en inglés y tres en español) sobre los cuales aplicaremos tres algoritmos de agrupamiento jerárquico, en cada caso aplicaremos dichos algoritmos haciendo uso de nuestra metodología y sin la metodología.

La cuestión principal que queremos demostrar, es que nuestra propuesta funciona independientemente de la fuente de datos de los textos, del idioma en que se encuentren los textos, así como del algoritmo de agrupamiento utilizado, y que es capaz de mejorar los resultados cuando no se aplica la metodología propuesta en el presente capítulo basada en el uso de una ontología.

3.2.1. Conjunto de datos

Los conjuntos de datos utilizados para la experimentación forman parte de Reuters-21578 [Lewis et al., 2004], Semcor (Inglés y Español) [Banea et al., 2008], la Agencia de Noticias EFE, Twitter y Dreamcatchers. En la Tabla 3.5 se describen la principales características de cada conjunto (**cantidad de documentos, idioma, tipo de documento, cantidad de términos diferentes y cantidad total de términos**). Como se mencionó, tres de ellos presentan los textos en español, para así demostrar que nuestra propuesta no depende del idioma de los textos.

Reuters-21578 contiene artículos de noticias y ha sido ampliamente utilizado para evaluar algoritmos de agrupamiento. Por su parte, Semcor es un corpus que cubre varios subtópicos dentro de tópicos más generales como *Deporte*, *Política*, *Moda*, *Educación*, entre otros. Para los experimentos seleccionamos los textos en Inglés y Español de todas las sentencias *Objective* presentes en Semcor. El conjunto de datos EFE es una colección de artículos, disponibles gracias a la Agencia Española de Noticias EFE . Los artículos son de mayo del 2000.

TABLA 3.5: Descripción de los conjuntos de datos

Conjunto	Cantidad de documentos	Idioma	Tipo de Documento	Cantidad de términos diferentes	Número total de términos
Reuters-21578	4340	Inglés	Noticias	5520	106337
Semcor	4113	Inglés	Noticias	3786	14509
Twitter	5000	Inglés	Tweets	3189	12915
Semcor	3933	Español	Noticias	2476	12121
EFE	6234	Español	Noticias	3329	33158
Dreamcatchers	5000	Español	Comentarios	1661	8851

Además se han utilizado textos de dos redes sociales (Twitter y Dreamcatchers) con el objetivo de demostrar que la metodología propuesta para la detección automática de contextos, tiene aplicación en cualquier entorno y dominio. La primera es una de las redes sociales más populares y de las más utilizadas en investigaciones relacionadas con el tema. La segunda ha sido desarrollada bajo un enfoque colaborativo entre sus miembros, y se cuenta con la base de datos que le da soporte. Los datos seleccionados de Twitter y Dreamcatchers se encuentran en inglés y español respectivamente. Los datos de Twitter fueron descargados de Sentiment140[2], se encuentran en formato CSV y consta de seis campos, entre ellos el texto de los tweets el cual será utilizado en el presente trabajo.

3.2.2. Evaluación

Como ya hemos mencionado, una fase vital de nuestra metodología es el proceso de evaluación, ya que nos permitirá conocer si la metodología es viable o no y en qué medida. Para cada conjunto de datos, hemos experimentado con tres algoritmos clásicos de agrupamiento jerárquico (Complete Link o Enlace Completo, Average o Media y Ward's Method o Método de Ward) y los mismos algoritmos utilizando nuestra propuesta basada en el uso de una ontología (Enlace Completo(OB), Media(OB), y Método de Ward(OB)) OB por la traducción al inglés de basado en ontología, y así poder establecer una comparación de los resultados de dichos algoritmos cuando aplicamos nuestra metodología y cuando no es aplicada.

La calidad de los grupos obtenidos, es evaluada mediante el Coeficiente de Silueta. En el presente trabajo se han realizado cortes en 3, 5, 7, 10, 13, 17 y 25 grupos para los conjuntos de datos Reuters y EFE, para Semcor se han realizado cortes en 17, 25, 30, 35, 40, 45 y 50 grupos, mientras para Twitter y Dreamcatchers se han hecho cortes en 17, 25, 40, 60, 80, 100 y 120 grupos, y en cada caso se ha determinado el Coeficiente de Silueta, para de esta forma establecer la cantidad de grupos para cada conjunto o lo que es lo mismo la cantidad de grupos de contextos.

3.2.3. Resultados y discusión

De forma general, los resultados obtenidos para todos los conjuntos de datos son buenos. Las Figuras 3.6-3.11 muestran los valores de Coeficiente de Silueta para cada conjunto. Como se puede apreciar, cuando se realiza el agrupamiento de los

[2]http://www.sentiment140.com/

textos empleando nuestra propuesta basada en el uso de una ontología de dominio (Enlace Completo(OB), Media(OB), y Método de Ward(OB)), los resultados mejoran de forma general. Los valores del Coeficiente de Silueta están en el rango de -0.06 a 0.52, el cual es un buen resultado si tenemos en cuenta que se trata de conjuntos de datos reales.

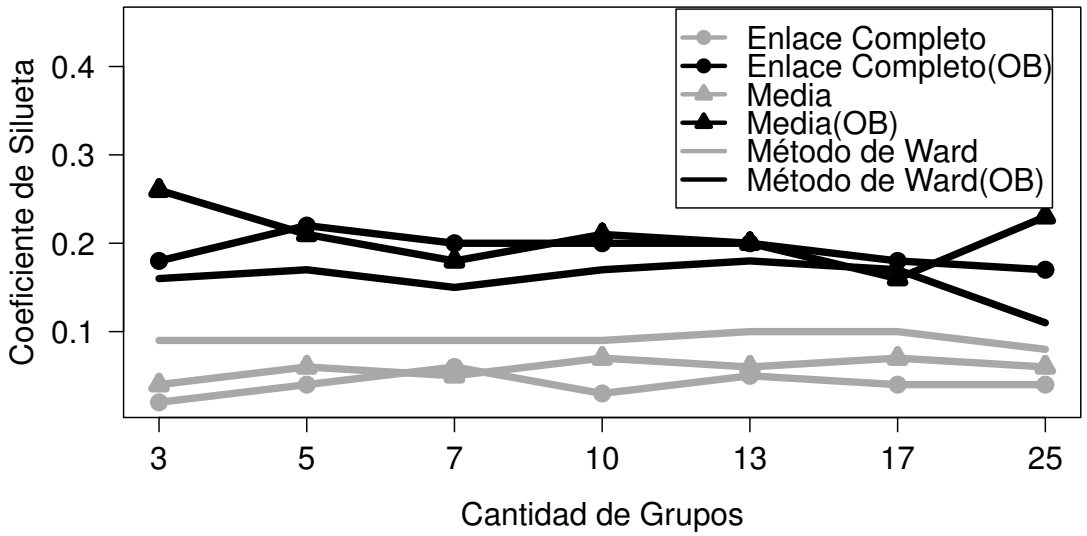


FIGURA 3.6: Coeficiente de Silueta para el conjunto Reuters (inglés)

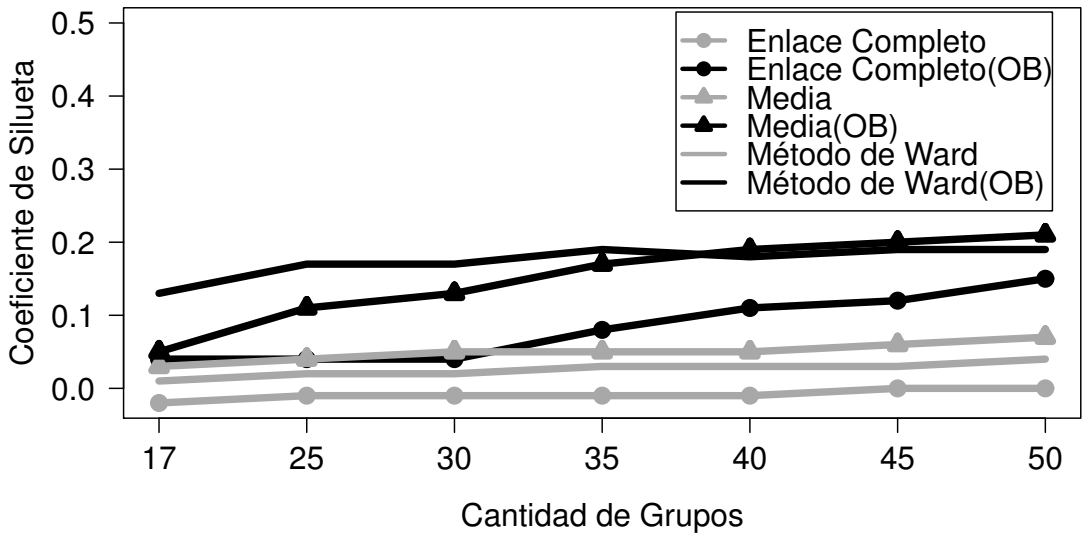


FIGURA 3.7: Coeficiente de Silueta para el conjunto Semcor (inglés)

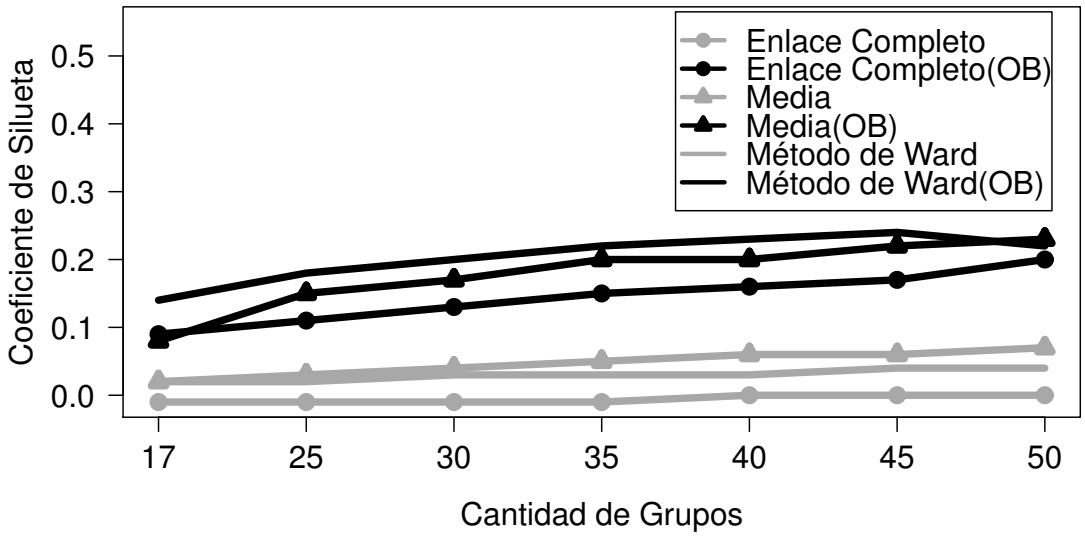


FIGURA 3.8: Coeficiente de Silueta para el conjunto Semcor (español)

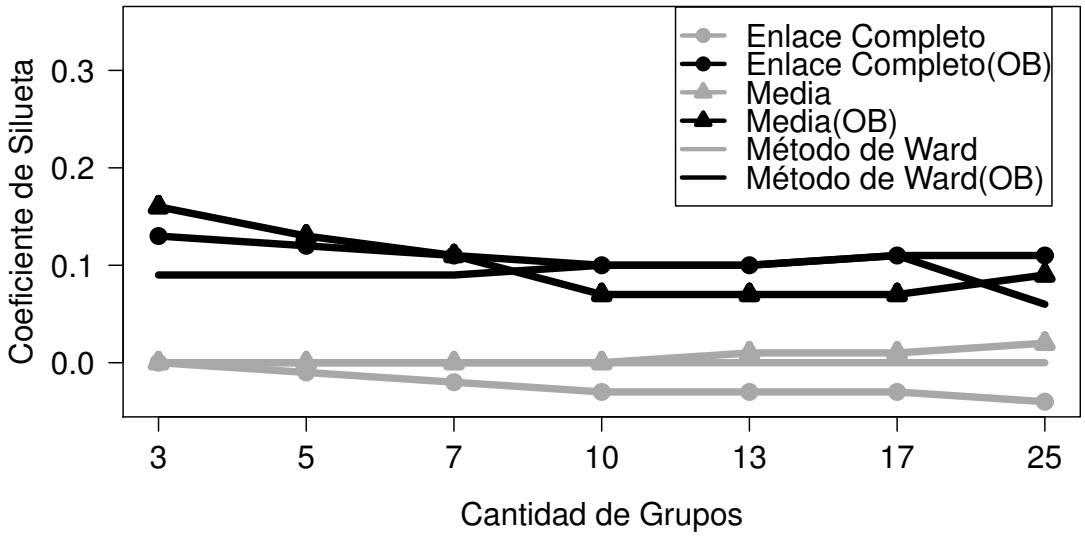


FIGURA 3.9: Coeficiente de Silueta para el conjunto EFE (español)

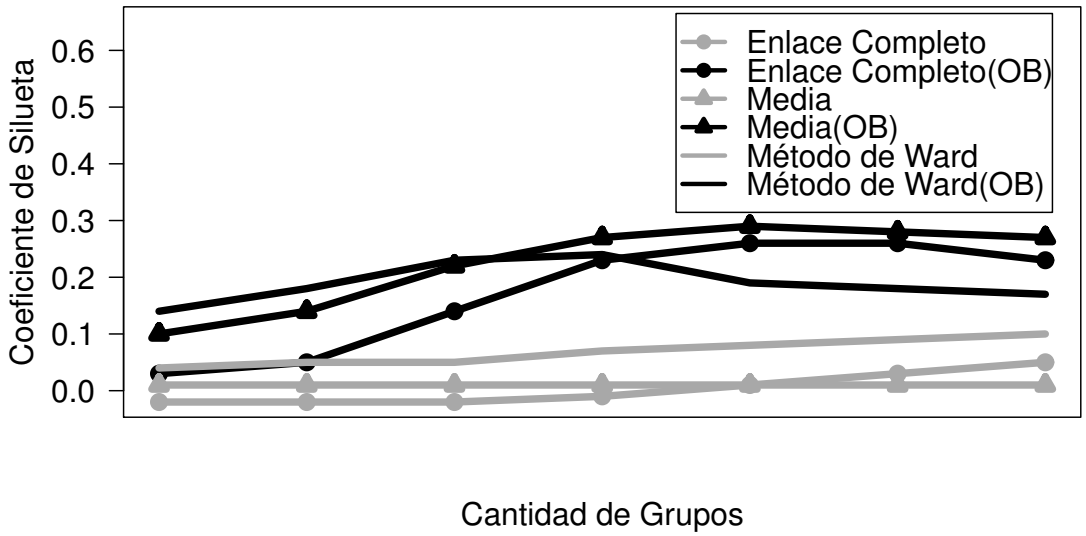


FIGURA 3.10: Coeficiente de Silueta para el conjunto Twitter (inglés)

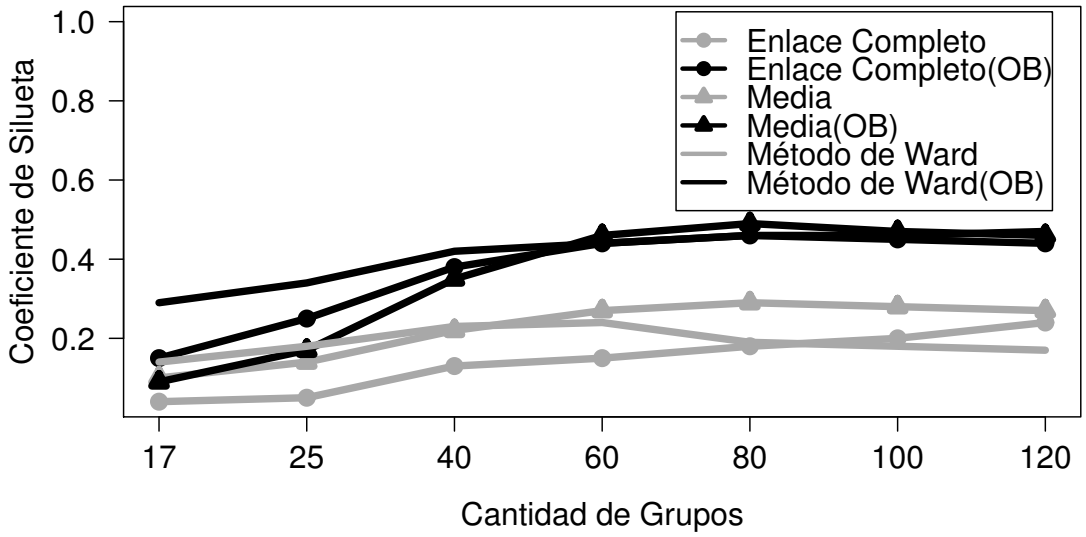


FIGURA 3.11: Coeficiente de Silueta para el conjunto Dreamcatchers (español)

Para un análisis más detallado de los resultados, hemos realizado un estudio estadístico para determinar si existen diferencias significativas entre los valores obtenidos teniendo en cuenta los diferentes algoritmos de agrupamiento jerárquico utilizados, si se ha usado o no la metodología basada en ontología y el idioma de los

textos. La Figura 3.12, muestra la gráfica correspondiente al Coeficiente de Silueta con respecto a los algoritmos de agrupamiento. Para ello se ha realizado la prueba no paramétrica Kruskal-Wallis [Kruskal and Wallis, 1952]. Se puede observar que existe una pequeña diferencia entre el algoritmo Complete Link y los demás.

De igual forma la Figura 3.13 muestra la gráfica que relaciona el Coeficiente de Silueta y si fue aplicada o no la metodología propuesta para mejorar el proceso de detección automática de contextos. En este caso se utilizó la prueba no paramétrica Wilcoxon [Wilcoxon, 1945], ya que existen sólo dos grupos de valores (**Método clásico** y **Basado en Ontología**), y podemos concluir que existen diferencias significativas, donde nuestra propuesta brinda los mejores resultados.

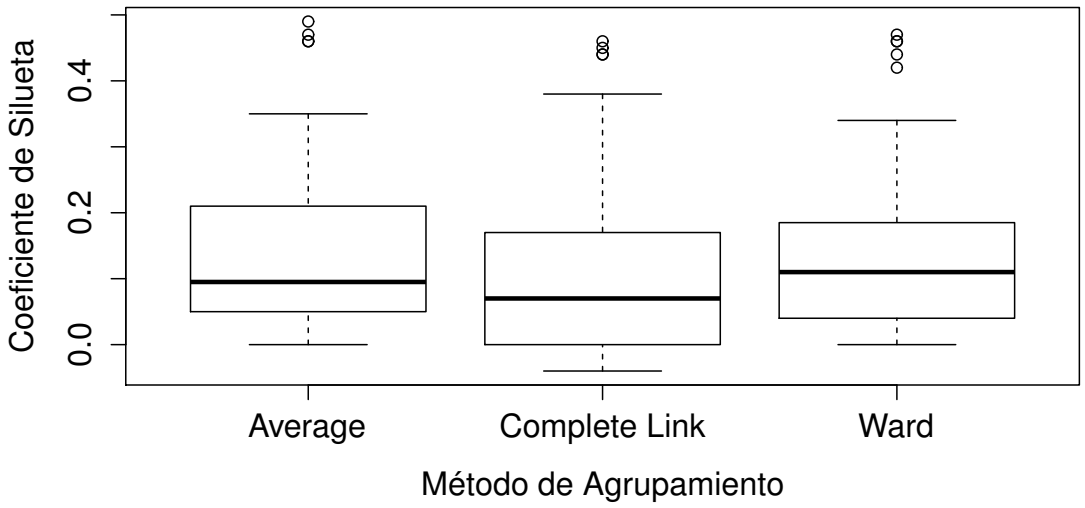


FIGURA 3.12: Gráfica que relaciona el Coeficiente de Silueta y los algoritmos de agrupamiento jerárquico

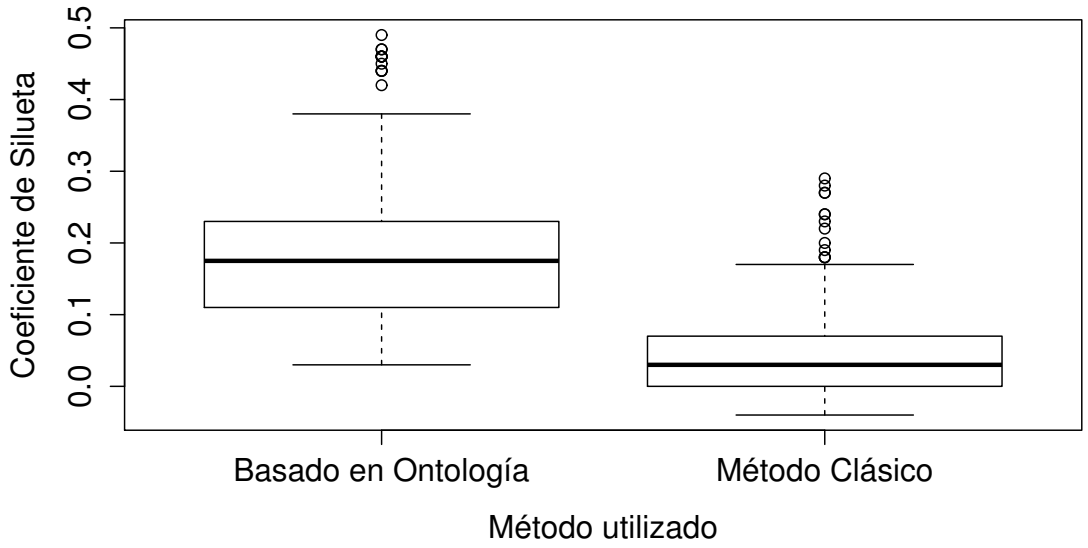


FIGURA 3.13: Gráfica que relaciona el Coeficiente de Silueta y el método utilizado para la detección de contextos

Como se mencionó en la Tabla 3.5, los textos de tres conjuntos de datos están en Inglés y los otros tres en Español. Además, como se muestra en las Figuras 3.7 y 3.8 (valores del Coeficiente de Silueta para el conjunto Semcor en Inglés y Español respectivamente), los resultados para ambos conjuntos son muy similares. La Figura 3.14 muestra la gráfica del Coeficiente de Silueta con respecto al idioma de los conjuntos de datos. Al igual que el caso anterior se utilizó la prueba de Wilcoxon y se debe resaltar que no existen diferencias significativas. Por tanto, podemos decir que nuestra propuesta basada en ontología para la detección de contextos es independiente del idioma, siempre y cuando la base de conocimiento utilizada brinde soporte para el idioma en que se encuentren los textos a analizar.

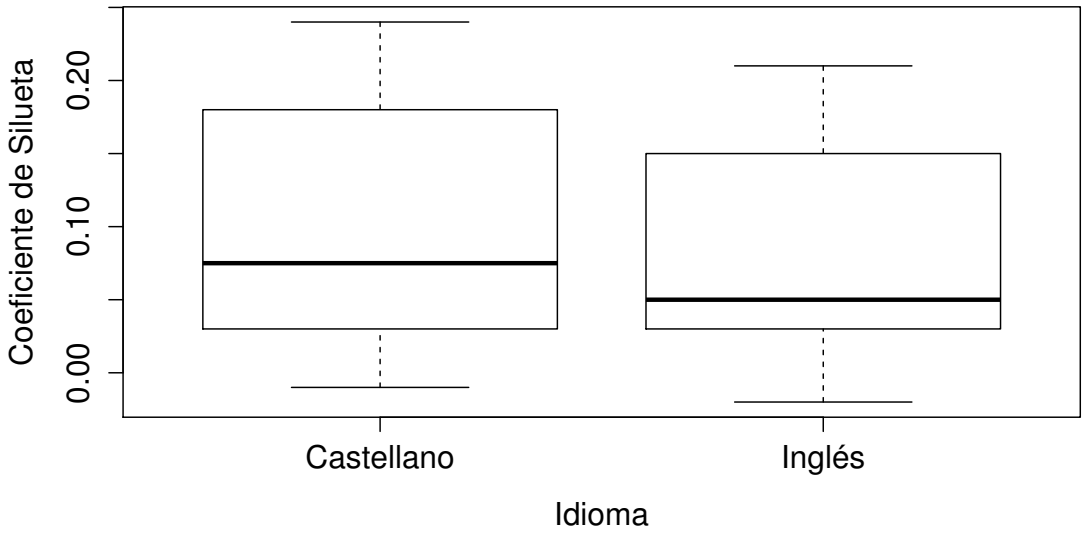


FIGURA 3.14: Gráfica que relaciona el Coeficiente de Silueta y el idioma para el conjunto Semcor

3.3. Conclusiones

En este capítulo, se presenta una propuesta para la detección automática de contextos basada en el uso de una ontología. El aporte principal consiste en permitir a los usuarios organizar los textos automáticamente en grupos (contextos) sin tener información previa, donde cada grupo constituye un conjunto de tópicos. Nuestra metodología puede ser aplicada en diferentes dominios sin importar los temas abordados en ellos, así como tampoco depende del idioma de los textos, ya que la base de conocimiento utilizada (MCR 3.0) es un recurso multilingüe.

Los resultados obtenidos con seis conjuntos de datos (Reuters-21578, Semcor en Inglés y Español, EFE, Twitter y Dreamcatchers) permiten mostrar la viabilidad de la propuesta. Se evaluó el comportamiento de tres algoritmos de agrupamiento jerárquico (Enlace Completo, Media y método de Ward) sin aplicar la metodología y cuando es aplicada. De forma general, los resultados obtenidos por los algoritmos son mejores cuando se aplica nuestra propuesta basada en el uso de una ontología. En cuanto a los métodos de agrupamiento el Método de Ward brinda resultados ligeramente superiores a los demás algoritmos y por último, no existen diferencias significativas en cuanto al idioma en el cual estén escritos los textos.

Capítulo 4

Influencia de los términos con orientación sentimental en la detección de contextos: Análisis en redes sociales

En el presente capítulo, se propone analizar la influencia de los términos con orientación sentimental en la detección automática de contextos en redes sociales. Teniendo en cuenta que los datos textuales de redes sociales se encuentran redactados de una manera más coloquial, y los usuarios tienden a expresar sus sentimientos u opiniones sobre determinado producto, servicio, entidad, etc., resulta útil detectar y eliminar aquellos términos con una determinada orientación sentimental, ya que dichos términos no aportan información útil para la detección de contextos.

4.1. Marco de trabajo

El crecimiento alcanzado por las redes sociales, y como consecuencia de ello, el aumento del número de usuarios interactuando en dichas redes, ha provocado la acumulación de grandes volúmenes de datos textuales no estructurados. Por tal motivo, las redes sociales constituyen una fuente de información de gran importancia, por lo que es de esperar que organizaciones, investigadores, etc., empleen tiempo y recursos en el estudio de éstas.

Sin embargo, el gran cúmulo y falta de estructura de los textos, hace que sea prácticamente imposible su procesamiento y análisis automático de forma masiva,

motivo por el cual resulta conveniente tener los textos previamente organizados teniendo en cuenta la temática abordada. En este entorno, resulta particularmente útil, detectar automáticamente los principales contextos que son abordados y que constituyen información relevante para los usuarios.

La detección de contextos a partir de textos no estructurados, permite organizar dichos textos por temáticas, lo cual facilita su posterior análisis integrado con datos convencionales. Segmentar los datos textuales constituye una tarea de gran dificultad, ya que su heterogeneidad complica en gran medida su procesamiento y análisis de forma automática. En el capítulo anterior, hemos propuesto una metodología multilingüe para la detección automática de contextos en datos textuales. Mediante la experimentación, se demostró la viabilidad de la propuesta. Aunque dicha propuesta es independiente del idioma y de los temas abordados en los textos, se debe resaltar que los resultados obtenidos mediante dicha metodología pueden ser mejorados, principalmente si los textos a analizar provienen de redes sociales.

Esto se debe a que la detección de contextos en textos más elaborados (librerías digitales, sitios web de noticias, etc.) no es exactamente el mismo proceso que cuando los textos pertenecen a redes sociales, donde los usuarios expresan ideas y sentimientos sobre determinado tema utilizando un lenguaje coloquial, por lo que es de esperar que en los textos aparezcan con alta frecuencia términos que permiten expresar sentimientos relacionados con determinados productos, servicios, etc. Estos términos, constituyen una fuente de información de vital importancia en áreas de investigación como el análisis de sentimientos y los sistemas de recomendación, no siendo así para la detección de contextos donde pueden introducir ruido.

Motivado por la problemática anterior, en el presente capítulo se propone realizar un estudio para analizar la influencia que tienen los términos que expresan sentimientos en la detección automática de contextos en redes sociales. Para ello se propone un nuevo enfoque para mejorar la metodología para la detección automática de los principales contextos presentes en datos textuales propuesta en el Capítulo 3, la cual utiliza técnicas de minería de datos, recursos relacionados con el análisis de sentimientos, así como una base de conocimiento multilingüe. El nuevo enfoque permite identificar y eliminar los términos con orientación sentimental, con el objetivo de mejorar la detección automática de contextos en textos de redes sociales.

Para lograr esta tarea, se propone aplicar un filtro para detectar los términos con orientación sentimental durante la etapa de preprocesamiento semántico presente en

la metodología. Para ello, se utilizan los recursos SentiWordNet 3.0 [Baccianella et al., 2010], SenticNet 3 [Cambria et al., 2014] y WordNet Affect [Valitutti, 2004]. En el caso específico de los dos primeros, al contar con el valor de la polaridad para cada sentido o concepto, hemos decidido realizar un estudio el cual nos va permitir observar el comportamiento de la metodología cuando en vez de eliminar todos los términos de sentimientos, sólo eliminamos aquellos términos cuya polaridad esté por encima de un umbral determinado. Esto último resulta de gran interés, ya que en ocasiones los textos analizados contienen una gran cantidad de términos de sentimientos, por lo que muchos textos son descartados influyendo negativamente en la detección de contextos.

Además fue necesario crear un listado especial de palabras vacías (stop words) para inglés y español donde se tiene en cuenta términos usados con mucha frecuencia y que no aportan información útil para la detección de contextos (ej. términos utilizados para saludos, despedirse, preguntar por estado, etc.). Para la experimentación se han utilizado ocho conjuntos de datos, los cuales pertenecen a las redes sociales Twitter y Dreamcatchers, en inglés y español respectivamente.

En la Figura 4.1 se muestra el marco de trabajo para la detección automática de contextos. Como se puede apreciar, el marco de trabajo es el mismo que para la metodología propuesta anteriormente. Excepto para el caso específico de la fase de Preprocesamiento Semántico, donde se ha incorporado el filtro *Removal of sentiment related terms*, el cual constituye el principal aporte del presente capítulo, y será explicado detalladamente en la Sección 4.2.

El resto del capítulo está estructurado de la siguiente forma. La Sección 4.1 describe brevemente la propuesta para la detección automática de contextos. Sección 4.2, brinda una descripción detallada del proceso que permite detectar y eliminar los términos que expresan sentimientos en datos textuales de redes sociales. Seguidamente la Sección 4.3 presenta y analiza los resultados experimentales. Finalmente en la Sección 4.4 se presentan las conclusiones derivadas del análisis realizado.

4.2. Filtro para detectar y eliminar los términos con orientación sentimental

En la presente sección explicaremos en detalle el proceso (filtro) que permite detectar y descartar los términos con orientación sentimental (positiva o negativa) en

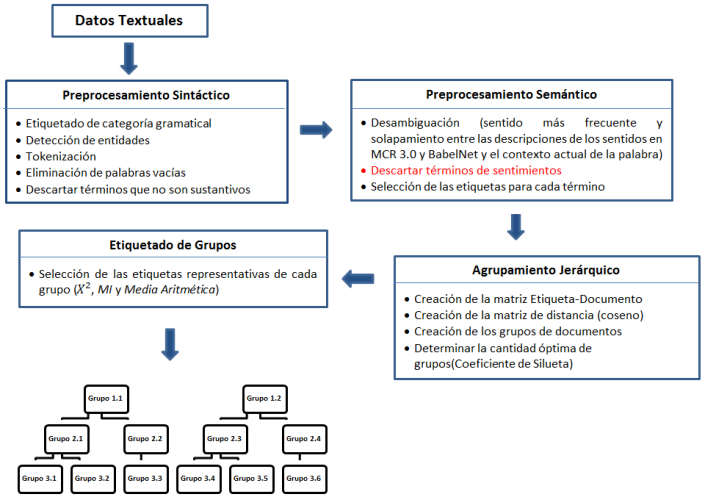


FIGURA 4.1: Marco de trabajo para el análisis de la influencia de los términos con orientación sentimental en la detección de contextos

textos de redes sociales, lo cual constituye el principal aporte del presente capítulo. La principal motivación para descartar dichos términos, parte del hecho de que se pretende detectar los principales contextos abordados en textos de redes sociales donde los usuarios mayormente expresan sentimientos u opiniones sobre determinado tema.

Estos términos, al igual que las palabras vacías, tienen una frecuencia muy elevada en estos tipos de textos, por lo que al aplicar determinadas técnicas de minería de datos para la detección de contextos, dichos términos pueden enmascarar los verdaderos contextos presentes en los textos. Dicho lo anterior, cobra gran importancia detectar y eliminar el ruido provocado por los términos con orientación sentimental.

Para detectar los términos con orientación sentimental, se han utilizado los recursos SentiWordNet 3.0 [Baccianella et al., 2010], SenticNet 3 [Cambria et al., 2014] y WordNet Affect [Valitutti, 2004], los cuales están basados en WordNet y permiten determinar si un término en un contexto determinado expresa algún tipo de sentimiento. El primer paso sería desambiguar el término, así de esta forma se conoce el verdadero significado del término en cuestión y finalmente determinar si tiene o no orientación sentimental.

De forma general, cuando los términos que expresan sentimientos son descartados, independientemente del recurso léxico utilizado, los términos más frecuentes que pertenecen a un tema determinado son más afines al mismo. Además se debe señalar que son muchos los textos que luego de aplicar el filtro de sentimiento no pueden ser procesados, ya que son descartados todos sus términos. Esto se debe

también en gran medida a que la gran mayoría de los textos de redes sociales sólo expresan sentimientos sobre cierta temática. Por este motivo, hemos decidido establecer α-cortes para eliminar sólo los términos cuya polaridad exceda un umbral determinado y de esta forma aumentar el número de textos final para la detección de contextos.

Debemos resaltar que la idea del filtro aplicado para detectar los términos que expresan sentimientos, es totalmente novedosa, ya que permite separar los términos con información relevante para la detección de contextos, de aquellos términos vinculados con algún tipo de sentimiento, útiles para muchos sistemas computacionales, pero no así para el objetivo de este trabajo. En la experimentación se realiza una comparación de los resultados obtenidos por la metodología para la detección de contextos aplicando el filtro para los términos que expresan sentimientos y sin aplicarlo. A continuación se describen las principales características y el modo de empleo de cada uno de los recursos utilizados para detectar los términos que expresan sentimientos.

4.2.1. SentiWordNet 3.0

Es un recurso léxico creado especialmente para tareas relacionadas con la clasificación de sentimientos, así como en aplicaciones basadas en la minería de opinión [Baccianella et al., 2010]. Constituye una versión mejorada de SentiWordNet 1.0 [Andrea and Fabrizio, 2006] y se encuentra disponible públicamente para propósitos de investigación. SentiWordNet 3.0 es el resultado de asignar a todos los sentidos (synsets) de WordNet dos valores numéricos que indican el valor de polaridad (positivo y negativo), y dichos valores están en el rango [0,1] [Baccianella et al., 2010].

Las Figuras 4.2 y 4.3 muestran los histogramas de las polaridades positivas y negativas en SentiWordNet respectivamente. Se puede apreciar que la gran mayoría de los sentidos presentan polaridades en el rango [0.0,0.2].

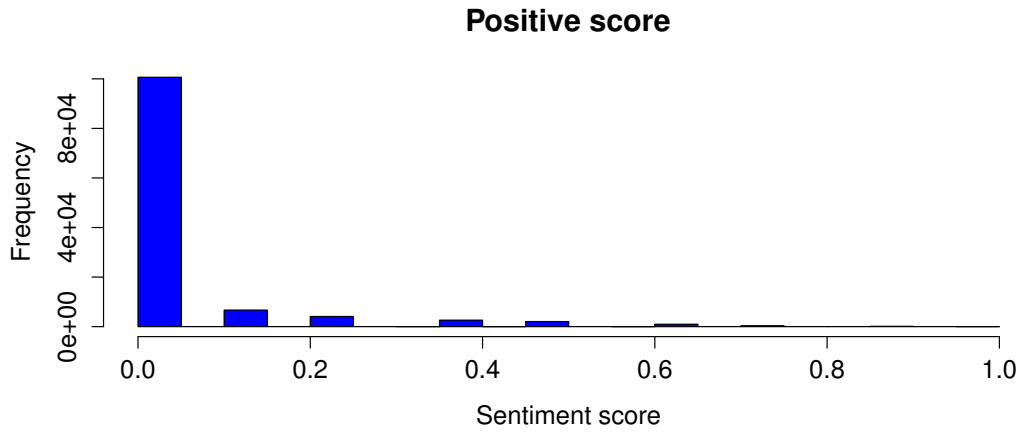


FIGURA 4.2: Histograma que muestra las polaridades positivas en SentiWordNet 3.0.

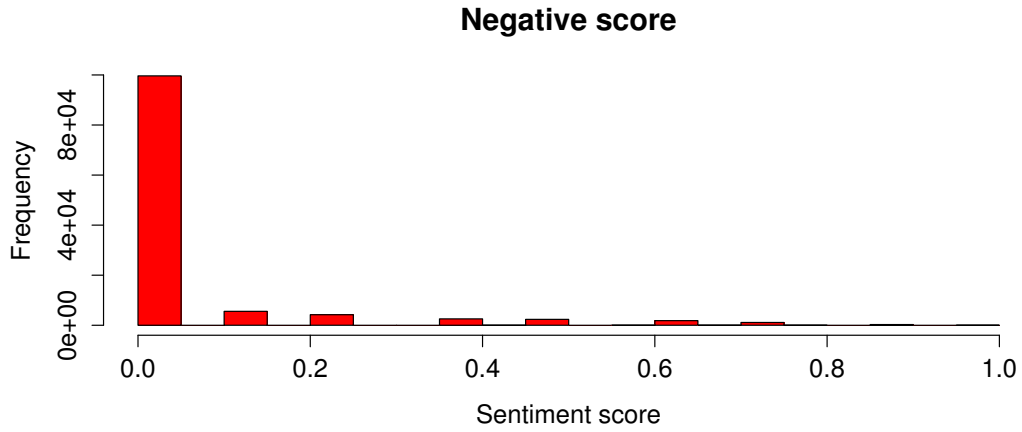


FIGURA 4.3: Histograma que muestra las polaridades negativas en SentiWordNet 3.0.

En nuestro caso, una vez desambiguado cada término, determinamos el valor positivo y negativo asignado en SentiWordNet al sentido correspondiente para cada uno de los términos analizados. En caso de que el sentido presente un valor positivo mayor que el umbral de polaridad positivo establecido o un valor negativo mayor que el umbral de polaridad negativo establecido, este término queda totalmente descartado y no se tiene en cuenta para el posterior análisis mediante el cual se detectan los principales contextos abordados en los textos.

4.2.2. SenticNet 3

Por su parte SenticNet 3 [Cambria et al., 2014], constituye uno de los recursos léxicos relacionados con el análisis de sentimientos más relevantes. Es un recurso semántico y afectivo basado en WordNet y es empleado principalmente para el análisis de sentimientos a nivel de conceptos. Esta característica, permite inferir la semántica y las sensaciones presentes en las opiniones en lenguaje natural, por lo que constituye una herramienta de gran utilidad a la hora de realizar el análisis de sentimientos basado en las características de productos y servicios [Cambria et al., 2014].

Dicho en otras palabras, SenticNet 3 más allá de evaluar una opinión sobre determinado elemento brinda la posibilidad a los usuarios de comparar una a una las características de dicho elemento. Está compuesto por 30,000 expresiones las cuales tienen un nivel de polaridad entre [-1,1].

La Figura 4.4 muestra el histograma de polaridades en SenticNet 3. De igual forma que en SentiWordNet, la gran mayoría de los sentidos presentan polaridades en el rango [0.0,0.2]. En este caso se debe mencionar que a diferencia de SentiWordNet, donde cada sentido tiene asociado un valor positivo y un valor negativo, en SenticNet para determinar la polaridad de un sentido fue necesario calcular el promedio de las polaridades que presenta cada uno de los términos que conforman dicho sentido.

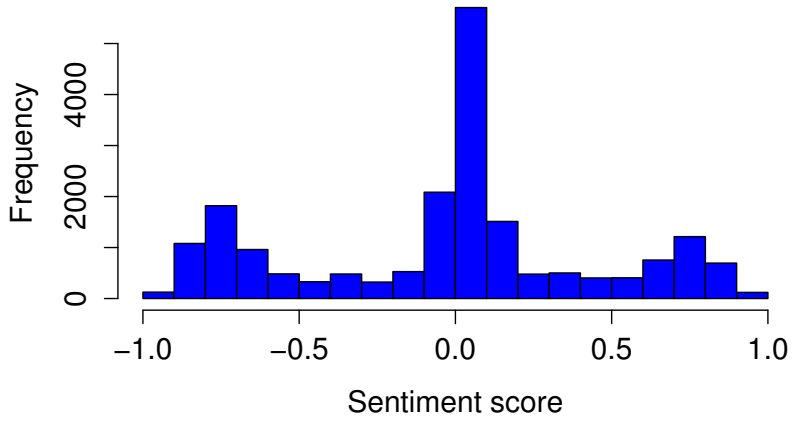


FIGURA 4.4: Histograma de polaridades en SenticNet 3.

Ha sido utilizado de forma similar a SentiWordNet, es decir, cuando un término es desambiguado, éste es descartado si su sentido en el contexto actual presenta un

valor de polaridad mayor que el umbral de polaridad positivo establecido o menor que el umbral de polaridad negativo establecido.

4.2.3. WordNet Affect

Es un recurso lingüístico para la representación léxica del conocimiento afectivo. Al igual que SentiWordNet 3.0 y SenticNet 3, está basado en WordNet. Fue creado mediante la selección y etiquetado de synsets de WordNet, que representan conceptos afectivos, y luego fue extendido mediante el uso de las relaciones entre términos y conceptos presentes en WordNet [Valitutti, 2004].

A diferencia de los dos recursos anteriores, WordNet Affect no presenta la polaridad de los distintos conceptos, en su lugar etiqueta los synsets con un conjunto de 1,903 categorías que constituyen estados afectivos. Actualmente consta de 2,874 synsets y 4,787 términos [Valitutti, 2004].

La forma en la que ha sido utilizado es similar a los dos recursos anteriores, cuando los términos son desambiguados éste es descartado si su significado en el contexto actual ha sido etiquetado con algún estado mental en WordNet Affect. Al utilizar este recurso en específico, no se puede establecer un umbral de polaridad, por lo que su uso obliga a descartar todos los términos de sentimientos,

4.2.4. Análisis gradual de la polaridad de los términos de sentimientos en la detección de contextos

Como ya hemos mencionado, merece la pena estudiar cómo se comporta la metodología para la detección de contextos cuando en lugar de eliminar todos los términos con orientación sentimental, sólo se eliminan los términos cuya polaridad supere un umbral determinado. Lógicamente, para ello sólo se podrán utilizar los recursos SentiWordNet 3.0 y SenticNet 3, los cuales proporcionan la polaridad de los diferentes sentidos.

Dicho estudio es de gran importancia, ya que los términos de sentimientos tienen un frecuencia elevada sobre todo en textos de redes sociales y al eliminarlos una gran cantidad de textos pueden quedar fuera del proceso de detección de contextos. Es por ello que en ocasiones sea conveniente eliminar una parte de los términos de sentimientos y que permanezcan más textos para la detección de contextos.

A continuación presentamos la modelización de la polaridad para construir conjuntos difusos de términos de sentimientos. Dicho modelo constituye la base de

nuestra propuesta, la cual nos va a permitir analizar de forma gradual la influencia de los términos de sentimientos en la detección de contextos.

Caso de SentiWordNet 3.0

Sea $\mathcal{V}$ el conjunto de términos difusos incluidos en SentiWordNet: $\forall t \in \mathcal{V}\ \exists \alpha_1, \alpha_2 \in [0,1]$ tal que α_1 es la intensidad de t como representación de sentimiento positivo y α_2 la intensidad de t como representación de sentimiento negativo.

Definimos $\hat{F}$ como el conjunto difuso de términos derivados de $\mathcal{V}$ cuya función de pertenencia es:

$$\forall t \in \mathcal{V}, \mu_{\hat{F}(t)} = \text{máx}\,(\alpha_1, \alpha_2) \tag{4.1}$$

Es decir para cada término recogemos la intensidad de su representación como sentimiento independientemente de que sea positivo o negativo.

Caso de SenticNet

Sea $\mathcal{W}$ el conjunto de términos difusos incluidos en SenticNet: $\forall t \in \mathcal{W}\ \exists \alpha_1 \in [-1,1]$ tal que α_1 es la intensidad de t como representación de sentimiento positivo o negativo. Por lo que t sólo puede ser de una polaridad.

Definimos $\hat{S}$ como el conjunto difuso de términos derivados de $\mathcal{W}$ cuya función de pertenencia es:

$$\forall t \in \mathcal{W}, \mu_{\hat{S}(t)} = \quad \text{mód}\ \alpha_1 \tag{4.2}$$

Obviamente también en este caso recogemos la intensidad de t como representación de sentimiento independientemente de que sea positivo o negativo.

La obtención de contextos se puede hacer ahora considerando los α-cortes de los conjuntos $\hat{F}$ y $\hat{S}$, en lugar de eliminar todos los términos de sentimientos (caso para el cual el α-corte sería 0). Esto permite flexibilizar el proceso, incluyendo un mayor número de documentos sin empeorar el Coeficiente de Silueta del agrupamiento resultante.

4.3. Experimentos

A continuación demostraremos de forma experimental la validez de nuestro enfoque. Como se ha mencionado, no contamos con información previa de los contextos presentes en los textos (categorías o etiquetas), por tal motivo debemos utilizar una medida no supervisada. En este caso se ha seleccionado el Coeficiente de Silueta [Rousseeuw, 1987], que permite determinar la cantidad de grupos para la cual los algoritmos de agrupamiento brindan un mejor resultado.

Los conjuntos de datos para evaluar el sistema pertenecen a las redes sociales Twitter y Dreamcatchers. Lo primera es una de las redes sociales más populares y de las más utilizadas en investigaciones relacionadas con el tema. La segunda ha sido desarrollada bajo un enfoque colaborativo entre sus miembros, y se cuenta con la base de datos que le da soporte. Los datos seleccionados de Twitter y Dreamcatchers se encuentran en inglés y español respectivamente, demostrando que la propuesta es independientemente del idioma.

4.3.1. Conjunto de datos

Se han seleccionado ocho conjuntos de datos, cuatro de forma aleatoria (Conjuntos 3, 4, 7 y 8) y cuatro de forma intencionada (Conjuntos 1, 2, 5 y 6) Tabla 4.1, pertenecientes a Twitter y a la Red Social Dreamcatchers. Los datos escogidos aleatoriamente, demuestran que el sistema funciona sin importar la procedencia de los datos, mientras que los seleccionados (con más de dos términos que expresan sentimientos en el caso de Twitter y más de uno para Dreamcatchers) permiten demostrar la certeza de nuestra propuesta asegurando la presencia de aquellos términos que influyen de forma negativa en la detección automática de contextos.

Además se ha variado la cantidad de documentos con el objetivo de validar la viabilidad independientemente de la cantidad de textos procesados.

Los datos de Twitter fueron descargados de Sentiment140[1], se encuentran en formato CSV y consta de seis campos, entre ellos el texto de los tweets el cual será utilizado en el presente trabajo. Se debe mencionar que se seleccionaron estos datos por estar orientados al Análisis de Sentimientos, y constituyen una fuente de gran importancia para la experimentación.

[1] http://www.sentiment140.com/

Por otra parte, se cuenta con la base de datos de Dreamcatchers, con un total de 61 tablas. La información recogida es toda la relacionada con los datos personales y de afiliación del usuario, así como, las interacciones que realiza en su perfil y con otros usuarios. Fundamentalmente aportan información textual los *Post* y sus *Comentarios*, las *Ideas* y sus *Sueños*, y el *Chat*.

4.3.2. Evaluación

En esta sección, se explica cómo se evaluará el funcionamiento de la metodología para la detección automática de contextos aplicando el filtro para eliminar los términos que expresan sentimientos con los distintos recursos y sin aplicarlo.

Primeramente, analizaremos los términos más frecuentes para los Conjuntos 4 y 8 aplicando el filtro y sin aplicarlo. Mediante el uso de nubes de etiquetas (*Tag Cloud* en inglés) podremos establecer una comparación visual de dichos términos.

Por último evaluaremos los resultados de los algoritmos de agrupamiento en la detección de contextos. Para ello se ha utilizado como medida de bondad el Coeficiente de Silueta [Rousseeuw, 1987] Ecuación 3.5, para así poder establecer una comparación entre los diferentes recursos utilizados. Primero se calculará el Coeficiente de Silueta cuando son eliminados todos los términos de sentimientos utilizando los tres recursos. Luego será calculado para los distintos umbrales de polaridad (α-cortes) seleccionados para los recursos que soportan dicha característica (SentiWordNet 3.0 y SenticNet 3).

La experimentación se ha llevado a cabo utilizando tres métodos de agrupamiento (Complete Link, Average y Ward's Method) y realizando cortes para las siguientes cantidades de grupos (17, 25, 40, 60, 80, 100 y 120) y para siete α-cortes de polaridad (0, 0.2, 0.3, 0.4, 0.5, 0.7 y 1). Se ha experimentado con estas cantidades de grupos, ya que mediante experimentos previos se comprobó que para valores menores de 17 y mayores de 120 grupos, el desempeño de los algoritmos es inferior a los valores mostrados anteriormente. En todos los casos se utilizó como medida de similitud la distancia del coseno.

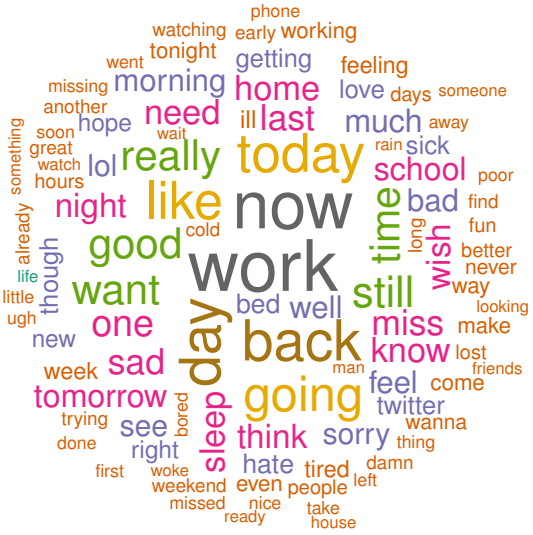
4.3.3. Resultados y discusión

Las Figuras 4.5 y 4.6 muestran los términos más frecuentes para los Conjuntos 4 y 8 respectivamente. En el caso de las Figuras 4.5 (a) y 4.6 (a) no se ha aplicado el filtro, por lo que se pueden observar un gran número de términos de sentimientos,

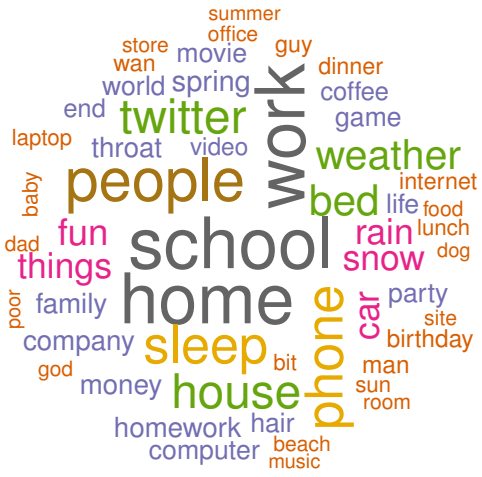
TABLA 4.1: Descripción de los conjuntos de datos

Conjunto	Cantidad de documentos	Fuente	Idioma	Intervención	Cantidad de términos diferentes	Cantidad total de términos
Conjunto 1	1000	Twitter	Inglés	Tweets	1665	4875
Conjunto 2	2000	Twitter	Inglés	Tweets	2417	9542
Conjunto 3	5000	Twitter	Inglés	Tweets	3189	12915
Conjunto 4	10000	Twitter	Inglés	Tweets	4597	25634
Conjunto 5	1000	Dreamcatchers	Español	Comentarios	808	3138
Conjunto 6	2000	Dreamcatchers	Español	Comentarios	1284	6806
Conjunto 7	5000	Dreamcatchers	Español	Comentarios	1661	8851
Conjunto 8	10000	Dreamcatchers	Español	Comentarios	2218	17141

no siendo así en las Figuras 4.5 (b) y 4.6 (b), pues al aplicar el filtro (en este caso utilizando el recurso SentiWordNet 3.0 y eliminando todos los términos de sentimientos) sólo permanecen los términos útiles para la detección de contextos. En la Tabla 4.2 se muestra el listado de términos que expresan sentimientos para cada conjunto de datos.



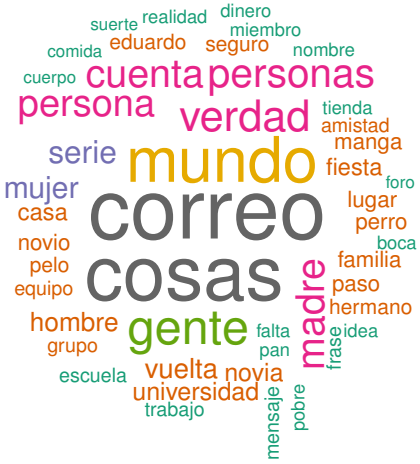
(a)



(b)

FIGURA 4.5: Nube de etiquetas para el Conjunto 4 **a)** No se aplica el filtro de sentimientos **b)** Se aplica el filtro de sentimientos (con el recurso SentiWordNet 3.0 y eliminando todos los términos de sentimientos)

(a)



(b)

FIGURA 4.6: Nube de etiquetas para el Conjunto 8 **a)** No se aplica el filtro de sentimientos **b)** Se aplica el filtro de sentimientos (con el recurso SentiWordNet 3.0 y eliminando todos los términos de sentimientos)

TABLA 4.2: Términos más frecuentes para los Conjuntos 4 y 8 cuando no se aplica el filtro para descartar los términos con orientación sentimental

Términos más frecuentes Conjunto 4 (Twitter)	**Términos más frecuentes Conjunto 8 (Dreamcatchers)**
like, good, sad, miss, wish, well, bad, feel, lol, sorry, hate, sick, love, ill, tired, feeling, better, fun, lost, great	gracias, bueno, bien, mejor, amor, gusta, buen, lindo, quiero, súper, bienvenido, mal, felicidades, feliz, amistad, preciosa

En la Tabla 4.3 se muestra la estadística relacionada con la cantidad de términos de sentimientos descartados para cada conjunto de datos y para cada recurso utilizado. En este caso se han descartados todos los términos de sentimientos (es decir α-corte es 0) y el recurso mediante el cual se han descartado más términos es con SenticNet 3.

TABLA 4.3: Cantidad de términos descartados con orientación sentimental por recurso para cada conjunto de datos

Conjunto	**SWN**	**%**	**SN**	**%**	**WA**	**%**
Conjunto 1	2904	59.57	4086	83.83	34	0.7
Conjunto 2	5775	60.52	7949	83.31	57	0.6
Conjunto 3	1696	13.13	9160	70.9	12	0.09
Conjunto 4	3281	12.8	18141	70.77	21	0.08
Conjunto 5	2081	66.32	2137	68.1	43	1.37
Conjunto 6	4147	60.93	4592	67.47	86	1.26
Conjunto 7	1314	14.85	5122	57.87	31	0.35
Conjunto 8	2533	14.78	10101	58.93	55	0.32

Las Tablas 4.4 y 4.5 muestran los valores del Coeficiente de Silueta de cada conjunto experimental. Para cada conjunto se ha experimentado con diferentes cantidades de grupos, tres algoritmos de agrupamiento jerárquico y cuando no es aplicado el filtro para eliminar los términos que expresan sentimientos **(NF)** y cuando es aplicado utilizando SentiWordNet 3.0 (**SWN**), SenticNet 3 (**SN**) y WordNet Affect (**WA**). En este caso también se han descartados todos los términos de sentimientos. Se puede observar a simple vista que cuando se aplica el filtro de sentimientos, se obtienen

mejores resultados que cuando no es aplicado, en especial cuando se utiliza el recurso SenticNet.

Con el fin de realizar un análisis más detallado de los resultados, se ha realizado un análisis estadístico para así poder determinar si existen diferencias significativas entre los valores obtenidos para las distintas cantidades de grupos, los recursos utilizados, la cantidad de textos analizadas y la fuente de los datos textuales. Las Figuras 4.7-4.9 muestran los valores del Coeficiente de Silueta con respecto a la cantidad de grupos, el recurso utilizado para detectar los términos de sentimientos y los métodos de agrupamiento jerárquico respectivamente, utilizando en todos los casos la prueba Kruskal-Wallis [Kruskal and Wallis, 1952]. A continuación se resumen las conclusiones del análisis:

- A partir de 60 grupos en adelante los valores del Coeficiente de Silueta se estabilizan, mostrando diferencias significativas con las cantidades anteriores (Figura 4.7).
- Existe una notable diferencia entre los resultados obtenidos cuando se aplica el filtro para eliminar los términos de sentimientos, especialmente SentiWordNet y SenticNet y cuando no se aplica el filtro. El recurso **SN** brinda los mejores resultados, mientras el recurso que peores resultados brinda es **WA** (Figura 4.8).
- Finalmente, no existen diferencias significativas entre los tres métodos de agrupamiento utilizados, por lo que podemos utilizar cualquiera de los algoritmos de agrupamiento jerárquico, ya que no influye en los resultados del enfoque aquí propuesto (Figura 4.9).

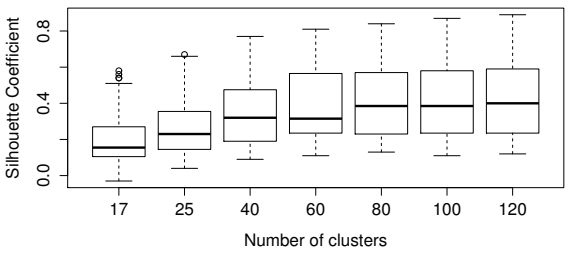


FIGURA 4.7: Boxplot entre el Coeficiente de Silueta y la cantidad de grupos

TABLA 4.4: Coeficiente de Silueta para los conjuntos del 1-4

Método de agrupamiento	**17**	**25**	**40**	**60**	**80**	**100**	**120**
Coeficiente de Silueta para el conjunto 1							
Complete Link(NF)	0.06	0.1	0.17	0.19	0.18	0.17	0.18
Complete Link(WA)	0.07	0.11	0.17	0.19	0.17	0.17	0.18
Complete Link(SWN)	0.07	0.1	0.19	0.3	0.28	0.26	0.27
Complete Link(SN)	0.24	0.43	0.54	**0.61**	0.59	0.59	0.6
Average(NF)	0.07	0.13	0.19	0.22	0.2	0.17	0.17
Average(WA)	0.07	0.12	0.19	0.22	0.2	0.17	0.17
Average(SWN)	0.09	0.15	0.26	0.31	0.29	0.26	0.26
Average(SN)	0.15	0.29	0.52	**0.62**	0.58	0.58	0.58
Ward's Method(NF)	0.12	0.11	0.11	0.15	0.15	0.17	0.19
Ward's Method(WA)	0.1	0.11	0.11	0.15	0.17	0.18	0.19
Ward's Method(SWN)	0.12	0.18	0.25	0.28	0.25	0.25	0.26
Ward's Method(SN)	0.41	0.47	0.56	0.59	0.62	0.64	**0.66**
Coeficiente de Silueta para el conjunto 2							
Complete Link(NF)	0.08	0.1	0.14	0.18	0.16	0.14	0.13
Complete Link(WA)	0.08	0.11	0.14	0.18	0.16	0.15	0.15
Complete Link(SWN)	0.07	0.1	0.17	0.25	0.3	0.29	0.26
Complete Link(SN)	0.27	0.33	0.52	0.63	0.63	**0.65**	**0.65**
Average(NF)	0.06	0.04	0.16	0.2	0.19	0.16	0.15
Average(WA)	0.04	0.09	0.17	0.21	0.2	0.16	0.15
Average(SWN)	0.1	0.14	0.24	0.32	0.33	0.31	0.28
Average(SN)	0.19	0.35	0.53	**0.65**	0.62	0.62	0.63
Ward's Method(NF)	0.12	0.1	0.09	0.11	0.14	0.12	0.13
Ward's Method(WA)	0.12	0.11	0.12	0.12	0.13	0.11	0.12
Ward's Method(SWN)	0.14	0.2	0.26	0.27	0.24	0.23	0.24
Ward's Method(SN)	0.43	0.5	0.55	0.61	0.64	0.66	**0.69**
Coeficiente de Silueta para el conjunto 3							
Complete Link(NF)	0.03	0.06	0.14	0.22	0.25	0.25	0.23
Complete Link(WA)	0.04	0.07	0.15	0.22	0.26	0.26	0.24
Complete Link(SWN)	-0.03	0.06	0.16	0.24	0.28	0.29	0.28
Complete Link(SN)	0.36	0.41	0.53	0.62	0.65	0.66	**0.67**
Average(NF)	0.13	0.15	0.23	0.26	0.28	0.28	0.26
Average(WA)	0.12	0.15	0.23	0.27	0.29	0.28	0.26
Average(SWN)	0.11	0.16	0.23	0.29	0.31	0.31	0.29
Average(SN)	0.12	0.25	0.43	0.64	**0.69**	0.67	0.63
Ward's Method(NF)	0.14	0.17	0.22	0.23	0.19	0.18	0.18
Ward's Method(WA)	0.14	0.18	0.23	0.24	0.2	0.18	0.18
Ward's Method(SWN)	0.16	0.2	0.25	0.25	0.21	0.22	0.23
Ward's Method(SN)	0.46	0.54	0.59	0.66	0.66	0.69	**0.71**
Coeficiente de Silueta para el conjunto 4							
Complete Link(NF)	0.02	0.05	0.12	0.21	0.2	0.23	0.22
Complete Link(WA)	0.03	0.06	0.13	0.22	0.23	0.24	0.22
Complete Link(SWN)	0.06	0.07	0.16	0.25	0.27	0.28	0.27
Complete Link(SN)	0.32	0.39	0.56	0.64	0.66	**0.67**	**0.67**
Average(NF)	0.04	0.1	0.16	0.23	0.26	0.26	0.25
Average(WA)	0.06	0.11	0.17	0.23	0.26	0.27	0.26
Average(SWN)	0.06	0.12	0.18	0.25	0.29	0.31	0.29
Average(SN)	0.11	0.23	0.42	0.61	0.66	**0.67**	0.63
Ward's Method(NF)	0.12	0.15	0.21	0.2	0.16	0.15	0.15
Ward's Method(WA)	0.13	0.16	0.21	0.21	0.17	0.15	0.15
Ward's Method(SWN)	0.15	0.18	0.24	0.26	0.21	0.18	0.18
Ward's Method(SN)	0.45	0.54	0.59	0.64	0.65	0.66	**0.68**

TABLA 4.5: Coeficiente de Silueta para los conjuntos del 5-8

Método de agrupamiento	17	25	40	60	80	100	120
Coeficiente de Silueta para el conjunto 5							
Complete Link(NF)	0.18	0.23	0.27	0.27	0.27	0.32	0.34
Complete Link(WA)	0.2	0.26	0.29	0.27	0.27	0.32	0.35
Complete Link(SWN)	0.21	0.29	0.45	0.54	0.54	0.56	0.56
Complete Link(SN)	0.3	0.49	0.63	0.71	0.73	**0.81**	0.76
Average(NF)	0.18	0.23	0.16	0.28	0.25	0.2	0.21
Average(WA)	0.17	0.24	0.27	0.27	0.24	0.21	0.27
Average(SWN)	0.27	0.37	0.45	0.52	0.53	0.55	0.54
Average(SN)	0.13	0.37	0.64	0.72	0.79	**0.81**	0.77
Ward's Method(NF)	0.19	0.19	0.23	0.29	0.33	0.36	0.38
Ward's Method(WA)	0.17	0.19	0.23	0.3	0.34	0.37	0.39
Ward's Method(SWN)	0.37	0.42	0.48	0.53	0.55	0.58	0.61
Ward's Method(SN)	0.54	0.61	0.71	0.77	0.82	**0.83**	0.76
Coeficiente de Silueta para el conjunto 6							
Complete Link(NF)	0.19	0.22	0.24	0.2	0.22	0.24	0.26
Complete Link(WA)	0.22	0.23	0.25	0.24	0.21	0.24	0.26
Complete Link(SWN)	0.24	0.3	0.41	0.42	0.46	0.44	0.45
Complete Link(SN)	0.4	0.51	0.69	0.72	0.76	0.78	**0.8**
Average(NF)	0.13	0.15	0.25	0.26	0.23	0.22	0.2
Average(WA)	0.13	0.19	0.25	0.26	0.24	0.22	0.21
Average(SWN)	0.27	0.32	0.37	0.44	0.43	0.4	0.41
Average(SN)	0.3	0.42	0.64	0.69	0.71	0.76	**0.8**
Ward's Method(NF)	0.15	0.17	0.17	0.2	0.22	0.24	0.26
Ward's Method(WA)	0.16	0.18	0.18	0.21	0.23	0.26	0.28
Ward's Method(SWN)	0.31	0.36	0.41	0.4	0.43	0.44	0.47
Ward's Method(SN)	0.54	0.6	0.69	0.75	0.8	0.82	**0.84**
Coeficiente de Silueta para el conjunto 7							
Complete Link(NF)	0.19	0.26	0.4	0.47	0.49	0.44	0.44
Complete Link(WA)	0.2	0.27	0.4	0.48	0.5	0.46	0.44
Complete Link(SWN)	0.23	0.3	0.42	0.53	0.54	0.53	0.54
Complete Link(SN)	0.4	0.53	0.73	0.78	0.8	0.84	**0.86**
Average(NF)	0.15	0.23	0.35	0.25	0.5	0.49	0.47
Average(WA)	0.13	0.23	0.36	0.46	0.5	0.49	0.47
Average(SWN)	0.12	0.23	0.4	0.49	0.56	0.55	0.54
Average(SN)	0.07	0.2	0.59	0.78	0.81	0.81	**0.82**
Ward's Method(NF)	0.28	0.34	0.43	0.45	0.48	0.48	0.49
Ward's Method(WA)	0.28	0.35	0.43	0.46	0.48	0.49	0.49
Ward's Method(SWN)	0.34	0.42	0.48	0.53	0.55	0.57	0.58
Ward's Method(SN)	0.56	0.66	0.76	0.81	0.84	0.87	**0.89**
Coeficiente de Silueta para el conjunto 8							
Complete Link(NF)	0.22	0.3	0.4	0.48	0.49	0.47	0.46
Complete Link(WA)	0.23	0.31	0.41	0.48	0.49	0.49	0.47
Complete Link(SWN)	0.2	0.3	0.43	0.54	0.56	0.55	0.55
Complete Link(SN)	0.51	0.59	0.66	0.75	0.81	0.82	**0.83**
Average(NF)	0.14	0.22	0.4	0.45	0.49	0.48	0.47
Average(WA)	0.16	0.24	0.37	0.45	0.5	0.49	0.48
Average(SWN)	0.19	0.27	0.39	0.47	0.56	0.55	0.55
Average(SN)	0.06	0.15	0.47	0.71	**0.82**	0.81	0.81
Ward's Method(NF)	0.29	0.36	0.44	0.47	0.47	0.48	0.48
Ward's Method(WA)	0.3	0.37	0.44	0.47	0.49	0.49	0.49
Ward's Method(SWN)	0.33	0.42	0.5	0.54	0.56	0.57	0.57
Ward's Method(SN)	0.58	0.67	0.77	0.81	0.84	0.86	**0.88**

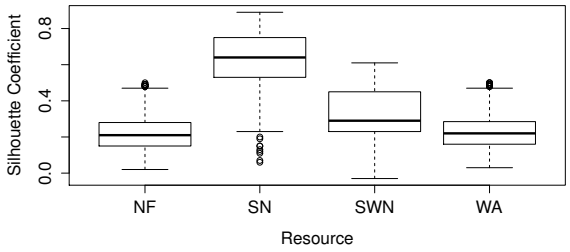


FIGURA 4.8: Boxplot entre el Coeficiente de Silueta y los recursos utilizados

Por otra parte, Figura 4.10 muestra la gráfica para el Coeficiente de Silueta con respecto a la red social de la cual provienen los textos. Se realizó la prueba de Wilcoxon [Wilcoxon, 1945] y se puede concluir que existe una gran diferencia para las redes sociales utilizadas, pues Dreamcatchers brinda los mejores resultados. Esto se debe en gran medida a que en dicha red social los textos están en español donde la riqueza del vocabulario es mayor, y por otra parte los usuarios de esta red pertenecen a un contexto universitario, por lo que el dominio de conversación es más restringido. Debemos destacar que aunque el idioma en el cual estén escritos los textos influye en cierta medida en la detección de contextos, esto no constituye un factor decisivo para dicha tarea.

En la Figura 4.11 se muestra la media de los valores del Coeficiente de Silueta con respecto al método utilizado para la selección de los conjuntos (inducido o aleatorio). A partir de este análisis se puede concluir que cuando los términos con orientación sentimental no son descartados, existen diferencias significativas entre los conjuntos de datos seleccionados aleatoriamente y los inducidos Figura 4.11(a). Sin embargo, cuando dichos términos son eliminados no existen diferencias significativas entre los resultados Figura 4.11(b), demostrando así la validez de la propuesta del presente trabajo.

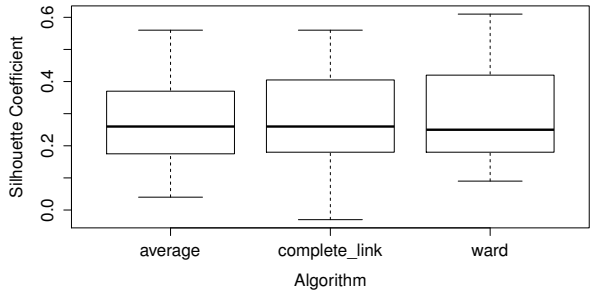


FIGURA 4.9: Boxplot entre el Coeficiente de Silueta y los algoritmos de agrupamiento

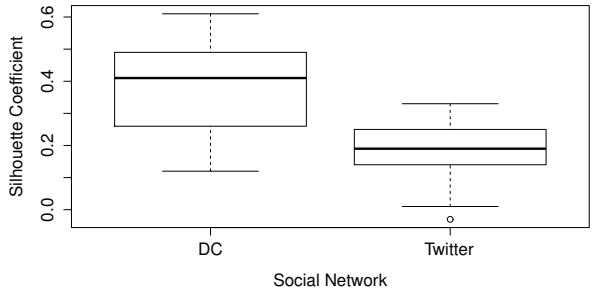
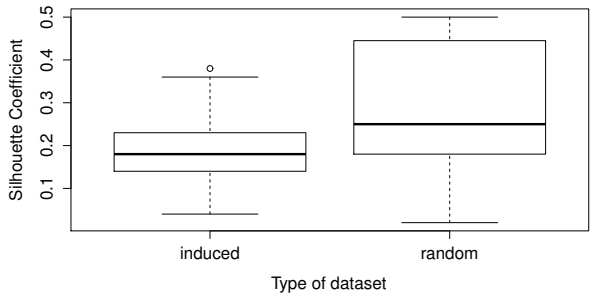
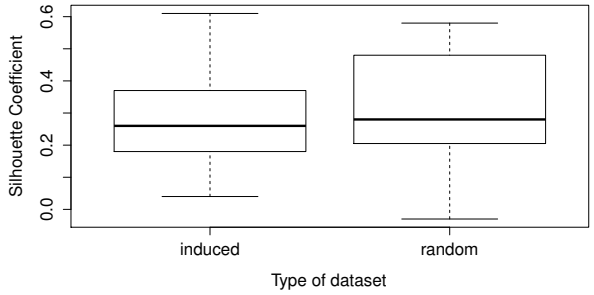


FIGURA 4.10: Boxplot entre el Coeficiente de Silueta y las redes sociales utilizadas



(a)



(b)

FIGURA 4.11: Valores del Coeficiente de Silueta con respecto al método de selección de los conjuntos de datos (inducido o aleatorio) (a) Valores de silueta con respecto al método de selección cuando los términos con orientación sentimental no son descartados (b) Valores de silueta con respecto al método de selección cuando los términos con orientación sentimental son descartados

A continuación estudiamos cómo se comporta nuestra propuesta para la detección de contextos cuando en lugar de eliminar todos los términos de sentimientos, sólo se eliminan los términos cuya polaridad supere un umbral determinado. Las Figuras 4.12-4.19 muestran los valores del Coeficiente de Silueta para los ocho conjuntos de datos. En cada caso se ha utilizado el método de agrupamiento jerárquico Complete Link, el recurso léxico SentiWordNet, se han realizado cortes en 17, 25, 40, 60, 80, 100 y 120 grupos y se han establecido siete umbrales de polaridad (0, 0.2, 0.3, 0.4, 0.5, 0.7 y 1), donde 0 implica eliminar todos los términos de sentimientos y 1 no eliminar ningún término. Como se puede apreciar para cada conjunto, a partir de la cantidad de 60 grupos que es cuando se estabilizan los valores del Coeficiente de Silueta, los mejores valores se obtienen para el umbral 0.

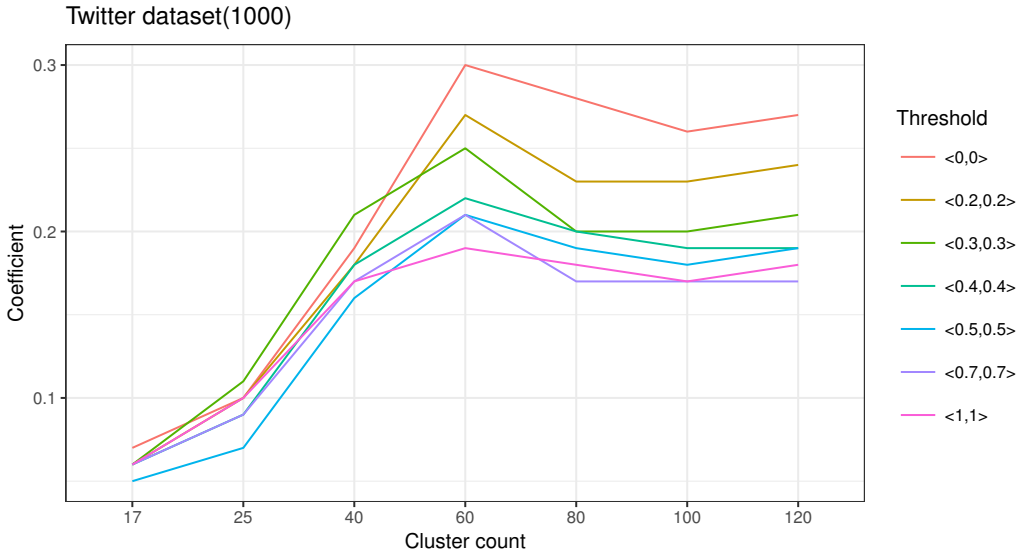


FIGURA 4.12: Coeficiente de Silueta para el Conjunto 1 utilizando el recurso SentiWordNet

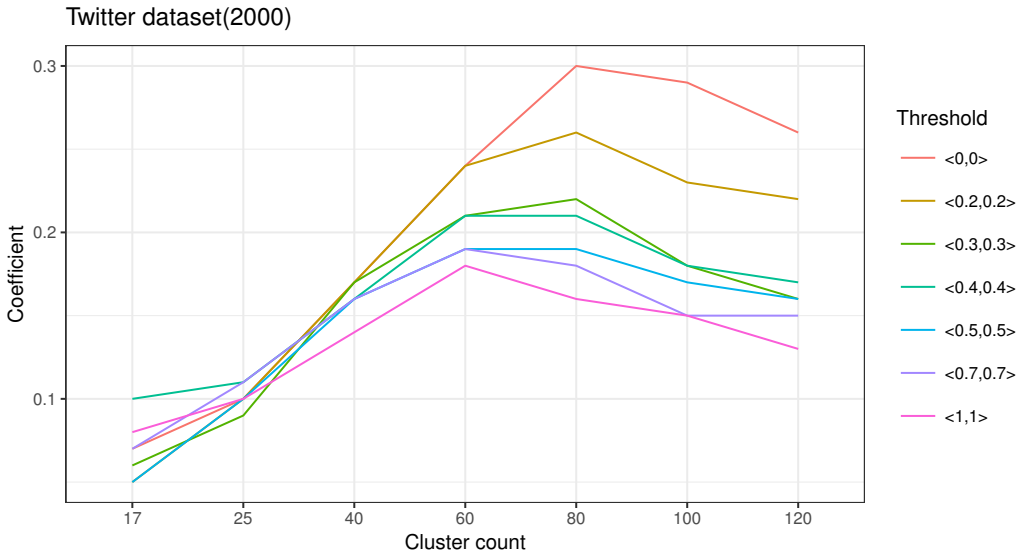


FIGURA 4.13: Coeficiente de Silueta para el Conjunto 2 utilizando el recurso SentiWordNet

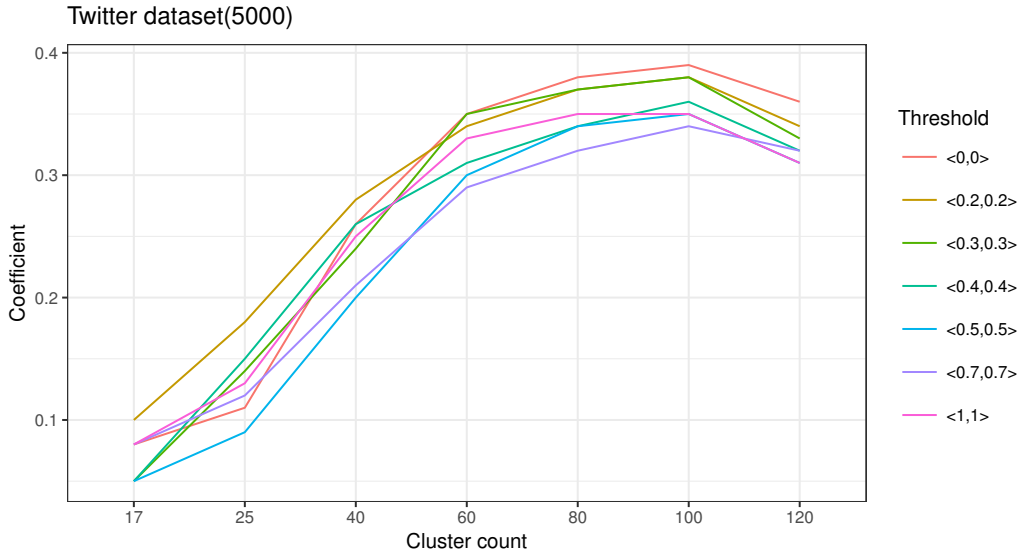


FIGURA 4.14: Coeficiente de Silueta para el Conjunto 3 utilizando el recurso SentiWordNet

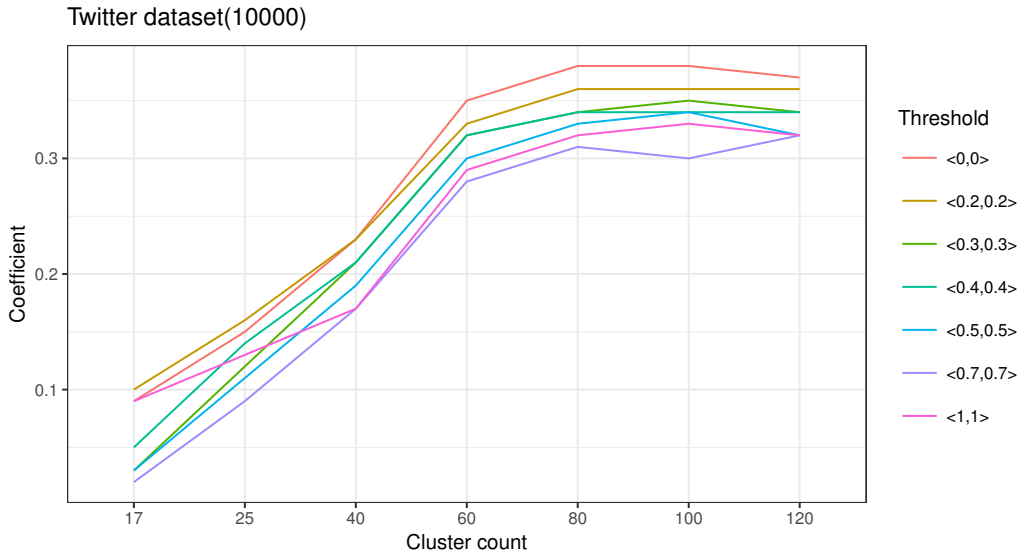


FIGURA 4.15: Coeficiente de Silueta para el Conjunto 4 utilizando el recurso SentiWordNet

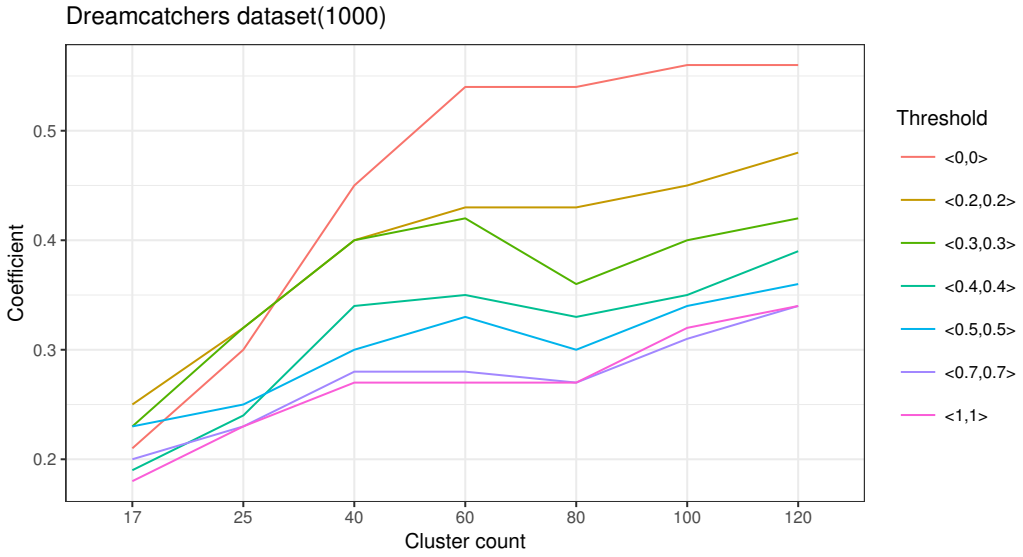


FIGURA 4.16: Coeficiente de Silueta para el Conjunto 5 utilizando el recurso SentiWordNet

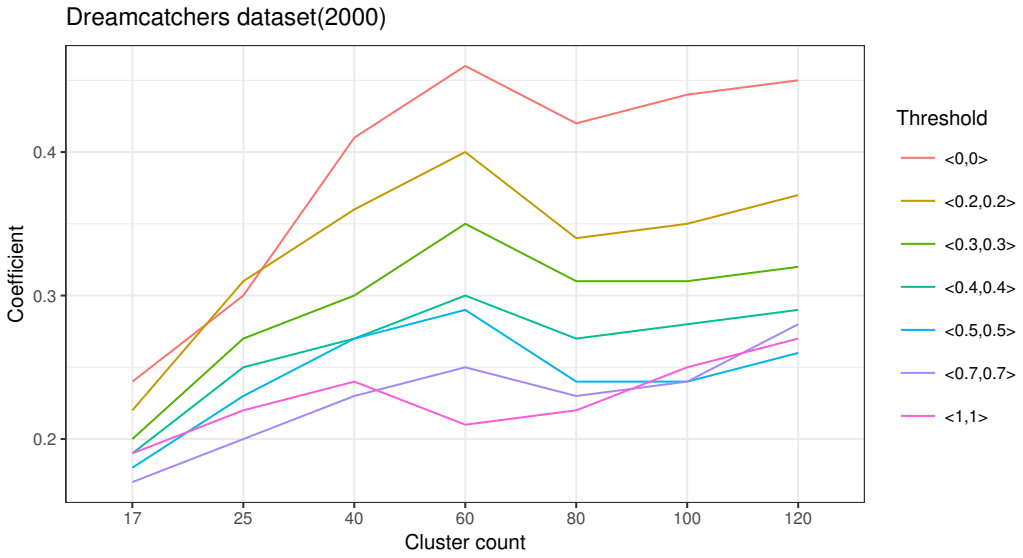


FIGURA 4.17: Coeficiente de Silueta para el Conjunto 6 utilizando el recurso SentiWordNet

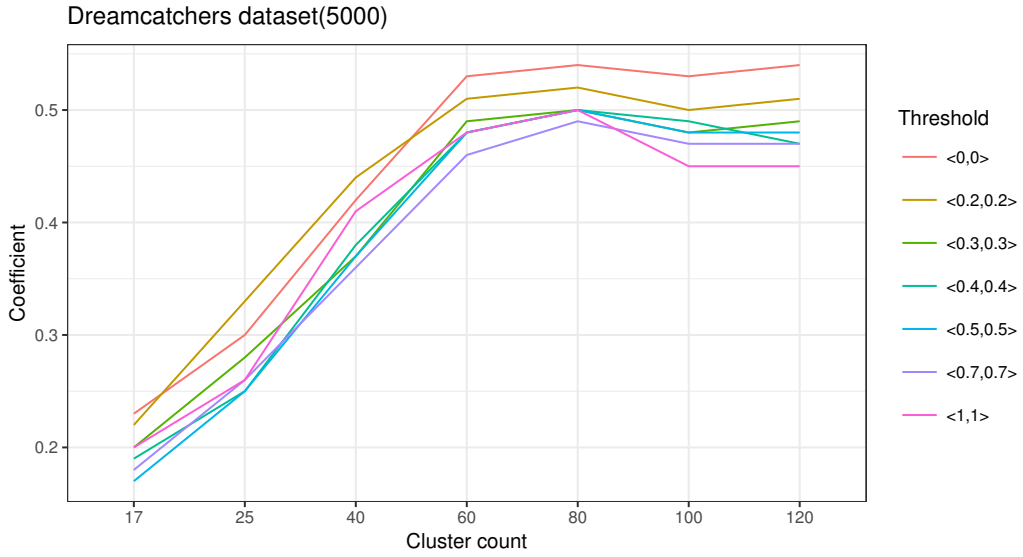


FIGURA 4.18: Coeficiente de Silueta para el Conjunto 7 utilizando el recurso SentiWordNet

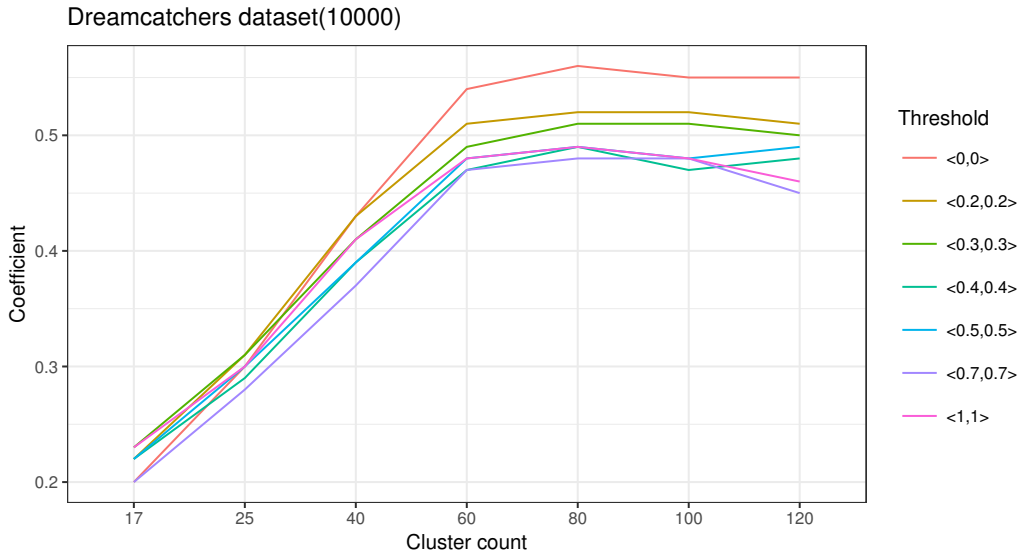


FIGURA 4.19: Coeficiente de Silueta para el Conjunto 8 utilizando el recurso SentiWordNet

Las Figuras 4.20-4.27, relacionan el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de aplicar el filtro que elimina los términos de sentimientos con respecto a los distintos umbrales seleccionados para cada conjunto de datos. En dichas gráficas se puede ver que aunque el Coeficiente de Silueta siempre alcanza el mayor valor cuando el umbral es 0, para este mismo valor la cantidad de documentos que permanecen luego de aplicar el filtro que elimina los términos de sentimientos siempre es la menor. Es por ello que resulta útil llegar a un consenso (punto de equilibrio) entre ambos parámetros con vista a mejorar el proceso

de detección de contextos. Dicho punto sería obviamente la intersección de ambas líneas.

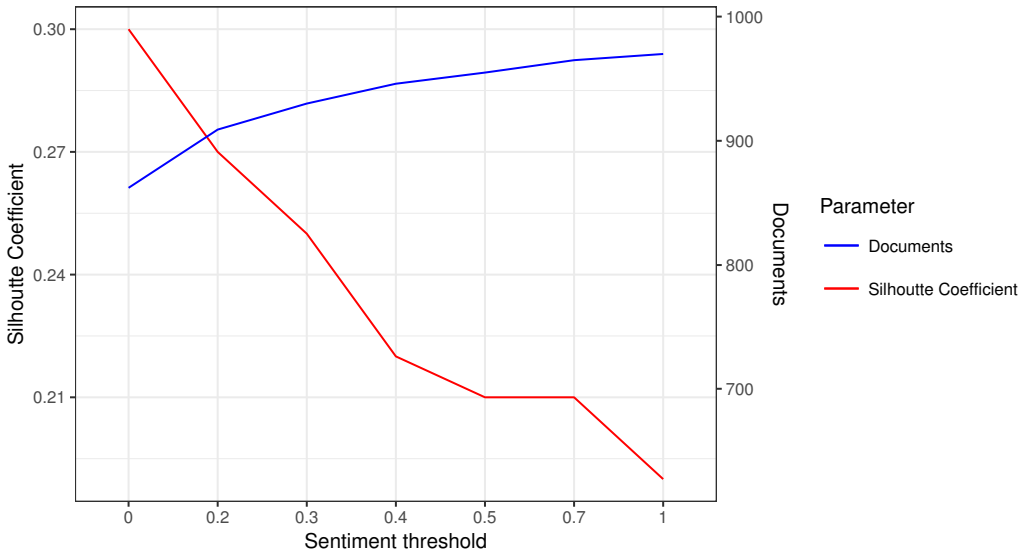


FIGURA 4.20: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 1)

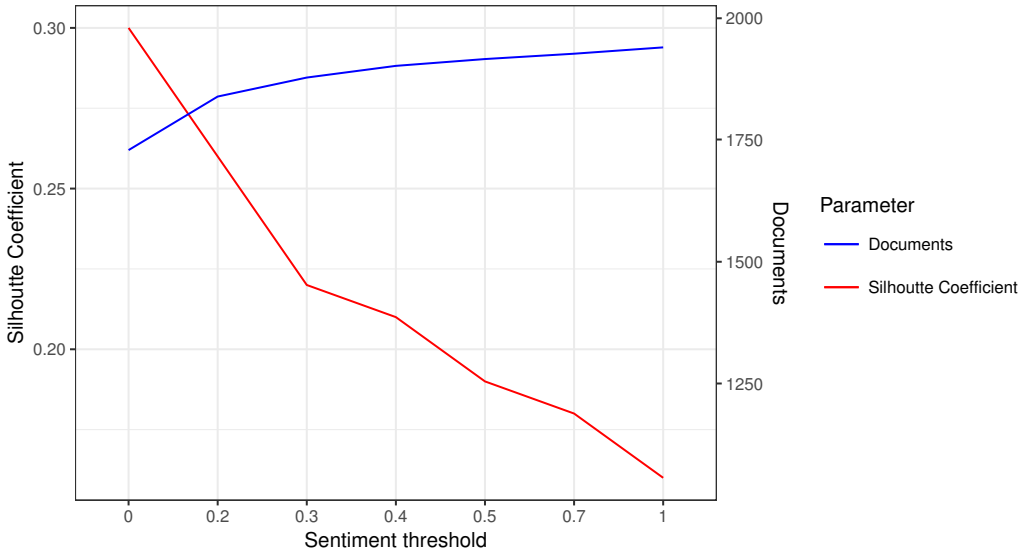


FIGURA 4.21: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 2)

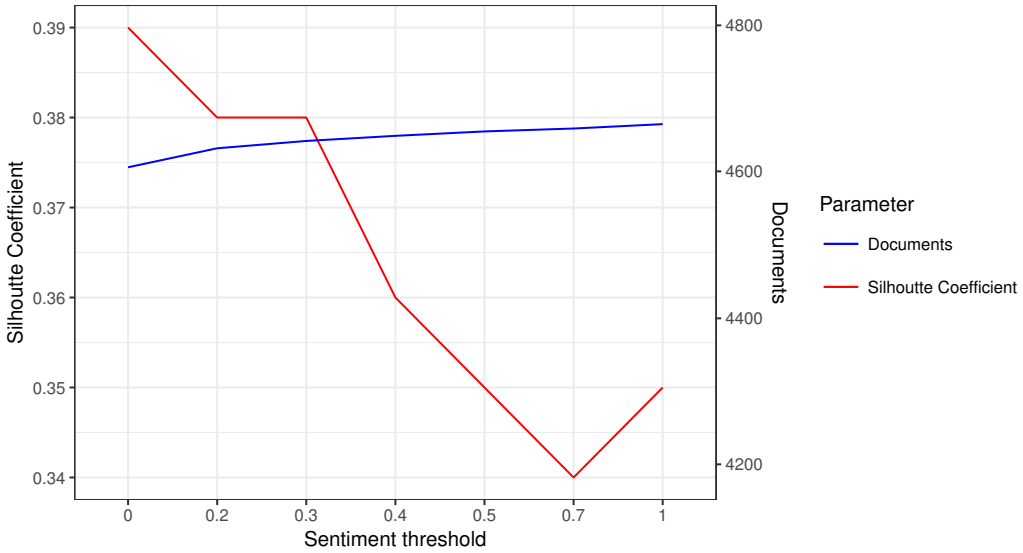


FIGURA 4.22: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 3)

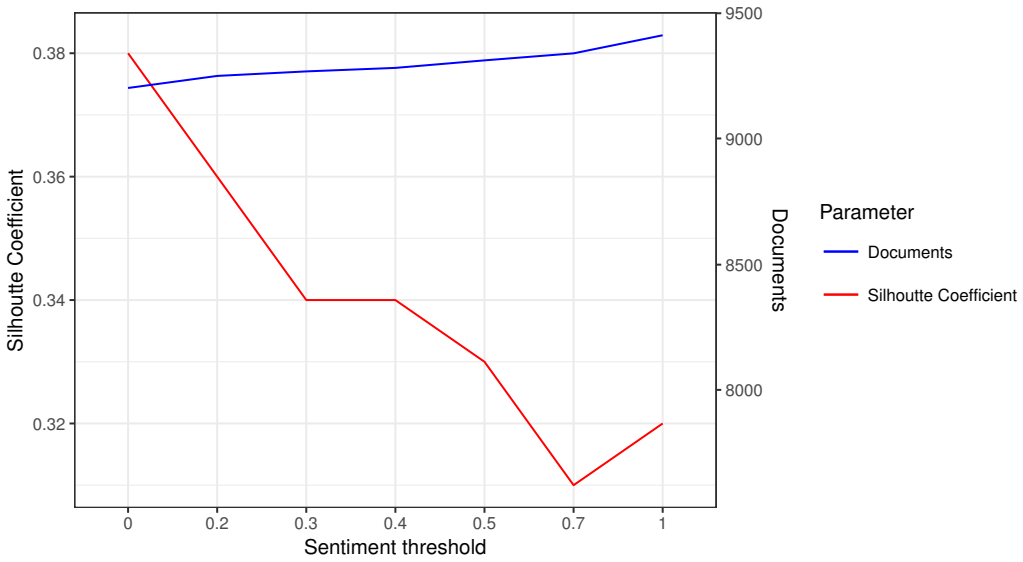


FIGURA 4.23: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 4)

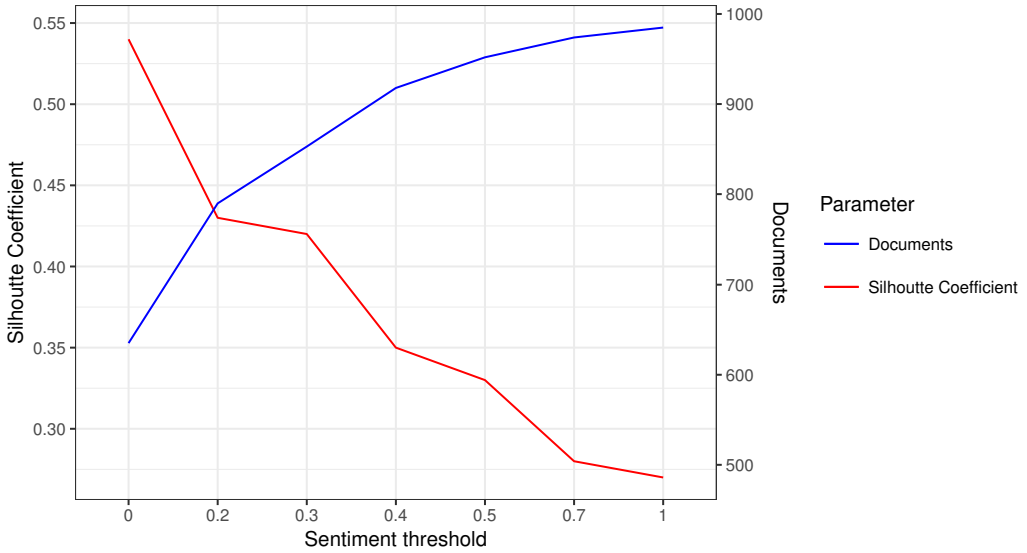


FIGURA 4.24: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 5)

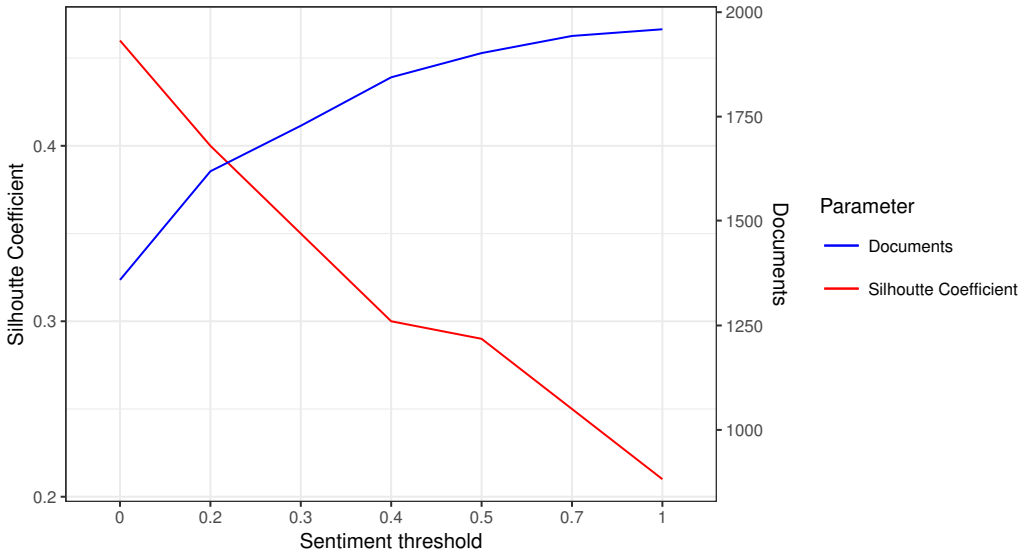


FIGURA 4.25: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 6)

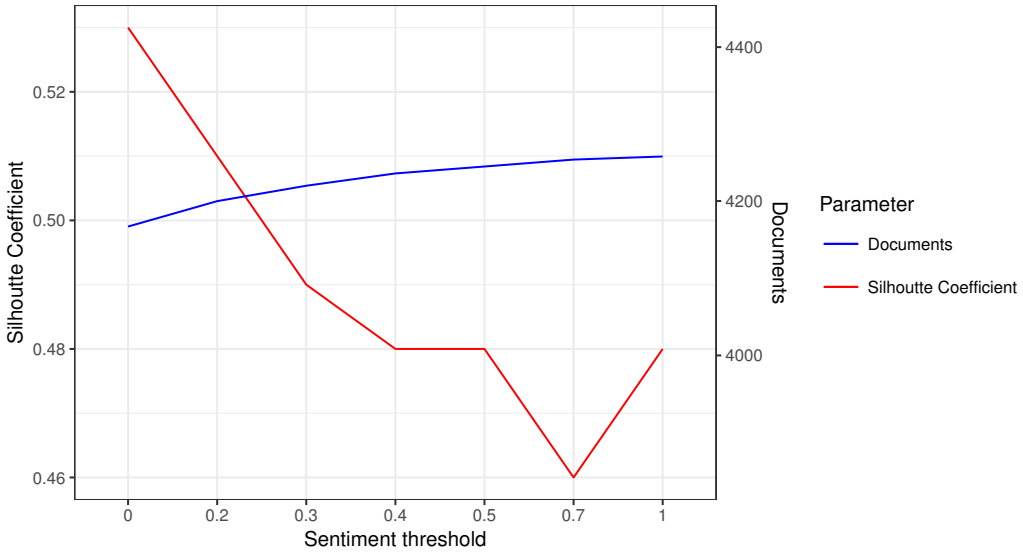


FIGURA 4.26: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 7)

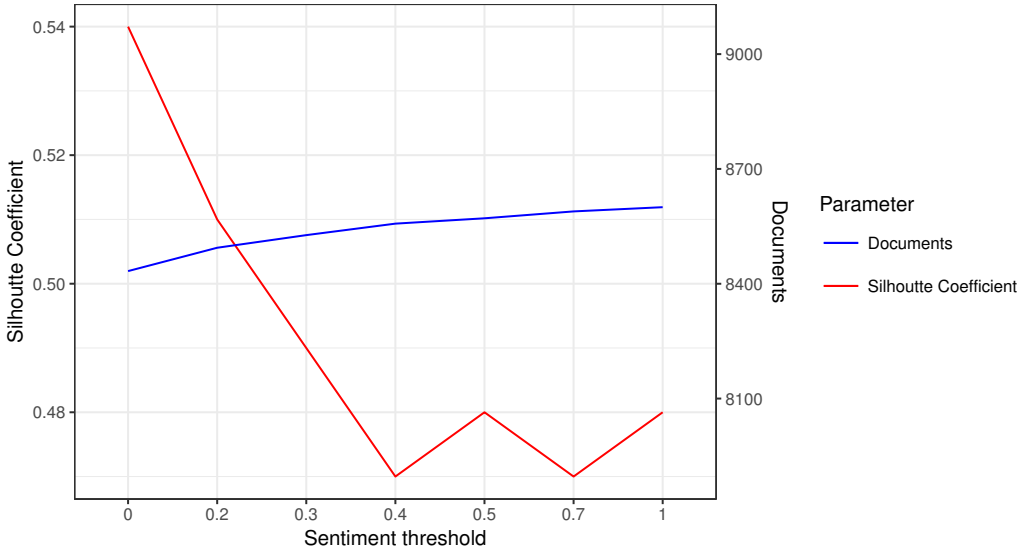


FIGURA 4.27: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SentiWordNet con los diferentes umbrales (Conjunto 8)

Al igual que en los ejemplos anteriores, las Figuras 4.28-4.35 muestran los valores del Coeficiente de Silueta para los ocho conjuntos de datos. En este caso se han utilizado los mismos parámetros excepto el recurso léxico para eliminar los términos de sentimientos, ya que se ha utilizado SenticNet 3. De igual forma para cada conjunto, a partir de la cantidad de 60 grupos que es cuando se estabilizan los valores del Coeficiente de Silueta, los mejores valores se obtienen para el umbral 0.

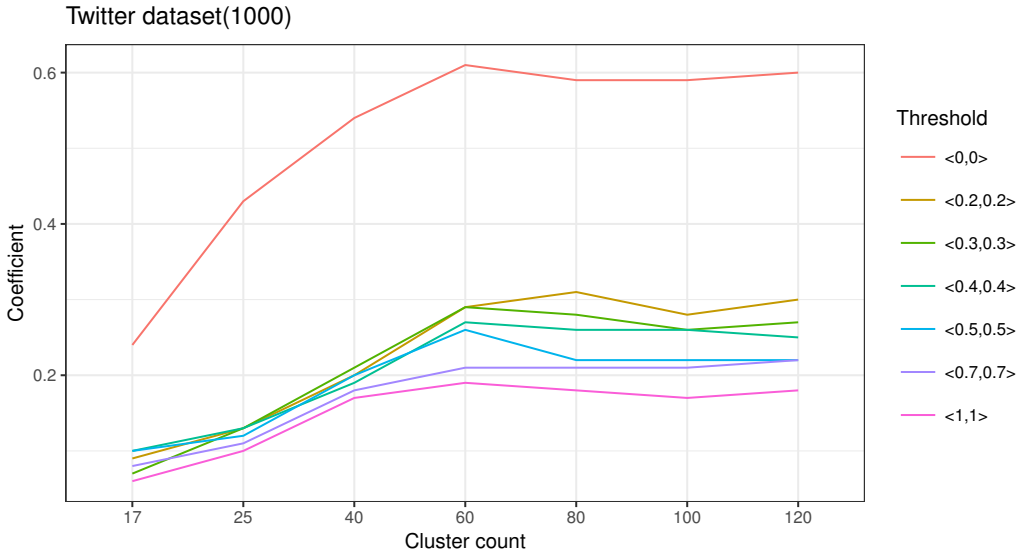


FIGURA 4.28: Coeficiente de Silueta para el Conjunto 1 utilizando el recurso SenticNet

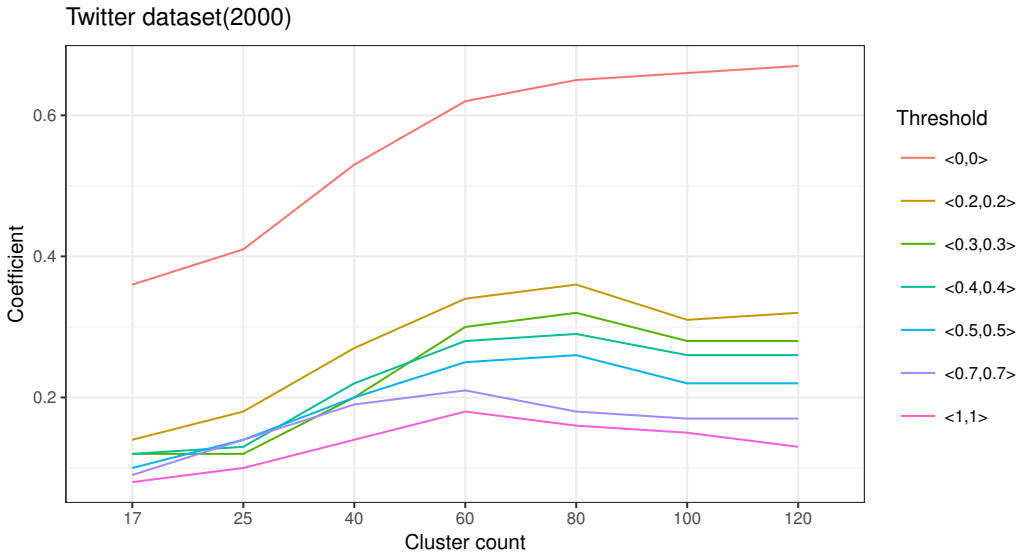


FIGURA 4.29: Coeficiente de Silueta para el Conjunto 2 utilizando el recurso SenticNet

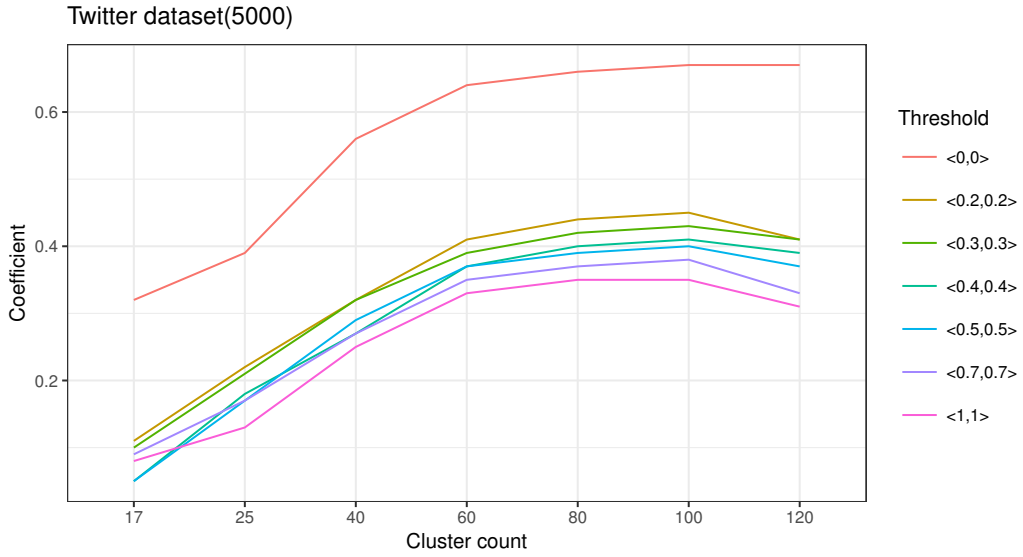


FIGURA 4.30: Coeficiente de Silueta para el Conjunto 3 utilizando el recurso SenticNet

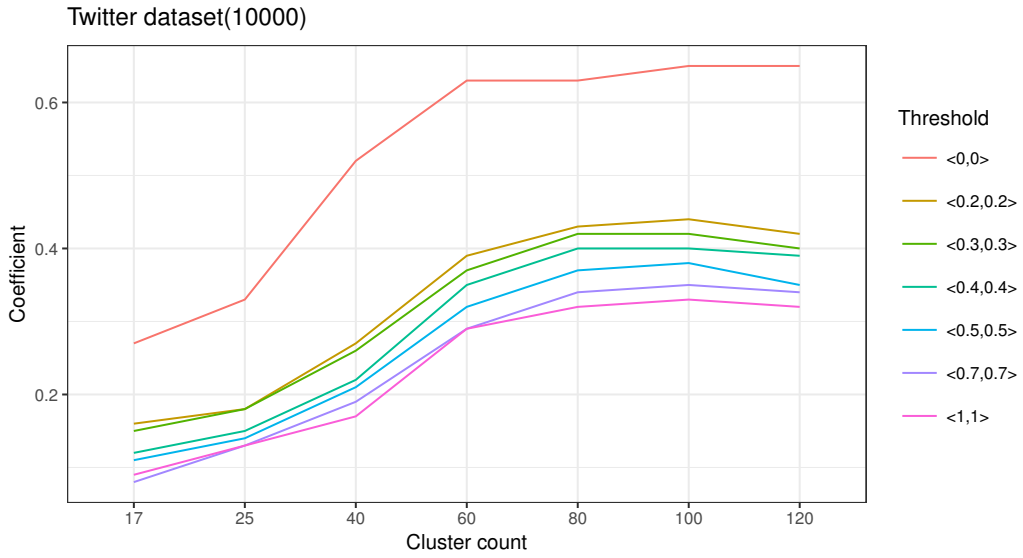


FIGURA 4.31: Coeficiente de Silueta para el Conjunto 4 utilizando el recurso SenticNet

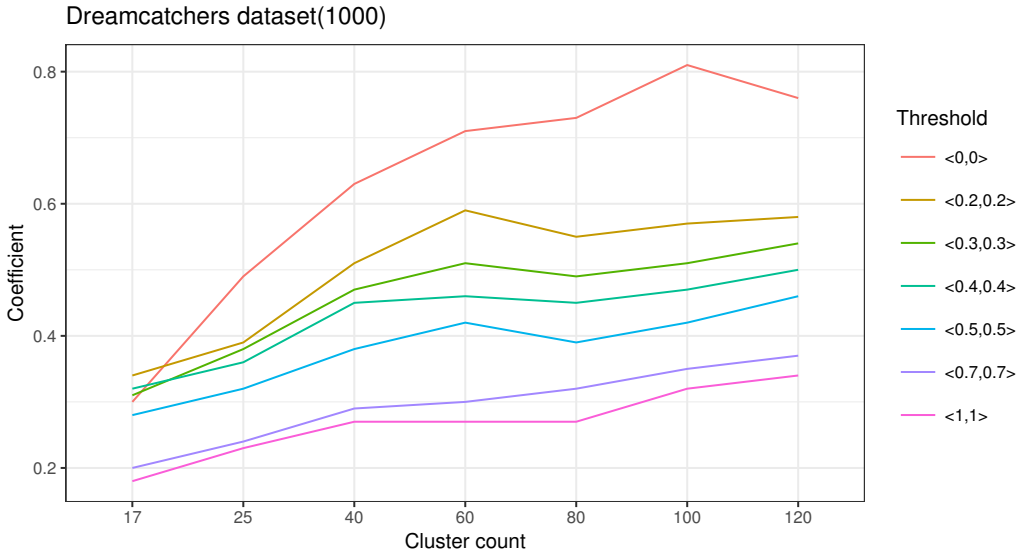


FIGURA 4.32: Coeficiente de Silueta para el Conjunto 5 utilizando el recurso SenticNet

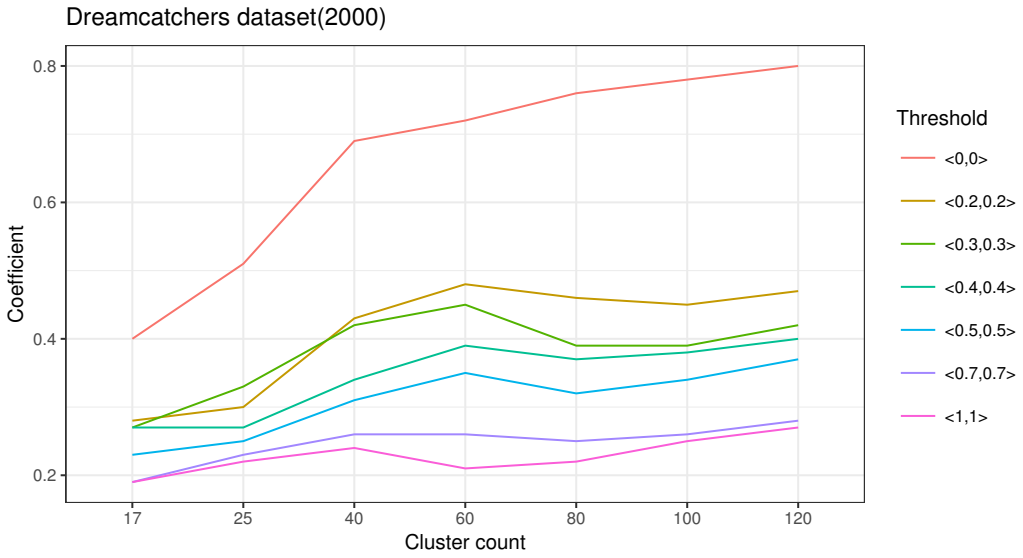


FIGURA 4.33: Coeficiente de Silueta para el Conjunto 6 utilizando el recurso SenticNet

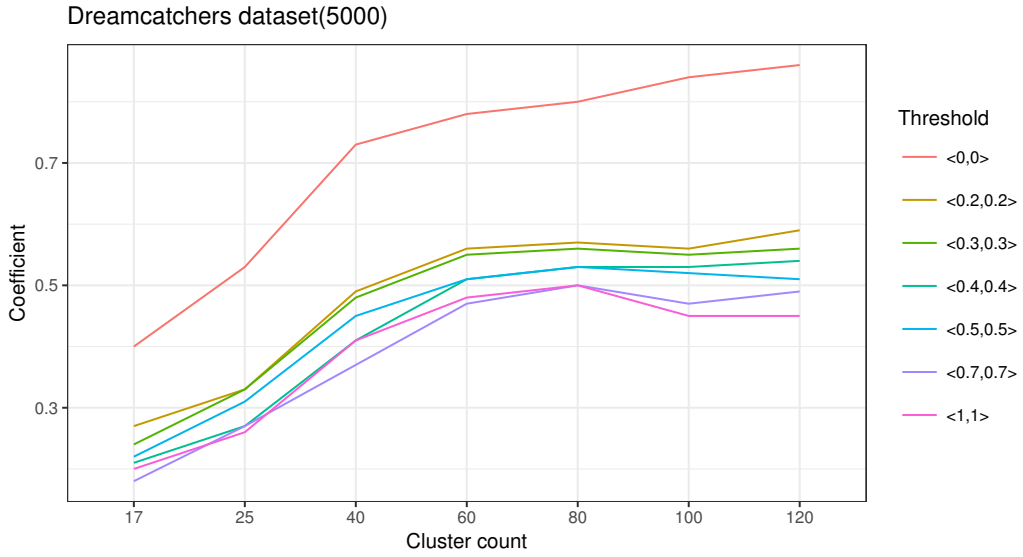


FIGURA 4.34: Coeficiente de Silueta para el Conjunto 7 utilizando el recurso SenticNet

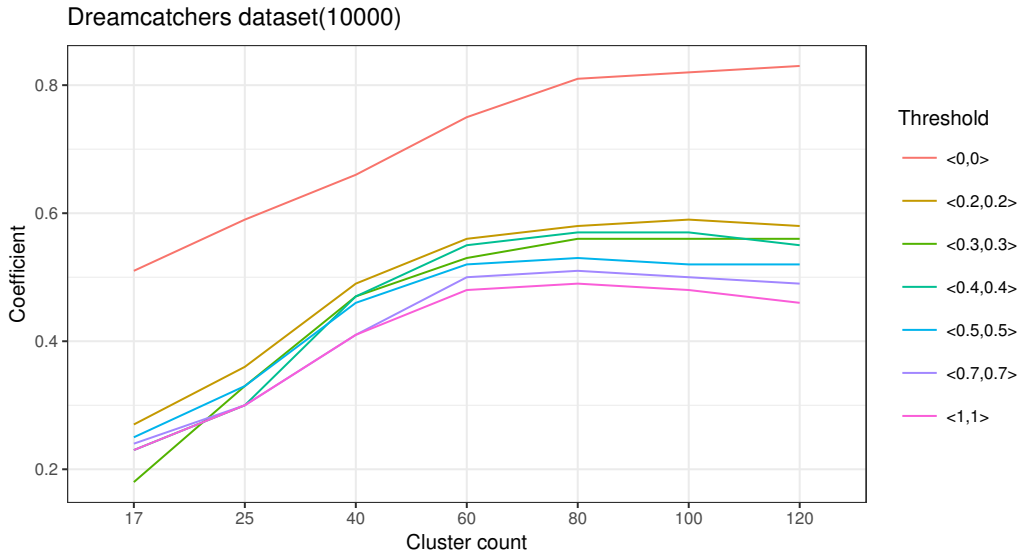


FIGURA 4.35: Coeficiente de Silueta para el Conjunto 8 utilizando el recurso SenticNet

Por su parte las Figuras 4.36-4.43, relacionan el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de aplicar el filtro que elimina los términos de sentimientos con respecto a los distintos umbrales seleccionados para cada conjunto de datos y las conclusiones de este análisis sería el mismo que para el caso anterior.

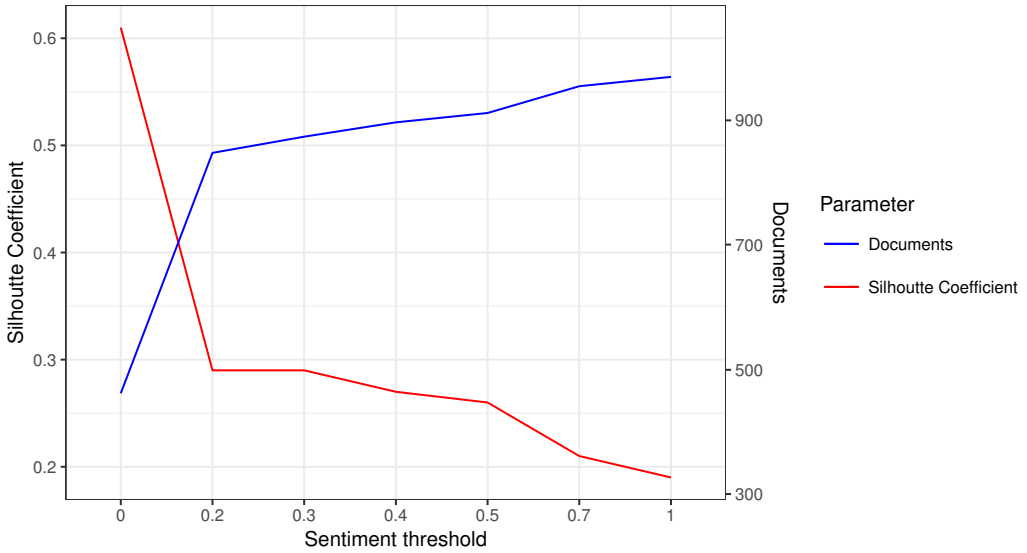


FIGURA 4.36: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 1)

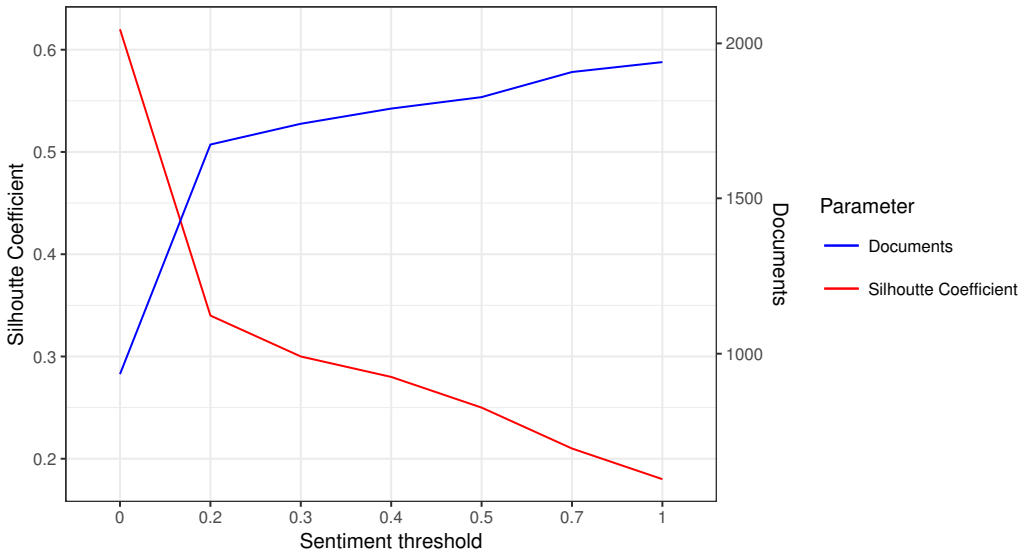


FIGURA 4.37: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 2)

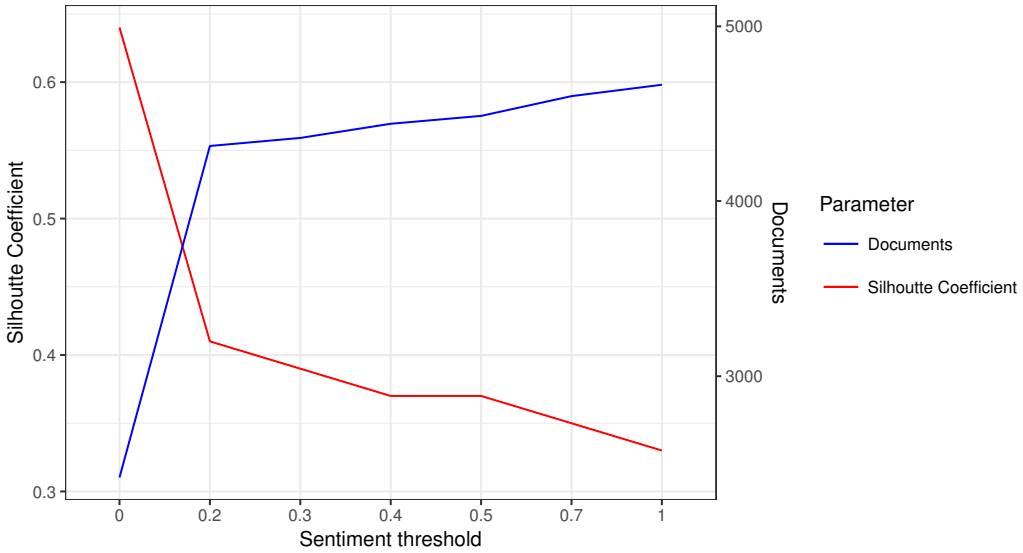


FIGURA 4.38: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 3)

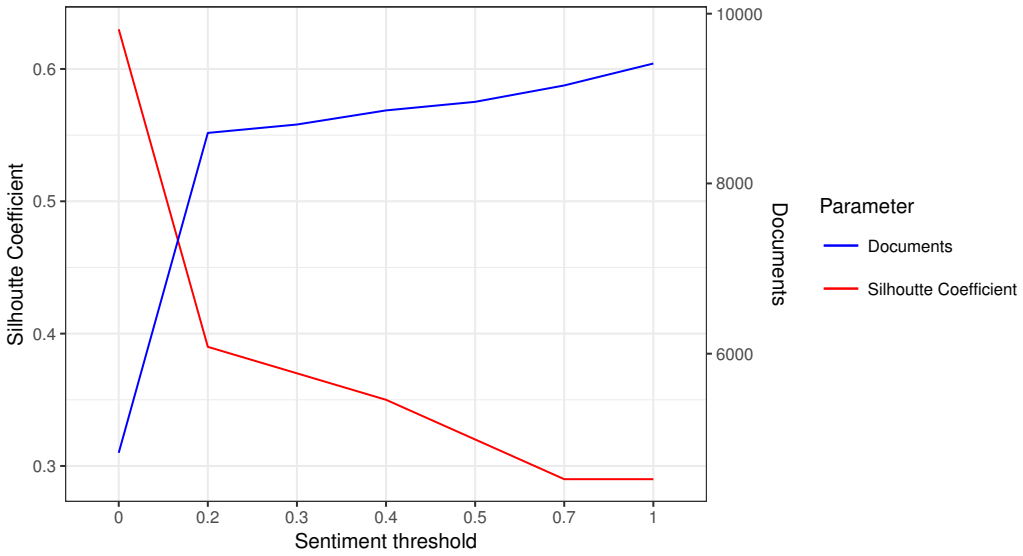


FIGURA 4.39: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 4)

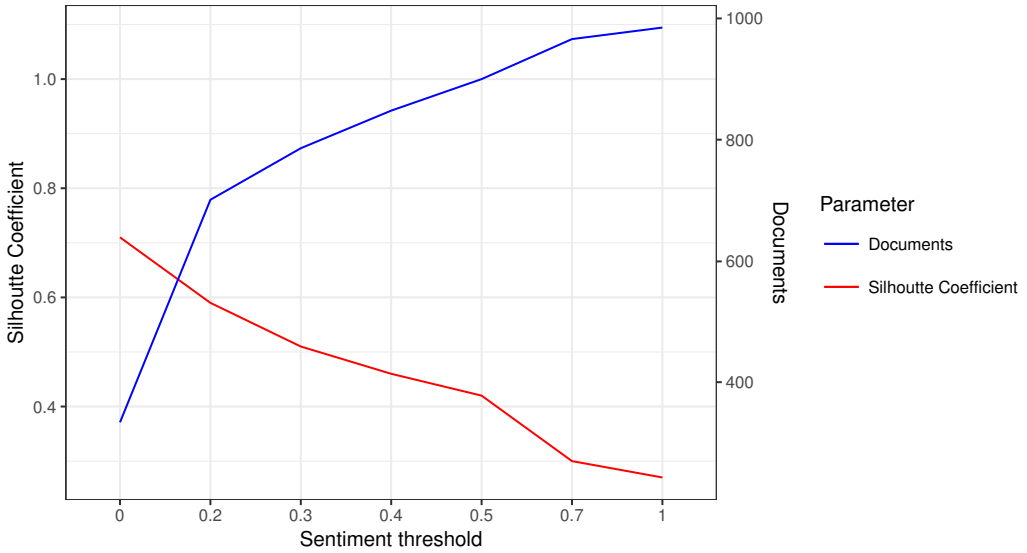


FIGURA 4.40: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 5)

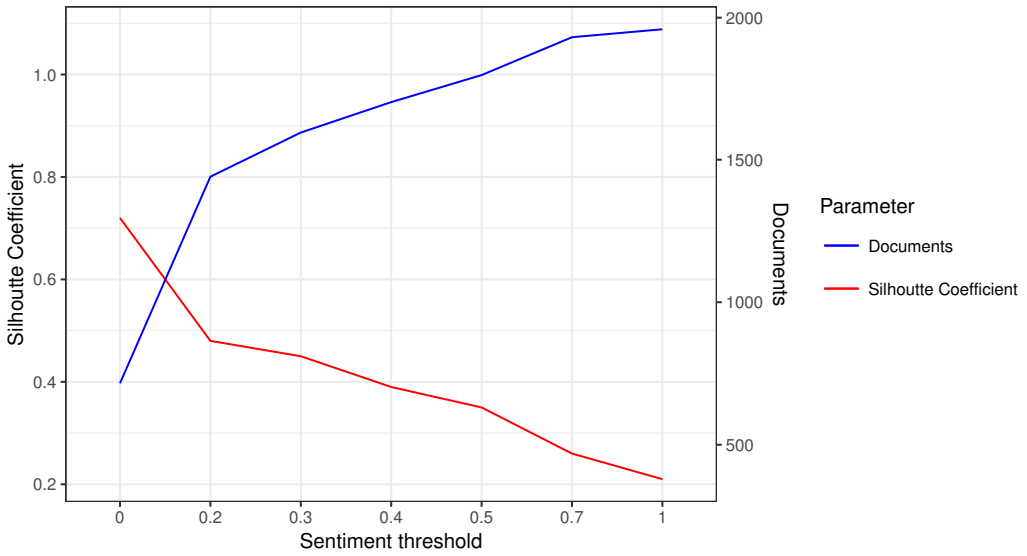


FIGURA 4.41: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 6)

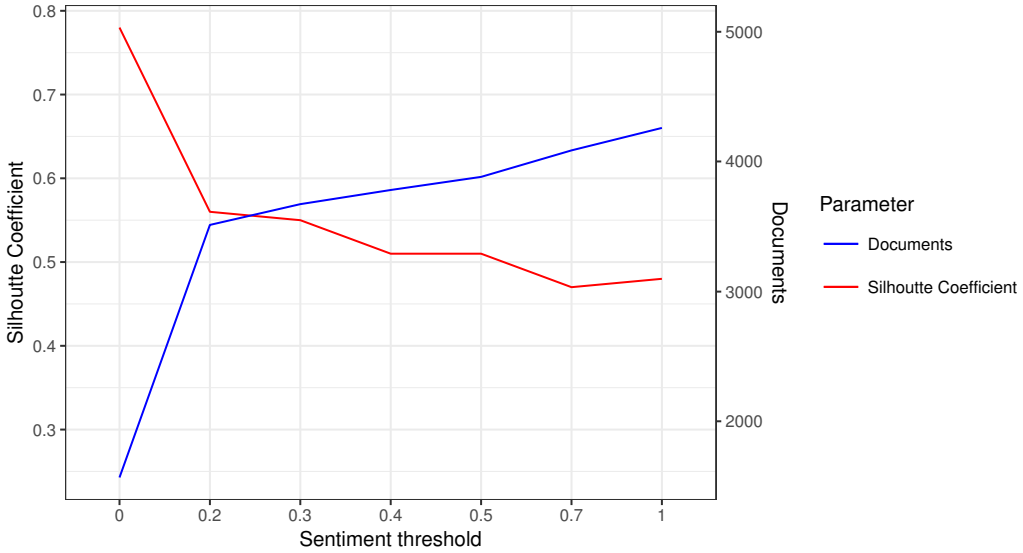


FIGURA 4.42: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 7)

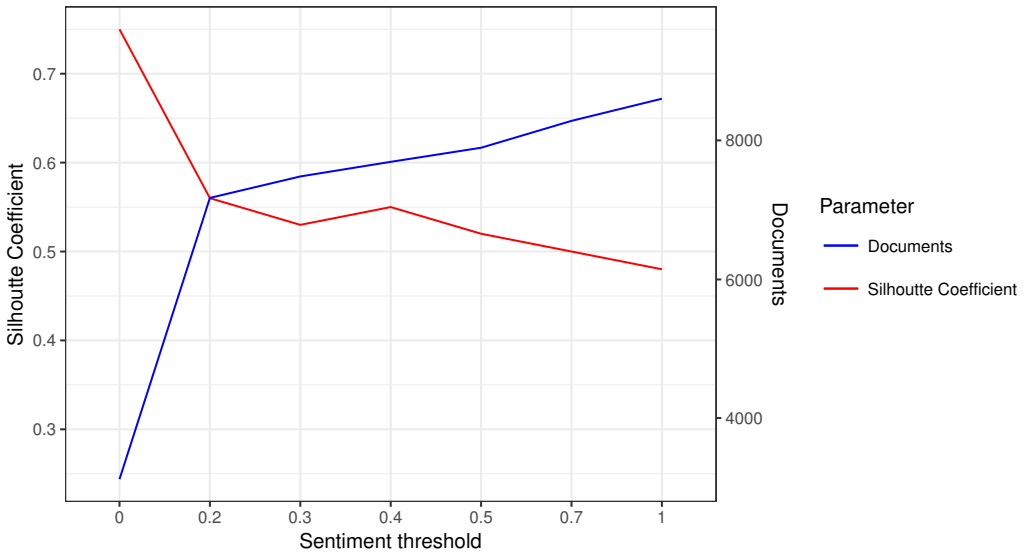


FIGURA 4.43: Relación entre el Coeficiente de Silueta y la cantidad de documentos que permanecen luego de descartar los términos de sentimientos con el recurso SenticNet con los diferentes umbrales (Conjunto 8)

Por último, en las tablas 4.6 y 4.7 se muestra un resumen con los resultados más relevantes para los ocho conjuntos de datos. En cada caso se ha reflejado el valor del Coeficiente de Silueta y la cantidad de documentos que permanecen luego de filtrar los términos de sentimientos para los α-cortes 0, 1 y el α-corte que proporciona un mejor balance entre los dos parámetros analizados (coeficiente y documentos).

TABLA 4.6: Coeficiente de Silueta y cantidad de documentos que permanecen para los conjuntos del 1-4

	SentiWordNet		**SenticNet**	
α-cortes	**Coeficiente**	**Documentos**	**Coeficiente**	**Documentos**
Conjunto 1				
1	0.19	970	0.19	970
0.2	0.25	930	0.29	874
0	0.3	862	0.61	462
Conjunto 2				
1	0.16	1940	0.16	1940
0.2	0.26	1878	0.34	1741
0	0.3	1729	0.62	933
Conjunto 3				
1	0.35	4665	0.35	4665
0.2	0.38	4632	0.41	4315
0	0.39	4606	0.64	2420
Conjunto 4				
1	0.32	9412	0.32	9412
0.2	0.36	9250	0.39	8595
0	0.38	9202	0.63	4829

TABLA 4.7: Coeficiente de Silueta y cantidad de documentos que permanecen para los conjuntos del 5-8

	SentiWordNet		SenticNet	
α-cortes	Coeficiente	Documentos	Coeficiente	Documentos
Conjunto 5				
1	0.27	985	0.27	985
0.3	0.42	853	0.51	786
0	0.54	635	0.71	334
Conjunto 6				
1	0.21	1959	0.21	1959
0.3	0.35	1728	0.45	1703
0	0.46	1359	0.72	715
Conjunto 7				
1	0.48	4258	0.48	4258
0.3	0.49	4220	0.55	3671
0	0.53	4167	0.78	1566
Conjunto 8				
1	0.48	8600	0.48	8600
0.2	0.51	8527	0.56	7481
0	0.54	8433	0.75	3122

4.4. Conclusiones

En este capítulo se desarrolla un nuevo enfoque para la detección automática de contextos en textos de redes sociales. Para ello se ha incorporado un filtro durante el preprocesamiento semántico de los textos, permitiendo detectar y eliminar los términos que presenten una orientación sentimental, ya que dichos términos no constituyen información relevante para la detección de contextos. Para ello se utilizan los recursos SentiWordNet 3.0, SenticNet 3 y WordNet Affect.

Los experimentos realizados tanto con Twitter como con Dreamcatchers, permiten mostrar la viabilidad de la propuesta. Con el fin de establecer una comparación, se ha experimentado sin aplicar el filtro para eliminar los sentimientos y aplicándolo.

Primero para cada conjunto de datos, se eliminaron todos los términos de sentimientos utilizando cada unos de los recursos antes mencionados. En cada caso se calculó el Coeficiente de Silueta realizando cortes en las cantidades de grupos ya mencionadas. Los resultados alcanzados cuando se aplica el filtro mejoran considerablemente en comparación cuando no se aplica, destacando los conjuntos de datos 1, 2, 5 y 6 que fueron seleccionados de forma intencionada. El recurso con el que se obtuvo un mejor rendimiento fue SenticNet 3.0, mientras que el recurso WordNet Affect brinda los peores resultados.

Finalmente, para cada conjunto de datos, se realizó un estudio con la finalidad establecer un consenso entre el valor del Coeficiente de Silueta y la cantidad de textos a analizar para la detección de contextos, ya que al eliminar todos los términos con orientación sentimental, se pueden descartar una gran cantidad de documentos los cuales no formarán parte del proceso de detección de contextos. Para ello se establecieron siete umbrales diferentes, y se llegó a la conclusión que para obtener un equilibrio entre los dos parámetros analizados, el umbral varía en el rango [0.2,0.3] en dependencia del caso en cuestión.

Capítulo 5

Construcción de una dimensión contextual para el análisis multidimensional de textos de redes sociales

En el presente capítulo se propone una solución original a la dificultad que existe hoy día para lograr la aplicación del análisis multidimensional sobre datos heterogéneos que integren datos textuales de redes sociales. Dicha solución se basa en detectar previamente los contextos que se abordan en los textos y de esta forma se tienen los textos segmentados y organizados por dichos contextos.

En el Capítulo 2 se hace referencia a un grupo de trabajos orientados al uso de las tecnologías DW y OLAP para el estudio de los datos textuales presentes en las distintas redes sociales, especialmente por la invaluable información que aportan sobre un determinado producto, servicio, etc. La gran mayoría de estos trabajos, realizan tareas tales como Recuperación de Información, Análisis de Sentimientos, Sistemas de Recomendación, etc., sin tener en cuenta previamente el contexto al cual pertenecen dichos datos textuales.

La base de nuestra propuesta lo conforman las estructuras de almacenamiento y de consulta presentadas en los trabajos previos [Martin-Bautista et al., 2013, Martin-Bautista et al., 2015]. Los autores definen e implementan una jerarquía textual, formalmente conocida como dimensión-AP, la cual permite tratar los datos no estructurados de igual manera que las dimensiones clásicas, razón por la cual dicha estructura se ajusta perfectamente para nuestros intereses. Para la implementación se ha utilizado el servidor OLAP de libre disposición Wonder OLAP Server 3.0, el

cual ha sido desarrollo por nuestro grupo de investigación [Avila et al., 2011].

Las contribuciones de este capítulo son las siguientes:

- Una nueva metodología para la creación de un modelo multidimensional con soporte para dimensiones textuales organizadas por contextos (los cuales a su vez son vistos como un conjunto de tópicos). Los detalles para la detección de contextos en textos de redes sociales se pueden encontrar en el Capítulo 3. Esta dimensión, denominada dimensión contextual, es creada automáticamente, e implementada en un sistema OLAP (Wonder 3.0). La dimensión contextual es extraída a partir de textos de redes sociales y está formada por dos componentes: una jerarquía de contextos compuesta por grupos de temas discutidos en los textos, y para cada nivel y contexto de esta jerarquía, una jerarquía de dominio que incluye los principales tópicos relacionados con dicho contexto.
- La dimensión contextual permitirá a los decisores analizar los datos de redes sociales a partir de un contexto determinado y por los principales tópicos relacionados con éste. La originalidad de nuestra propuesta radica en el hecho de que los responsables de la toma de decisiones no necesitan conocer de antemano los contextos para poder consultar los datos. En este análisis, es posible combinar información textual con otros atributos tradicionales, ej. hora del día y día de la semana cuando se publicaros los textos.
- El proceso de integración de la nueva dimensión contextual es completamente automático e independiente del idioma en el que se encuentren los textos. Esto permite el análisis de datos textuales de redes sociales mediante agregaciones que involucran dimensiones convencionales y no convencionales en datos heterogéneos.

5.1. Estructura de la dimensión contextual

Como se mencionó anteriormente, el principal aporte del presente capítulo consiste en crear e integrar de forma automática una nueva dimensión que recoja los principales contextos y tópicos tratados en textos de redes sociales. Esta dimensión (dimensión contextual) refleja o establece los tópicos presentes en los distintos contextos detectados en los textos mencionados anteriormente. De esta forma, al crear

un modelo multidimensional que brinde soporte a esta dimensión contextual, se dará la posibilidad de tratarla conjuntamente con las dimensiones tradicionales y de igual manera, permitiendo así realizar OLAP sobre datos de redes sociales a partir de los principales tópicos abordados en un contexto determinado.

En los apartados 5.1.1 y 5.1.2, se explica detalladamente la estructura de la dimensión contextual y se introducen las definiciones principales, que hemos establecido en trabajos previos y que serán la base para su construcción [Martin-Bautista et al., 2008, Martín-Bautista et al., 2010]. Como se puede observar en la Figura 5.1, las dos componentes por la que está formada la dimensión contextual son la jerarquía de contextos y la jerarquía de dominio o de consulta.

La primera no es más que la jerarquía de grupos etiquetados que se obtiene de la aplicación de algoritmos de agrupamiento jerárquico Capítulo 3. Para cada nivel de ésta y por cada grupo perteneciente a dicho nivel, se obtiene una Jerarquía de consulta, con los tópicos más relevantes que son abordados en dicho contexto y que serán la base de las búsquedas y operaciones OLAP que se definen en nuestro modelo multidimensional. Finalmente en la Subsección 5.1.3 se expone el proceso de integración de la dimensión contextual en el modelo multidimensional.

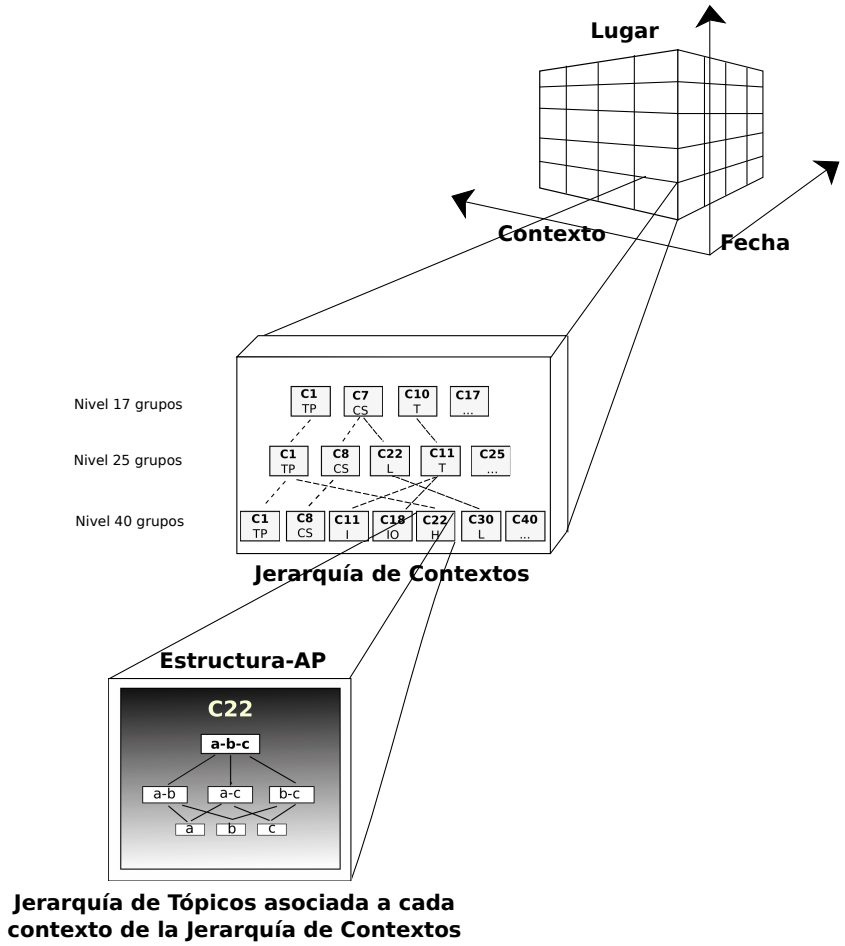


FIGURA 5.1: Componentes de la dimensión contextual

5.1.1. Jerarquía de contextos

Una de las componentes que conforma la dimensión contextual es la jerarquía de contextos. Es importante indicar que el proceso de construcción tanto de la jerarquía de contextos, como de su jerarquía de consulta asociada a cada contexto en cada nivel, es completamente automático. Dicha jerarquía de contextos se obtiene como resaltado de la aplicación de un algoritmos de agrupamiento jerárquico (Complete Linkage, Single Linkage, Centroid Linkage, Average Linkage, Ward's Method, etc.) a los datos textuales previamente preprocesados sintáctica y semánticamente Capítulo 3. De esta forma se obtiene un conjunto de grupos los cuales representan los principales contextos detectados.

Como se puede apreciar en la Figura 5.1, la jerarquía de contextos presenta una estructura en forma de árbol, donde cada nivel es el resultado de aplicar cortes durante la ejecución de un algoritmo de agrupamiento jerárquico sobre los textos preprocesados para una cantidad de grupos determinada (17, 25, 40, 60, 80 100, 120,

etc.). Estos cortes se realizan teniendo en cuenta el Coeficiente de Silueta [Rousseeuw, 1987], el cual es una medida de bondad de agrupamiento no supervisado.

Cada nivel está compuesto por los grupos o principales contextos en los que fueron agrupados los textos analizados. Además se debe resaltar que las relaciones entre los distintos niveles, se debe a la propia naturaleza de los algoritmos de agrupamiento jerárquico, donde un grupo o contexto de un nivel superior puede o no, ser dividido en dos o más grupos o contextos en el nivel inmediato inferior.

La obtención de la jerarquía de contextos es un proceso automático complejo, el cual incluye varias etapas o fases necesarias para su correcto funcionamiento. Aunque a continuación se dan los principales elementos para entender dicho proceso, en el Capítulo 3 se pueden encontrar todos los detalles del mismo.

En la Figura 5.2 se muestra la jerarquía de contextos que se obtiene para un con conjunto de datos reales de Twitter, utilizando el algoritmo de agrupamiento jerárquico enlace completo (Complete Linkage) y la distancia del coseno. En este caso se han hecho cortes en 17, 25 y 40 grupos. Cada cuadro en la jerarquía corresponde a un contexto, el cual se describe mediante el conjunto de etiquetas que contiene.

La jerarquía de contextos es un componente de vital importancia en la construcción de la dimensión contextual, ya que a partir de ella se crea la jerarquía de consulta asociada a un grupo y un nivel determinado. Esta jerarquía de consulta es la que nos permite tratar a la nueva dimensión contextual de igual forma que una dimensión clásica, pues brinda soporte para realizar las operaciones del modelo multidimensional. A continuación se introducen las principales definiciones que permiten establecer dicha jerarquía de consulta.

5.1.2. Jerarquía de consulta

La otra componente de la dimensión contextual es la jerarquía de consulta. Para cada contexto de cada nivel de la jerarquía de contextos, se crea una jerarquía de consulta. Para ello, se procesan los datos pertenecientes a un contexto dado, y se crea una estructura reticular la cual será utilizada como forma intermedia de representación y permitirá realizar consultas que contengan los principales tópicos abordados en dicho contexto.

FIGURA 5.2: Jerarquía de contextos para un conjunto de textos de Twitter utilizando el método de enlace completo (Complete Link)

Se debe resaltar que en esta estructura sólo aparecerán los conjuntos de tópicos frecuentes, que cumplen con un soporte determinado. Dicha estructura la denominamos estructura-AP, y está compuesta por conjuntos-AP [Martin-Bautista et al., 2006]. A continuación se incluyen las definiciones y algunas propiedades de dichas estructuras subyacentes en los textos, que capturan la semántica que encierran los mismos. Un estudio más detallado de las propiedades de estas estructuras y su uso en el análisis textual se presenta en [Martin-Bautista et al., 2006].

Conjunto-AP

Hablando de manera informal, un conjunto-AP está formado por un conjunto de términos y el retículo que contiene todos los subconjuntos posibles de menor cardinalidad. Todos los conjuntos-AP deben cumplir las dos condiciones siguientes:

- la primera condición es que todos los conjuntos-AP deben cumplir la propiedad Apriori [Agrawal and Srikant, 1994], y
- la segunda asegura la existencia de conjunto único llamado Y de máxima cardinalidad que caracteriza el conjunto-AP.

La propiedad Apriori es una característica intrínseca del Algoritmo Apriori propuesto en [Agrawal and Srikant, 1994], la cual establece que cualquier subconjunto de elementos frecuentes también debe ser frecuente.

Ejemplo 3 *Sea* $X = \{ACCESS, INTERNET, TWITTER, ..., WEBSITE\}$ *y*
$\mathcal{R} = \{\{ACCESS\}, \{INTERNET\}, \{WEBSITE\}, \{ACCESS, INTERNET\}, \{ACCESS, WEBSITE\}, \{INTERNET, WEBSITE\}, \{ACCESS, INTERNET, WEBSITE\}\}$,
luego el conjunto generador de $\mathcal{R}$ *es* $Y = \{ACCESS, INTERNET, WEBSITE\}$

La Figura 5.3 muestra el retículo de inclusión del Ejemplo 3. El conjunto generador Y es la raíz del retículo, cada uno de los elementos de Y se encuentran en los nodos hojas, mientras en los nodos intermedios, se encuentran las diferentes combinaciones de los elementos de Y con cardinalidad dos.

Los conjuntos-AP definen un conjunto de operaciones que nos permiten establecer un mecanismo formal de consulta a partir de su representación [Martin-Bautista et al., 2008].

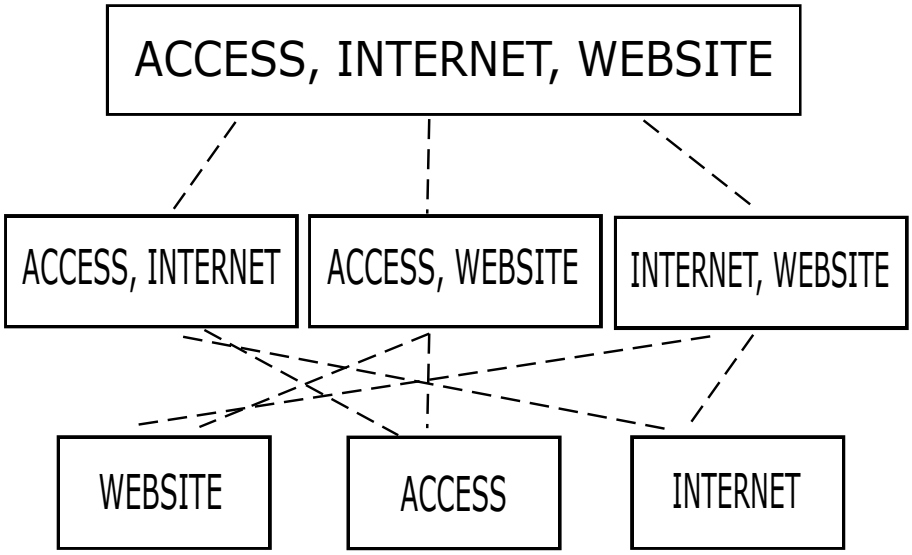


FIGURA 5.3: Ejemplo conjunto-AP para el contexto *Computer Science*

Estructura-AP

Una estructura-AP es un retículo de subconjuntos cuyos extremos superiores están formados por los conjuntos generadores. En la Figura 5.4 se muestra la estructura-AP $\mathcal{T} = g(\{PAPER, WEBSITE\}, \{ACCESS, INTERNET, WEBSITE\})$. Se debe resaltar que en este ejemplo los dos conjuntos generadores presentan un elemento en común, razón por la cual dicho elemento aparece como parte del retículo subyacente, concretamente el subconjunto $\mathcal{R} = g(\{WEBSITE\})$ de cardinalidad uno.

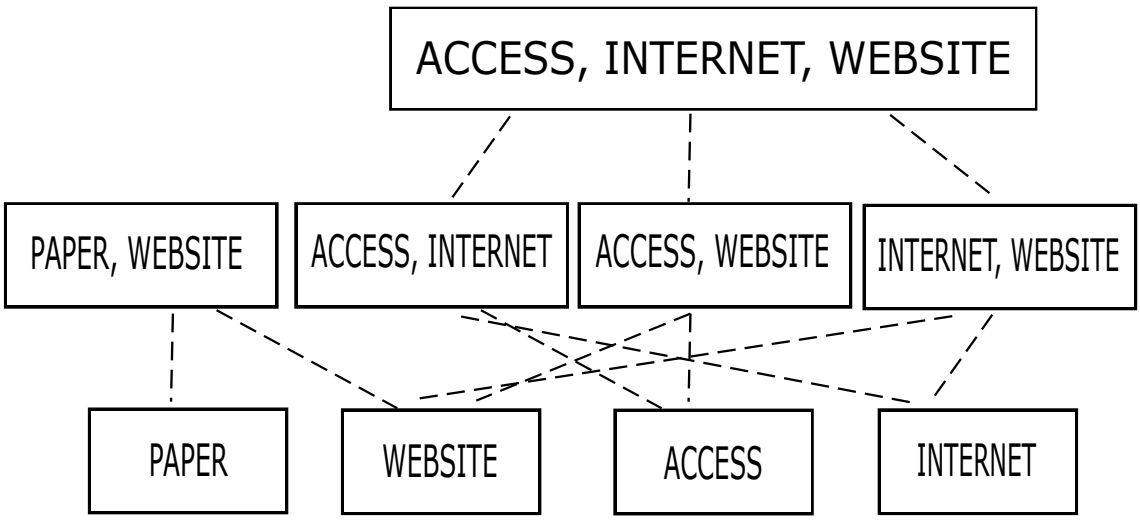


FIGURA 5.4: Ejemplo de estructura-AP para el contexto *Computer Science*

Si tenemos en cuenta que la estructura-AP se obtiene a partir de los términos más relevantes que aparecen en los textos procesados, podemos decir que dicha estructura contendrá la mayoría de los términos relevantes que aparecen en dichos textos. Como resultado, podemos afirmar que la estructura-AP es el dominio activo de la jerarquía de consulta que se obtuvo para un contexto y nivel dado dentro de la jerarquía de contextos.

Una de las operaciones más importantes que se pueden definir sobre la estructura-AP es la operación subestructura-AP inducida. Esta operación es particularmente significativa debido a que nos permitirá encontrar la representación de la estructura-AP, que corresponde a un texto dado para un determinado contexto. Esta estructura-AP resultante, no es más que la intersección de cualquier estructura-AP dada con un determinado conjunto de términos de búsqueda.

Una vez que han sido establecidas la estructura-AP y la subestructura-AP inducida, en la siguiente sección se explica la operación que nos permitirá determinar si un conjunto de términos aparece o no dentro de una estructura-AP dada. Esta operación será la base de la búsqueda implementada por nuestro modelo.

Búsqueda de conjuntos con la estructura-AP

La idea es que el usuario exprese sus búsquedas como conjuntos de términos, los cuales serán lanzados como una consulta en la jerarquía de dominio correspondiente con un nivel y contexto dado dentro de la jerarquía de contextos. Dicho lo anterior, una vez que los textos se han sido organizados por contextos, esta operación en particular, brindará la posibilidad al usuario final de realizar búsquedas semánticas por los tópicos más relevantes abordados en un contexto determinado.

5.1.3. Descripción del sistema OLAP

Antes de entrar en el proceso de integración de la dimensión contextual en un modelo multidimensional, daremos un breve vistazo al modelo multidimensional implementado en nuestro sistema. Para un mejor entendimiento, en la Figura 5.5 se muestra el esquema del modelo multidimensional creado por Wonder 3.0 para el ejemplo presentado en la Sección 5.2.2.

El modelo OLAP implementado por Wonder 3.0 está basado en el esquema *Copo de Nieve*, donde las dimensiones están normalizadas para ahorrar espacio de almacenamiento mediante la eliminación de datos redundantes (cada tabla representa un

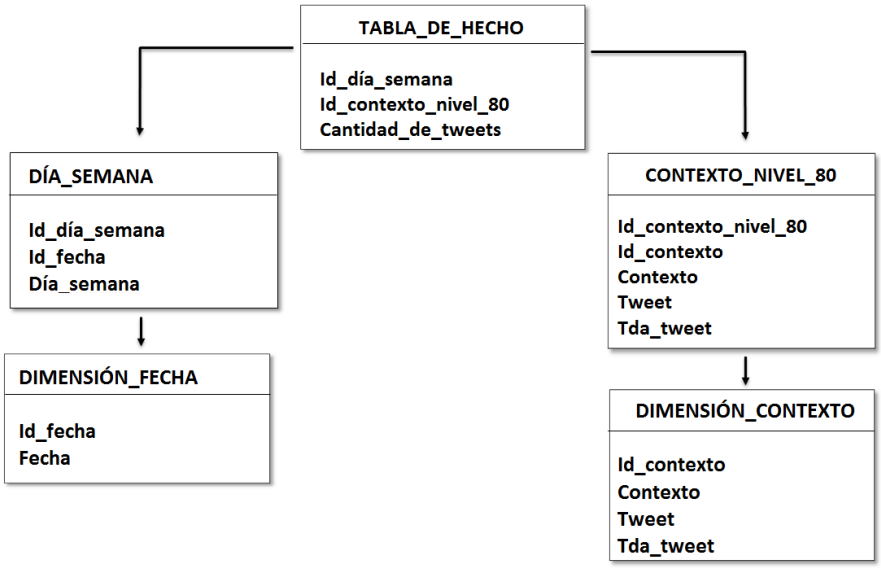


FIGURA 5.5: Ejemplo de esquema del modelo multidimensional implementado por Wonder 3.0

nivel en la jerarquía de la dimensión contextual). En este caso, se ha creado una tabla de hechos, que contiene referencias a las dimensiones (*FECHA* y *CONTEXTO*), así como también la variable cuantitativa a analizar (es decir, la cantidad de tweets publicados). Para el ejemplo que estamos tratando, cada dimensión tiene una jerarquía que permite el análisis de los datos a diferentes niveles de granularidad.

En el caso de la dimensión contextual, es necesario que el nivel más específico en la jerarquía de contexto sea el nivel 80, además de estar compuesto por el contexto al que pertenecen los tweets y su TDA asociado (tipo de dato abstracto) (*Tda_tweet*). Esta última, es la representación semántica de los textos mediante la cual se realizarán las búsquedas (jerarquía de consulta asociada).

La integración y la gestión de la nueva dimensión contextual en un modelo multidimensional, se realiza tomando como punto de partida el concepto de dimensión-AP presentado en [Martín-Bautista et al., 2010]. Es precisamente la dimensión AP y sus estructuras y operaciones asociadas, lo que nos permitirá integrar la dimensión contextual en el modelo multidimensional. Dicha integración se lleva a cabo mediante las dos acciones siguientes:

1. Crear una tabla de hechos para el nivel y contexto que se desea analizar dentro de la jerarquía de contextos combinada con las dimensiones tradicionales.

2. Para el caso de las dimensiones contextuales, la jerarquía de consulta asociada es incorporada y tratada exactamente igual que una dimensión-AP.

Por esta razón, el dominio de la jerarquía de consulta se definirá de la misma manera que el de una dimensión-AP, por lo tanto, el dominio será el conjunto de todas las subestructuras-AP de la estructura-AP global asociada a un atributo, los cuales son los posibles valores que puede alcanzar dicho atributo.

Operaciones dice, roll-up, y drill-down sobre la jerarquía de consulta

Al ser definida la jerarquía de consulta como una dimensión, las operaciones dice, roll-up y drill-down correspondientes al modelo multidimensional proceden de igual forma que en una dimensión-AP. En el ejemplo de la Figura 5.6 se muestra la funcionalidad de estas operaciones. Se debe aclarar que en este ejemplo sólo se muestra una parte de la jerarquía, es decir, que se pueden generar muchas más frases de búsquedas más específicas C^3. El ejemplo corresponde al contexto *COMPUTER SCIENCE* del nivel con 40 grupos de la jerarquía de contextos. Como se aprecia en el ejemplo, los datos pueden ser analizados por las frases iniciales C^1, así como por frases más o menos detalladas, conjuntos C^3 y C^2 respectivamente. Los detalles de las definiciones para estas operaciones se pueden consultar en [Martín-Bautista et al., 2010].

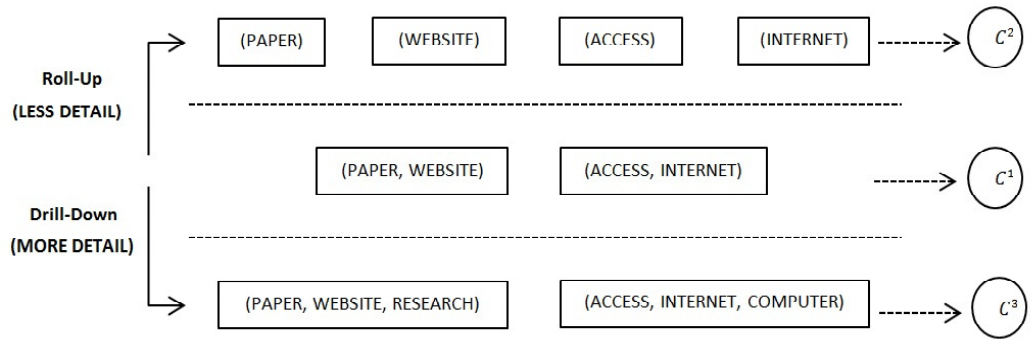


FIGURA 5.6: Ejemplo de uso de las operaciones roll-up y drill-down mediante la jerarquía de consulta

A continuación se muestra un ejemplo real de la estructura de la dimensión contextual que puede ser obtenida siguiendo todas las definiciones presentadas hasta este punto. En las secciones siguientes, se utilizarán las operaciones del modelo sobre estas estructuras para dar respuestas a posibles interrogantes del usuario.

En la Figura 5.7 se muestra la jerarquía de contextos correspondiente a la Figura 5.2(refinada), además por un problema de espacio, sólo se muestra un segmento de

la jerarquía de consulta que se obtiene para los contextos etiquetados como *COMPUTER_SCIENCE* y *PUBLISHING*. Cabe resaltar que con la integración de la dimensión contextual, y utilizando conjuntamente las operaciones tradicionales del modelo multidimensional y las que introduce nuestro modelo para dimensiones textuales, el usuario podrá buscar dentro de cada contexto perteneciente a un nivel dado, frases más y menos específicas que sean de su interés.

Como se ha mencionado, esto último constituye las aportación principal del presente capítulo, ya que permitirá al usuario estudiar los contextos obtenidos automáticamente, con la posibilidad de analizar conjuntamente, tanto dimensiones textuales como las tradicionales (fecha, lugar, etc.).

Una vez que han sido comentadas las definiciones y las principales operaciones que permiten la integración de la dimensión contextual en nuestro modelo multidimensional, en la Sección 5.2 explicaremos en detalle el proceso de implementación en la herramienta Wonder 3.0.

5.2. Metodología para crear e integrar la dimensión contextual en un modelo multidimensional

Para un mejor entendimiento hemos dividido nuestra metodología en tres pasos principales, los cuales corresponden a los módulos de la Figura 5.8. Esta figura recoge los procesos que conducen a la generación de la dimensión contextual a partir de un conjunto de datos textuales.

1. Creación de la jerarquía de contextos a partir de textos de redes sociales: los textos de redes sociales son procesados sintáctica y semánticamente. Luego mediante un algoritmo de agrupamiento jerárquico se crea la jerarquía de contextos, y cada contexto es etiquetado con las etiquetas más representativas presentes en los textos (etiquetas correspondientes a los términos originales). Este módulo no es explicado en el presente capítulo ya que los detalles se pueden encontrar en el Capítulo 3.

2. Creación de la jerarquía de consulta para un contexto y un nivel determinado en la jerarquía de contextos: para cada contexto de la jerarquía de contextos se crea una jerarquía de consulta a partir de los textos preprocesados mediante el algoritmo Apriori.

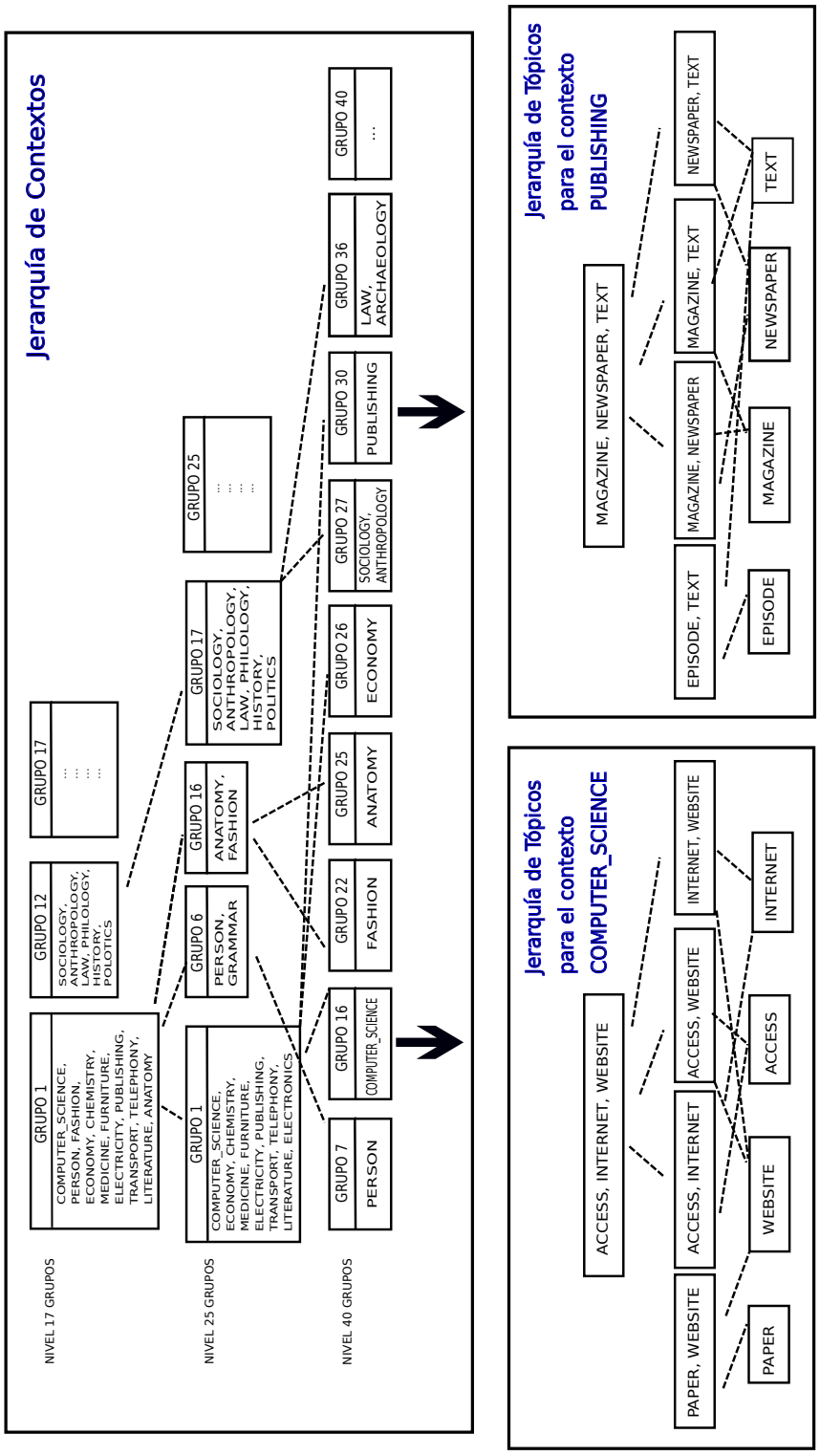


FIGURA 5.7: Ejemplo de una dimensión contextual para un conjunto de datos de Twitter

3. Integración de la dimensión contextual en un modelo multidimensional: la integración computacional de esta nueva dimensión en dicho modelo se lleva a cabo mediante la herramienta Wonder 3.0, que es un servidor OLAP de libre disposición. Los detalles de cada componente de este módulo se pueden verificar en [Martin-Bautista et al., 2013, Martin-Bautista et al., 2015]. Vale la pena mencionar que ha sido necesario la implementación de nuevas funcionalidades, ya que luego de crear un cubo OLAP que contiene la dimensión contextual, es necesario seleccionar previamente el nivel y el contexto en la jerarquía de contextos de los datos que se desea analizar.

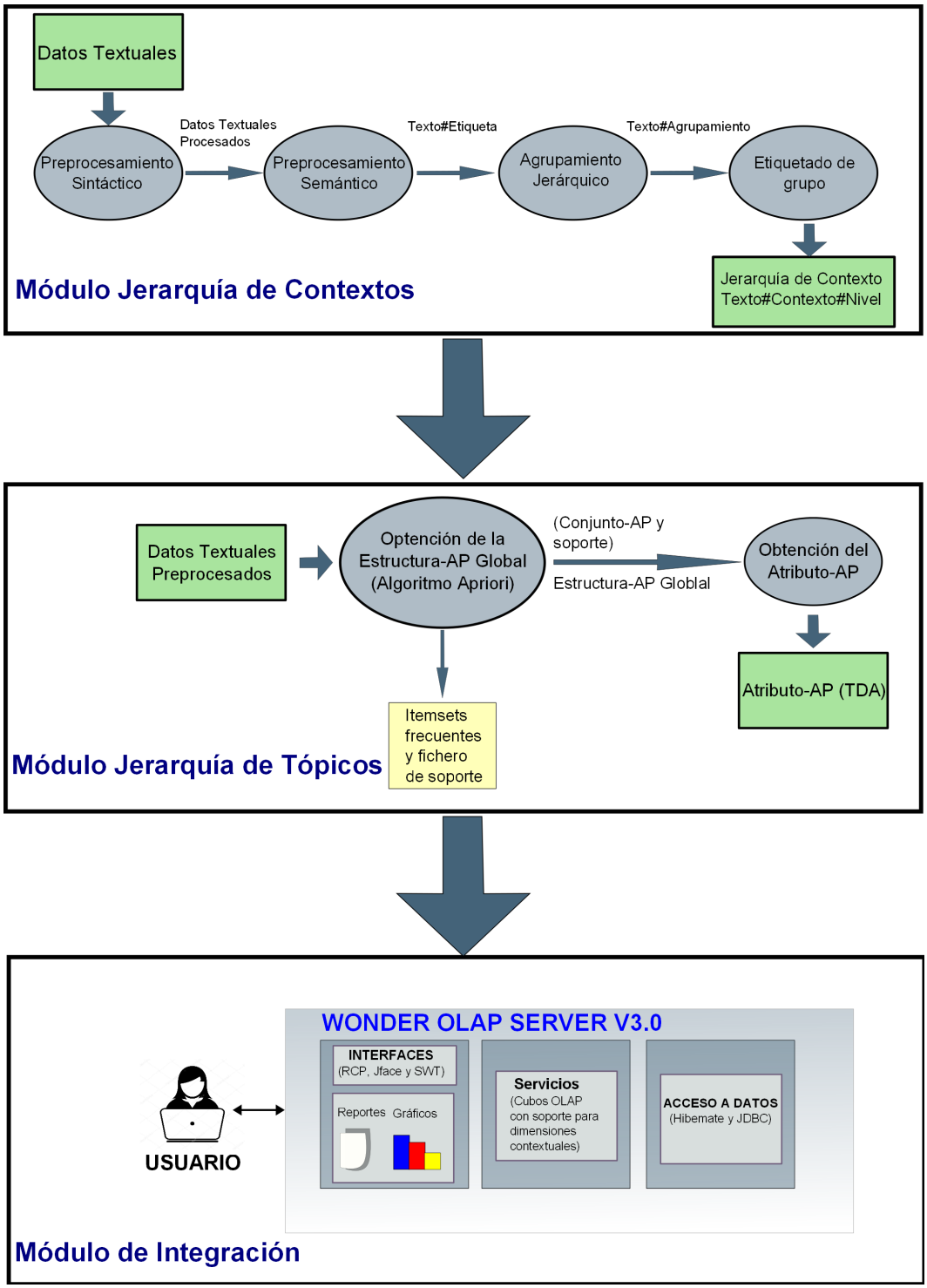


FIGURA 5.8: Metodología para la creación e integración de la dimensión contextual en el modelo multidimensional

5.2.1. Módulo para crear la jerarquía de consulta (jerarquía de dominio)

En la presente sección se explican los pasos que conforman el módulo que permite crear la jerarquía de consulta, así como los datos de entrada y de salida de cada proceso.

Tomando como punto de partida la jerarquía de contextos, es posible crear la jerarquía de consulta a partir de los textos preprocesados pertenecientes a un contexto y un nivel determinado de la jerarquía de contextos. El segundo módulo del sistema se encarga de construir dicha jerarquía utilizando como datos de entrada la salida del primer módulo.

Obtención de la estructura-AP global

Para este propósito, se utiliza el algoritmo Apriori tomando como estrada los datos textuales preprocesados para un contexto determinado. Con la implementación del algoritmo Apriori, se obtienen las dos estructuras de conocimiento fundamentales que encierran la semántica de los datos textuales, estas estructuras son los conjuntos-AP y la estructura-AP. Para la obtención de estas estructuras, se utiliza la herramienta Text Mining Tool V1.0 [Martin-Bautista et al., 2015].

La implementación del algoritmo Apriori, permite obtener los conjuntos de elementos que cumplan con el soporte que se le especifica como parámetro, así como los conjuntos-AP con sus correspondientes soportes. Luego se obtiene la estructura-AP global a partir de los conjuntos-AP maximales.

La herramienta implementada brinda la posibilidad de crear un fichero para cada una de las cardinalidades de los conjuntos de elementos generados, o almacenar en un único fichero todos los conjuntos de elementos obtenidos. La estructura de estos ficheros garantiza posteriormente, la optimización del algoritmo de intersección para la obtención del Tipo de Dato Abstracto (TDA).

En estos ficheros los términos aparecen ordenados alfabéticamente dentro de cada tupla, y por ende, el algoritmo de intersección puede realizar la búsqueda de una palabra dentro de una tupla mientras el cardinal del primer carácter de cada palabra que se busca sea menor o igual al cardinal del primer carácter de la palabra que se lee del fichero de conjuntos de elementos.

Obtención del atributo-AP

Hasta este punto, ya se cuenta con la estructura-AP global y los conjuntos que la forman. A partir de estos datos y de los textos preprocesados, se obtiene la estructura-AP inducida para cada tupla del atributo textual que se procesa. La estructura-AP

inducida se obtiene al intersectar la estructura-AP global y el atributo textual preprocesado sintácticamente. El resultado es almacenado como un nuevo atributo y se le conoce como TDA.

La implementación de la intersección, al igual que el proceso descrito anteriormente, se realiza de forma automática con la herramienta Text Mining Tool V1.0. Como se explicó anteriormente, la salida de este proceso es un nuevo atributo. Para cada valor de las tuplas se realiza el proceso de intersección con la estructura-AP global y el resultado final será el que se almacene como valor del TDA.

Para optimizar este proceso, el algoritmo implementado comienza buscando la intersección de cada tupla con los conjuntos-AP de máxima cardinalidad. De igual forma va calculando la intersección con cada uno de los conjuntos de las diferentes cardinalidades que componen la estructura-AP de forma decreciente. De esta manera se garantiza que el valor del TDA, se acopla con la frase de mayor cardinalidad en la estructura-AP.

La salida de este segundo módulo del sistema no es más que la jerarquía de consulta correspondiente a un grupo de textos preprocesados, de un contexto y un nivel determinado, en otras palabras, esta jerarquía de consulta está compuesta por el atributo textual preprocesado, así como el TDA asociado a dicho texto. Esta estructura almacena la semántica asociada a los principales términos (tópicos) presentes en los contextos detectados en los textos. Es por ello, que adquiere una gran importancia el hecho de lograr integrar esta jerarquía como una dimensión en un sistema multidimensional, y de esta forma permitir realizar OLAP con datos de redes sociales, teniendo en cuenta los tópicos abordados en atributos textuales de dichas redes.

5.2.2. Módulo de integración

El tercer y último módulo del sistema, permite integrar la dimensión contextual en un modelo multidimensional. Con este propósito, se utiliza la herramienta Wonder 3.0 [Martin-Bautista et al., 2013]. Entre las funcionalidades que brinda, está dar soporte a las dimensiones textuales, conocidas como dimensiones-AP. Si tenemos en cuenta que la dimensión contextual propuesta en el presente trabajo, es tratada de igual forma que una dimensión-AP, Wonder 3.0 se ajusta perfectamente para lograr la integración de dicha dimensión en un modelo multidimensional.

Sin embargo, se debe mencionar ciertas funcionalidades que han sido implementadas en Wonder 3.0 para garantizar su correcto funcionamiento. Para crear un cubo

OLAP con una dimensión contextual (cubo contextual), es necesario primeramente seleccionar el contexto y el nivel de la jerarquía de contextos para el cual se desea crear el cubo de datos, ya que el análisis de los datos se realizará por contextos. Una vez creado el cubo para un contexto específico, podemos realizar OLAP ya sea por la dimensión contextual, como por las dimensiones tradicionales que contenga el cubo.

La nueva dimensión contextual puede ser tratada de igual forma que las dimensiones clásicas, es decir, es posible realizar todas las operaciones básicas del modelo multidimensional sobre dicha dimensión. Es válido mencionar que debido a la propia estructura de la dimensión contextual (formada por la jerarquía de contextos y la jerarquía de consulta), es posible navegar tanto por los niveles de la jerarquía de contextos como por la jerarquía de consulta. Wonder 3.0 le permite al usuario seleccionar el nivel de la jerarquía de contextos, así como el contexto por el cual desea analizar los datos en el momento de crear el cubo OLAP. Asimismo, una vez que ha sido creado el cubo de datos, Wonder 3.0 brinda la posibilidad de realizar consultas más y menos detalladas mediante la jerarquía de consulta correspondiente al contexto seleccionado al crear el cubo.

Ejemplo ilustrativo

A continuación presentamos un ejemplo real de cómo crear y consultar un cubo contextual con datos de Twitter. Se creará un cubo con los tweets pertenecientes al contexto *Computer Science* y analizaremos la cantidad de publicaciones por día de la semana que corresponden con las frases de búsqueda *"COMPUTER"*, *"INTERNET"* y *"INTERNET COMPUTER"*. Se han seleccionado de forma aleatoria un total de 4,583 tweets. Además de los textos contamos con la fecha de publicación de los tweets, por lo que analizaremos los tweets teniendo en cuenta el día de la semana en que fue publicado.

Antes de crear el cubo contextual se debe construir la dimensión textual. Como se ha explicado en los apartados *"Módulo para crear la jerarquía de contextos"* y *"Módulo para crear la jerarquía de consulta"* dicha dimensión es creada de forma automática y permite organizar los textos o documentos a analizar en contextos, y a su vez para cada contexto se crea una jerarquía de consulta la cual brinda la posibilidad de estudiar los textos por los tópicos abordados en cada contexto.

La Figura 5.9, muestra la vista principal para crear un cubo contextual. Se puede apreciar que es necesario seleccionar la tabla o la vista donde se encuentras los datos

a analizar, así como el nivel de la jerarquía de contextos y el contexto deseado. En el presente ejemplo se ha seleccionado el nivel 80 y el contexto *Computer Science*, el cual consta de 1,708 tweets. En la TagCloud de dicha figura, aparecen los contextos a los que pertenecen los tweets.

De la misma forma la Figura 5.10 muestra la TagCloud generada cuando no se construye la dimensión contextual. Esta figura tiene el objetivo de mostrar la utilidad de nuestra propuesta a la hora de analizar datos no estructurados. En este caso, resulta más engorroso para el usuario realizar un análisis de los tweets mediante frases de búsquedas al no estar organizados en contextos.

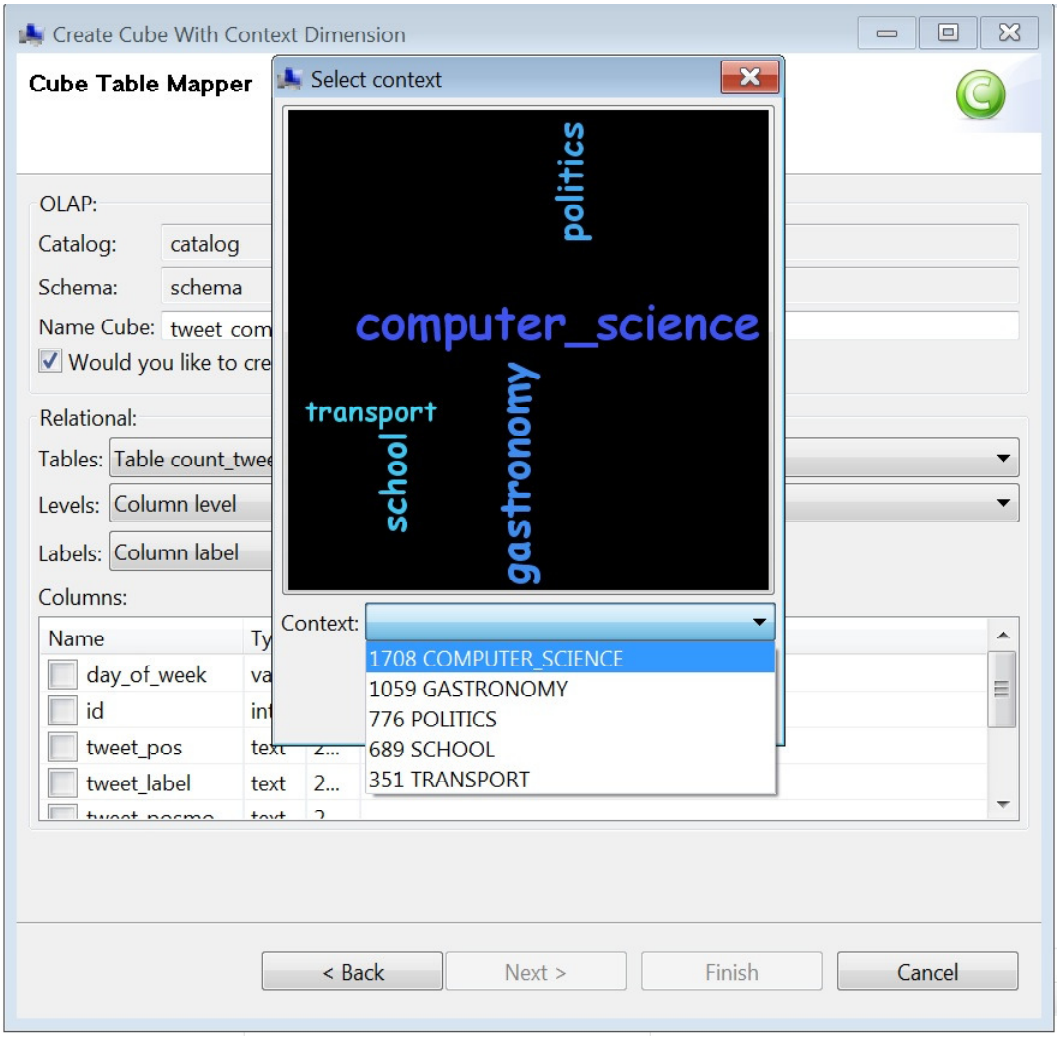


FIGURA 5.9: Interfaz principal para la creación de un cubo contextual

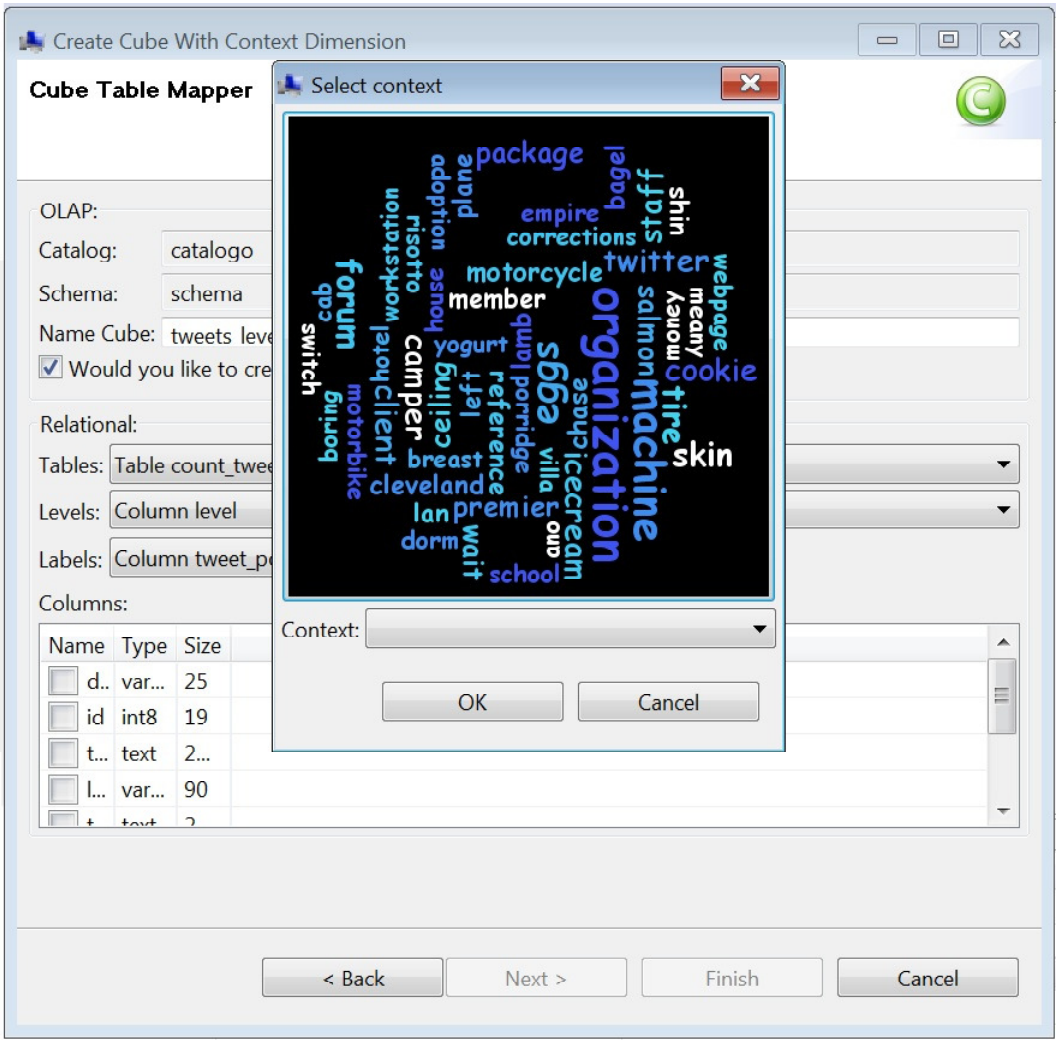


FIGURA 5.10: Interfaz principal para la creación de un cubo contextual (sin utilizar la metodología propuesta)

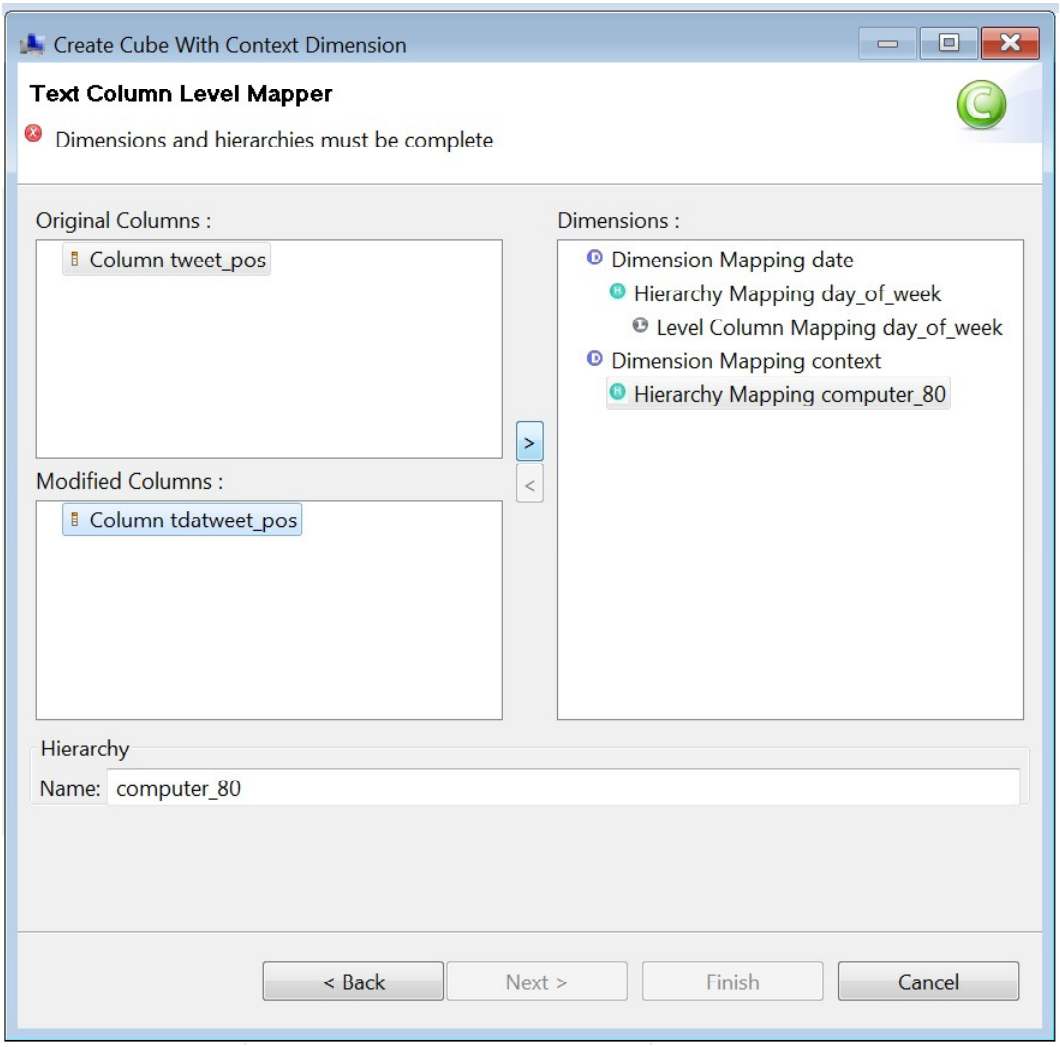


FIGURA 5.11: Página para seleccionar las dimensiones, jerarquías y niveles del cubo OLAP

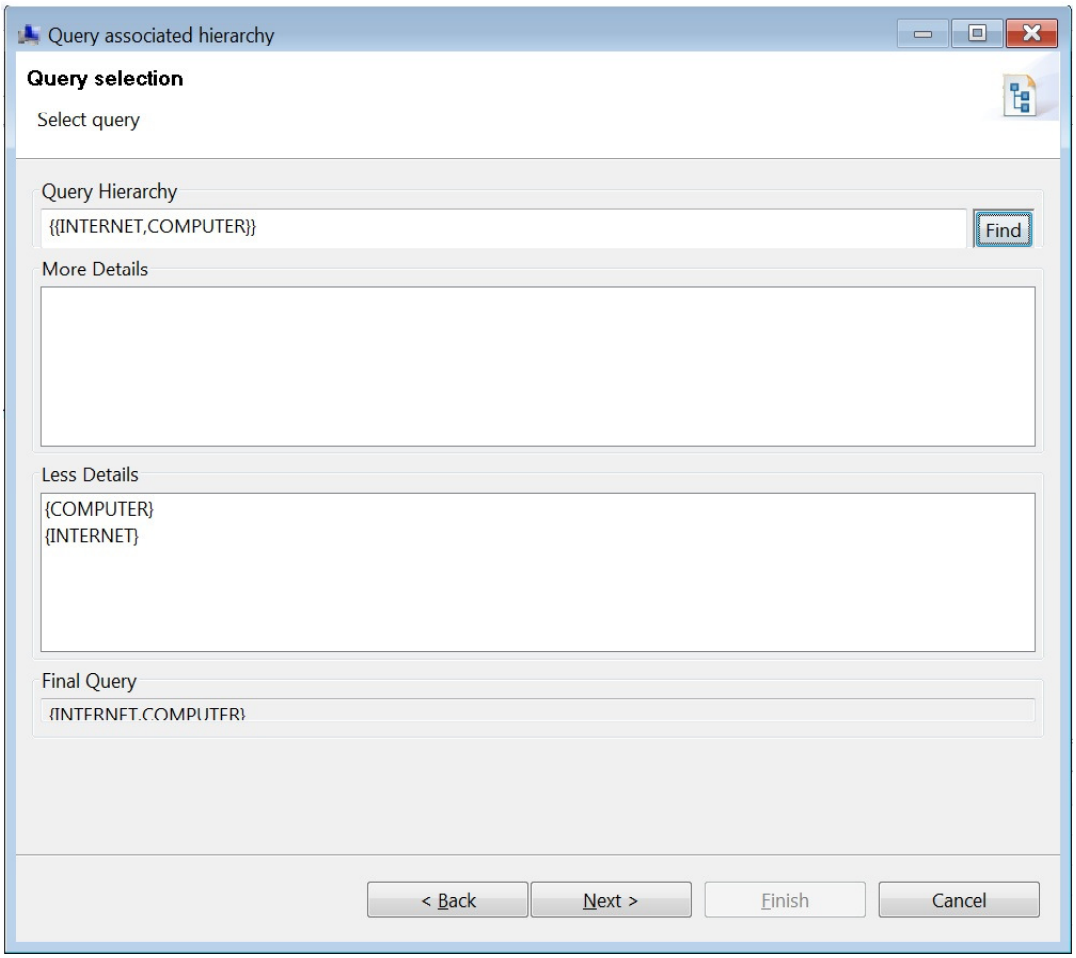


FIGURA 5.12: Página que permite realizar las operaciones roll-up y drill-down mediante la jerarquía de consulta

La Figura 5.11 muestra las dimensiones que formarán parte del nuevo cubo OLAP. En este caso, el cubo está formado por dos dimensiones (*date* e *context*), donde la segunda es una dimensión contextual formada por una jerarquía que contiene los valores correspondientes al nivel 80. Además se ha seleccionado como medida la cantidad de tweets.

Una vez creado el cubo de datos, realizaremos una consulta donde se ponga de manifiesto las ventajas de la dimensión contextual. La Figura 5.12 muestra la interfaz que brinda la posibilidad de realizar consultas sobre las dimensiones contextuales. Para el caso de la dimensión *"context"* podemos realizar las operaciones roll-up y drill-down mediante el uso de la jerarquía de tópicos. En este caso se desea conocer el número de intervenciones que contienen las frases *"COMPUTER"*, *"INTERNET"* e *"INTERNET COMPUTER"*.

Finalizado el proceso de creación de la consulta, el resultado puede ser analizado mediante el gráfico de la Figure 5.13. Dicho gráfico muestra la cantidad de tweets

por día de la semana que abordan los tópicos *"COMPUTER"*, *"INTERNET"* e *"INTERNET COMPUTER"*. Podemos apreciar como el número de tweets para las frases de búsquedas más generales lógicamente es mayor que para la frase más específica.

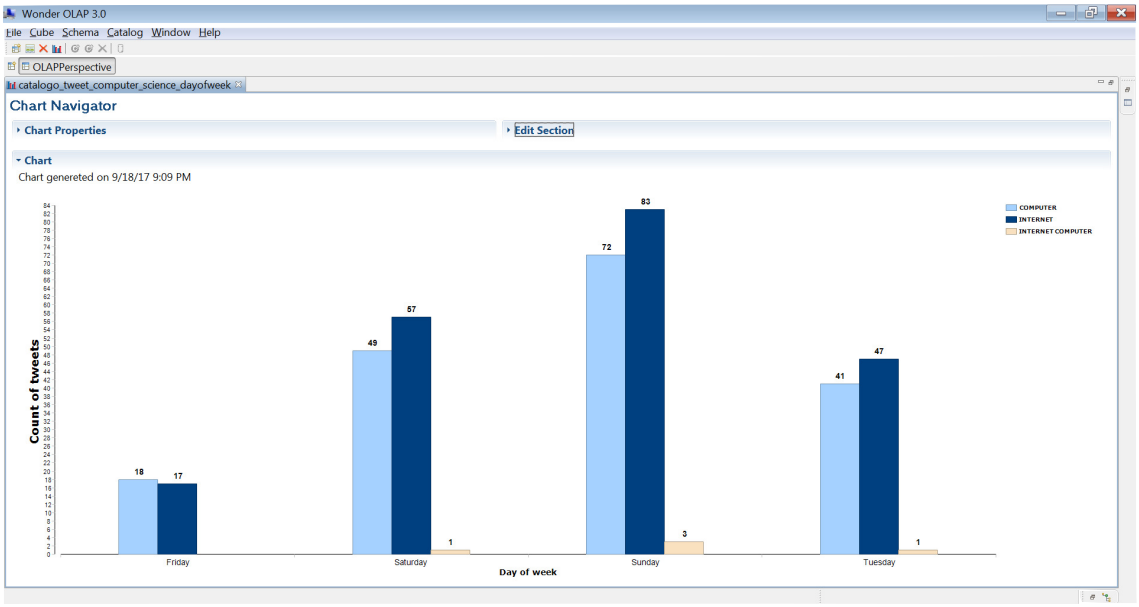


FIGURA 5.13: Interfaz del gráfico resultante al lanzar la consulta

5.3. Experimentos

En la presente sección se evalúa experimentalmente la metodología propuesta, para de esta forma demostrar las potencialidades que brinda la integración de la dimensión contextual en un modelo multidimensional. Primero, demostraremos que la metodología utilizada para detectar de forma automática los principales contextos presentes en datos textuales es correcta, y para ello hemos experimentado con datos reales de dos redes sociales con distintas cantidades de documentos. Luego mostraremos dos ejemplos de uso de la dimensión contextual mediante la herramienta OLAP Wonder 3.0.

5.3.1. Conjunto de datos

Se ha experimentado con varios conjuntos de datos reales, los cuales pertenecen a Twitter y a la Red Social Dreamcatchers en inglés y español respectivamente. Los datos de Twitter fueron descargados de Sentiment140[1], los cuales se encuentran en

[1]http://www.sentiment140.com/

formato CSV y constamos con seis campos, entre ellos el texto de los tweets y la fecha los cuales serán utilizados en el presente trabajo.

Por otra parte, DreamCatchers, desarrollada bajo un enfoque colaborativo entre sus miembros, se cuenta con la base de datos que le da soporte en PostgreSQL 9.4, con un total de 61 tablas. La información recogida es toda la relacionada con los datos personales y de afiliación del usuario, así como, las interacciones que realiza en su perfil y con otros usuarios. Dreamcatchers contiene varios campos con información textual. En este estudio hemos utilizado los comentarios de los usuarios, ya que proporcionan un conjunto de interacciones donde se discuten diversos tópicos.

5.3.2. Evaluación de la detección automática de contextos

Para demostrar que la metodología propuesta para la detección de contextos funciona independientemente del idioma y de la cantidad de documentos, se ha experimentado con cuatro conjuntos de datos de las dos redes sociales mencionadas.

En la Tabla 5.1 se presentan las estadísticas para dichos conjuntos de datos. Por un lado el número de documentos finales (**Cantidad de documentos**) luego de ser preprocesados sintáctica y semánticamente y sobre los cuales se detectarán los contextos. también el total de documentos preprocesados para cada red social (**Cantidad de documentos preprocesados (Twitter)** y **Cantidad de documentos preprocesados (Dreamcatchers)**). En el caso de Twitter se han utilizado los *tweets* y para Dreamcatchers los *comentarios*.

Se debe señalar, que en ambas redes sociales se descartan una gran cantidad de documentos. Esto se debe en gran medida a que la mayoría de los textos de las redes sociales expresan sentimientos u opiniones sobre determinado servicio o producto, y como se explicó anteriormente, los términos con orientación sentimental son descartados durante el preprocesamiento semántico.

Conjunto de datos	Cantidad de documentos	Cantidad de documentos preprocesados (Twitter)	Cantidad de documentos preprocesados (Dreamcatchers)
Conjunto 1	5000	8743	23125
Conjunto 2	10000	17672	50673
Conjunto 3	20000	35240	93911
Conjunto 4	30000	50399	109475

TABLA 5.1: Descripción de los conjuntos de datos utilizados

En las Figuras 5.14, 5.15, 5.16, y 5.17 se pueden apreciar los valores del coeficiente de silueta una vez aplicado el algoritmo de agrupamiento jerárquico Complete Link para 5000, 10000, 20000 y 30000 documentos para ambas redes sociales. Como se mencionó anteriormente, el coeficiente de silueta [Rousseeuw, 1987] es utilizado como medida para evaluar los algoritmos de agrupamiento jerárquico, ya que esta medida nos permite determinar la cantidad de grupos para la cual el algoritmo de agrupamiento tiene un mejor rendimiento.

Para cada conjunto de datos se realizaron cortes para siete cantidades de grupos (17, 25, 40, 60, 80, 100 y 120). En todos los casos los valores de la red social Dreamcatchers mejoran a los de Twitter, y además se estabilizan a partir de 60 grupos. De esta forma, se demuestra la validez de nuestra metodología para detectar los principales contextos independientemente de la cantidad de documentos analizados. Además se debe resaltar que es totalmente automática e independiente del idioma gracias al uso de recursos léxicos multilingüe como MCR 3.0 y de las herramientas Stanford POS y Stanford NER.

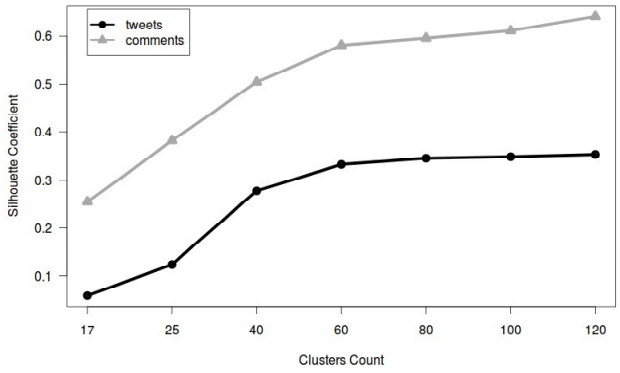


FIGURA 5.14: Valores del coeficiente de silueta para ambas redes sociales con 5000 documentos

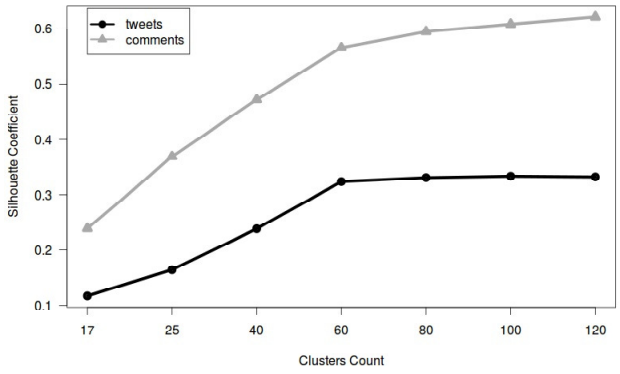


FIGURA 5.15: Valores del coeficiente de silueta para ambas redes sociales con 10000 documentos

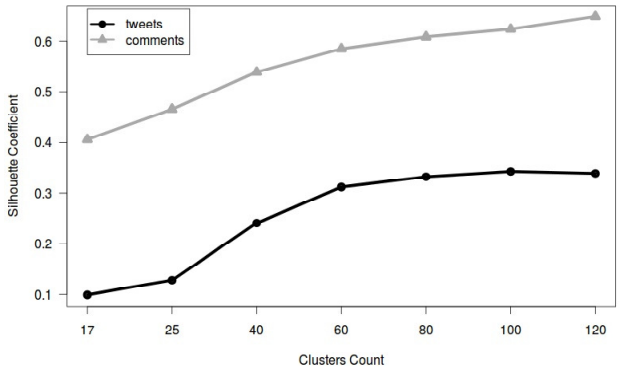


FIGURA 5.16: Valores del coeficiente de silueta para ambas redes sociales con 20000 documentos

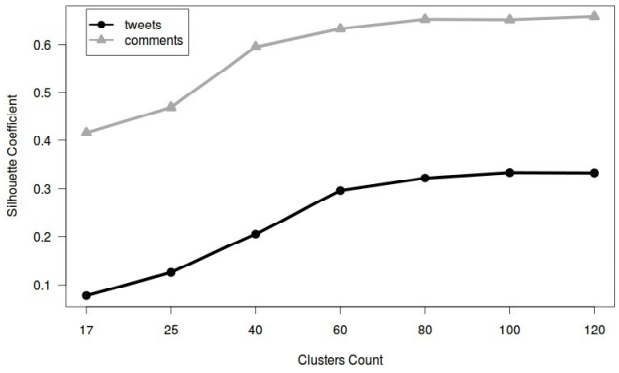


FIGURA 5.17: Valores del coeficiente de silueta para ambas redes sociales con 30000 documentos

5.3.3. Ejemplo práctico

Hemos seleccionado dos ejemplos prácticos para mostrar las ventajas que ofrece la integración de la dimensión contextual en un modelo multidimensional para el análisis de datos textuales (concretamente en redes sociales) conjuntamente con datos estructurados. Para el primer ejemplo se ha seleccionado el Conjunto 4 de Twitter, y una vez creada la dimensión contextual correspondiente a estos textos, se ha seleccionado el contexto *COMPUTER_SCIENCE* para la cantidad de 80 grupos.

En la Figura 5.18 se muestra una primera consulta general donde aparecen la cantidad de documentos (en este caso tweets) por días de la semana en el contexto seleccionado (*COMPUTER SCIENCE*). En total este contexto consta de 1680 tweets distribuidos en los cuatro días de la semana (Friday, Saturday, Sunday y Tuesday). Como se mencionó en la Sección 5.1.2 para cada contexto de la jerarquía de contextos se crea una jerarquía de consulta, la cual permite realizar consultas más y menos detalladas a partir de los principales tópicos presentes en dicho contexto.

La Figura 5.19 muestra el resultado de la consulta realizada en el contexto *COMPUTER SCIENCE* con varias frases de búsquedas. En este caso, los valores representan la cantidad de tweets que contienen todos los términos presentes en la frase de búsqueda por días de la semana.

Por otra parte, para apreciar el funcionamiento de las operaciones roll-up y drill-down del modelo multidimensional sobre la jerarquía de dominio, se han seleccionado frases de búsqueda más y menos específicas como son *"INTERNET PHONE*

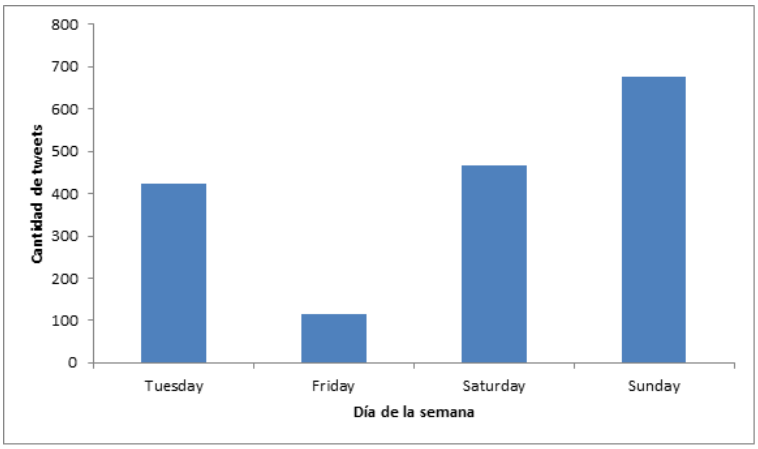


FIGURA 5.18: Ejemplo de consulta para el Conjunto 4 de Twitter (contexto = *COMPUTER SCIENCE* y nivel = 80 grupos)

TWITTER" e *"INTERNET"* respectivamente. Al analizar la cantidad de tweets de estas frases para cada día de la semana, podemos ver que en todos los casos la cantidad de tweets es mayor para la frase *"INTERNET"*.

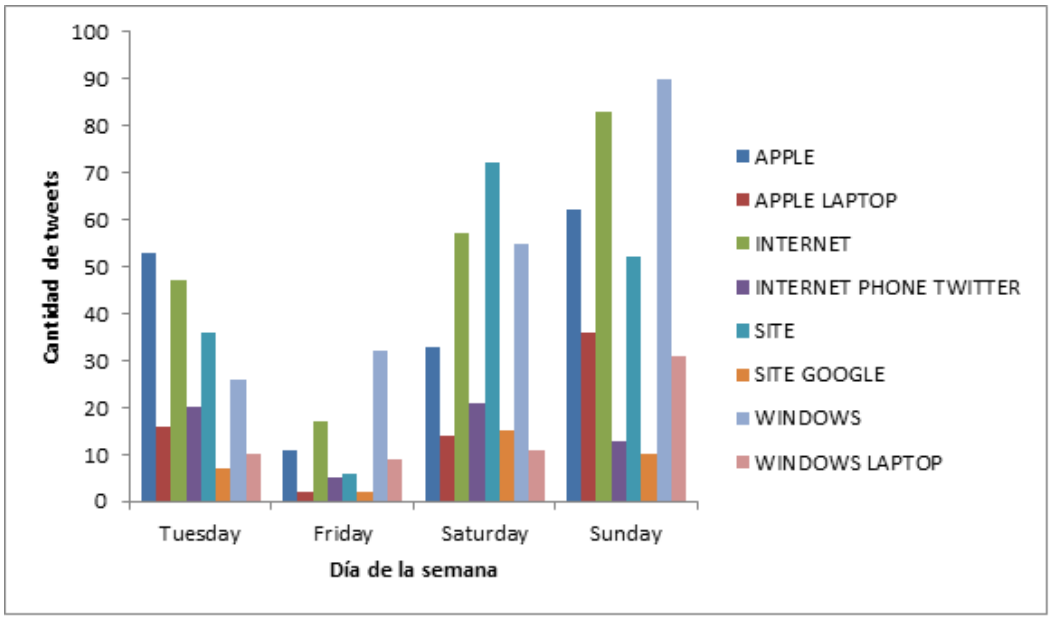


FIGURA 5.19: Ejemplo de consulta para el Conjunto 4 de Twitter utilizando la jerarquía de consulta (contexto = *COMPUTER SCIENCE* y nivel = 80 grupos)

Para el segundo ejemplo se ha seleccionado el Conjunto 4 de Dreamcatchers. Además se ha seleccionado el contexto *ANATOMY* y la cantidad de 100 grupos. En la Figura 5.20 se muestra la consulta general donde aparecen la cantidad de documentos (en este caso los comentarios) por ciudad en el contexto *ANATOMY* (910 comentarios).

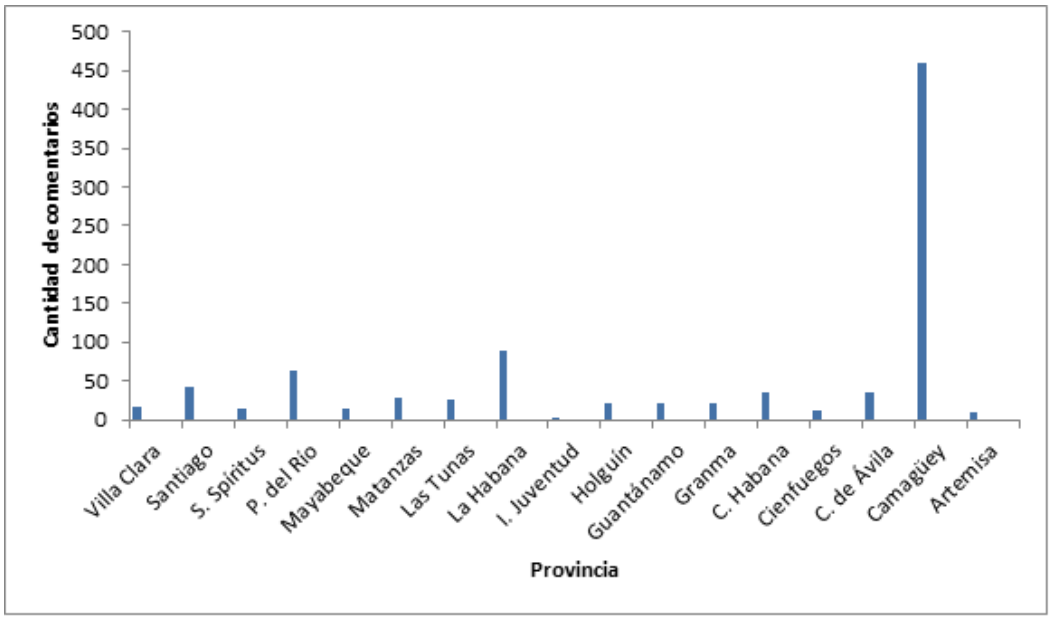


FIGURA 5.20: Ejemplo de consulta para el Conjunto 4 de Dreamcatchers (contexto = *ANATOMY* y nivel = 100 grupos)

Por su parte la Figura 5.21 muestra el resultado de la consulta realizada en el contexto *ANATOMY* con varias frases de búsquedas. En este caso, los valores representan la cantidad de comentarios que contienen todos los términos presentes en la frase de búsqueda por ciudad del usuario que realizó el comentario. Como se puede apreciar, se ha realizado un dice en la dimensión ciudad permaneciendo sólo las ciudades *'Camagüey, Ciego de Ávila, Ciudad Habana, La Habana, Pinar del Río y Santiago de Cuba*. También se han seleccionado frases de búsqueda más y menos específicas como *"BOCA NARIZ"* y *"BOCA"* respectivamente, para así poder comprobar las operaciones roll-up y drill-down del modelo multidimensional.

5.3.4. Evaluación del rendimiento de las consultas

En la presente sección hemos incluido el tiempo de respuesta del sistema para las cuatro consultas realizadas en la Sección 5.3.3 para dar una muestra del rendimiento del mismo. A continuación listamos las características del ordenador con el que se ha realizado la prueba.

- **Sistema operativo:** Ubuntu 16.04.
- **Tipo de sistema operativo:** 64 bits.
- **CPU (procesador):** 8 x Intel Core i7-4700MQ 2.4GHz.

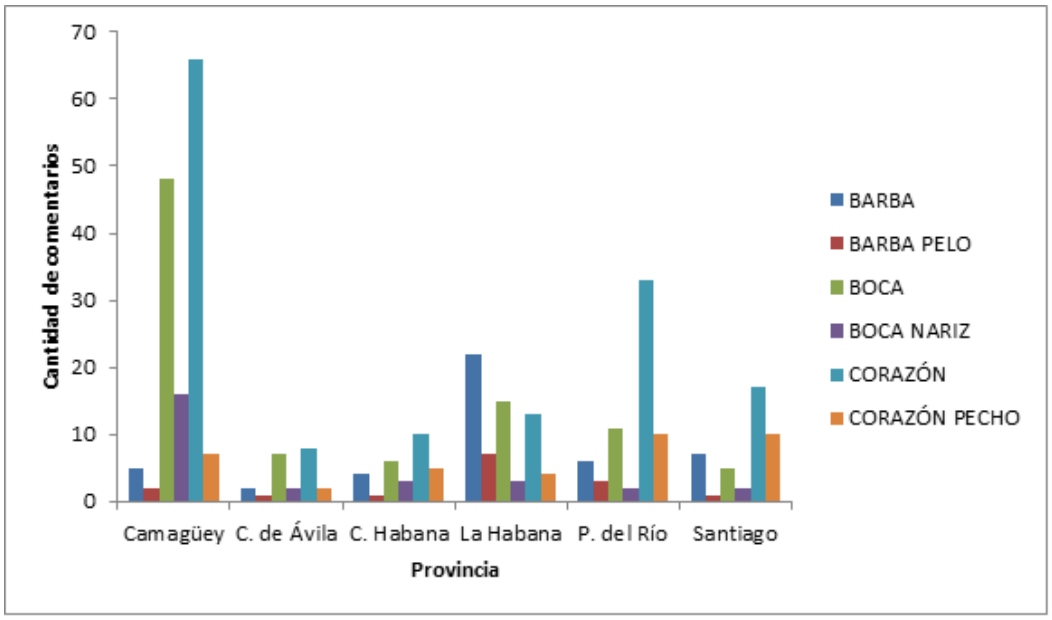


FIGURA 5.21: Ejemplo de consulta para el Conjunto 4 de Dreamcatchers utilizando la Jerarquía de consulta (contexto = *ANATOMY* y nivel = 100 grupos)

- **Memoria RAM:** 7.7GB.

La Tabla 5.2 muestra los resultados en milisegundos para cada una de las consultas realizadas en la sección anterior. Aunque el número de documentos con los que se ha realizado la evaluación no es elevado, debe quedar claro que el sistema es capaz de lidiar con grandes volúmenes de datos como en la gran mayoría de los problemas de la mundo real.

Además, se debe tener en cuenta que se ha experimentado con un máximo de 30000 documentos, pero una de las ventajas de nuestra propuesta, es precisamente que al tener los textos previamente agrupados por contextos, al analizar un contexto determinado la cantidad de textos a consultar se reduce considerablemente.

Consulta	**Cantidad de documentos**	**Tiempo de ejecución (milisegundos)**
Consulta 1	1680	1325
Consulta 2	1680	290
Consulta 3	910	255
Consulta 4	910	539

TABLA 5.2: Tiempo de ejecución de las consultas

5.4. Conclusiones

El presente capítulo tiene como objetivo facilitar el análisis multidimensional de datos textuales una vez que han sido previamente agrupados en contextos. Con este propósito se construye e integra una nueva dimensión en un modelo multidimensional a partir de datos textuales de redes sociales. La nueva dimensión la hemos llamado dimensión contextual y está formada por dos componentes: los contextos detectados en los textos (jerarquía de contextos) y los principales tópicos tratados en cada contexto de la jerarquía de contextos. Se obtiene a partir de técnicas de minería de datos (Algoritmos de Agrupamiento Jerárquico) y su construcción es completamente automática e independiente del idioma de los textos.

Luego esta dimensión es integrada en un modelo multidimensional permitiendo así el análisis de los datos textuales por contextos y tópicos de igual forma y conjuntamente con las dimensiones convencionales. Para facilitar la integración fue necesario extender el modelo multidimensional (extensiones de almacenamiento y consulta) implementado por el sistema OLAP utilizado (Wonder 3.0).

La experimentación fue realizada con datos reales de redes sociales, y se utilizó el sistema Wonder 3.0 como servidor OLAP para la integración de la dimensión contextual. Los resultados demuestran la viabilidad de la metodología para la construcción automática de la dimensión contextual, así como las posibilidades de consultas para el análisis multidimensional de los datos textuales de redes sociales integrados con dimensiones convencionales.

Capítulo 6

Explotación de la dimensión contextual mediante el servidor OLAP "Wonder OLAP Server 3.0"

El presente capítulo tiene como objetivo demostrar las diferentes posibilidades que proporciona el uso de la dimensión contextual propuesta en el capítulo anterior en un modelo multidimensional. Primero se describe la arquitectura general del sistema y se explican los componentes de dicha arquitectura y las herramientas utilizadas para la implementación del sistema.

Además se explica de forma detallada el proceso de integración de la dimensión contextual en un modelo multidimensional, así como las facilidades de consulta que brinda para analizar datos textuales de redes sociales combinados con datos convencionales. Para ello, se utilizará el sistema OLAP mencionado en capítulos anteriores, Wonder 3.0.

Los ejemplos presentados, muestran consultas sobre cubos de datos que contienen dimensiones contextuales (construidas con datos de las redes sociales Twitter y Dreamcatchers) y clásicas de forma conjunta. A través de ellos se plasma la viabilidad de la herramienta implementada y el cumplimiento del objetivo mencionado.

6.1. Descripción del sistema

A partir de un conjunto de datos textuales, los cuales pueden encontrarse en un fichero de textos o en una base de datos, se construye la dimensión contextual asociada al atributo textual. Estos textos, son preprocesados sintáctica y semánticamente

con el fin de facilitar la creación automática de la dimensión contextual. Posteriormente esta dimensión es almacenada en una base de datos, la cual se encuentra en el SGBD PostgreSQL.

Una vez creada la dimensión contextual se procede a integrarla en un servidor OLAP para su posterior análisis. Como ya mencionamos, en el presente trabajo utilizamos Wonder 3.0, el cual es un servidor OLAP de libre disposición. Fue implementado con técnicas y herramientas de software libre y brinda la posibilidad de gestionar cubos OLAP para cualquier base de datos en PostgreSQL. Tiene una particularidad que lo distingue y es que el modelo multidimensional que implementa, brinda soporte a datos textuales de base de datos mediante la dimensión-AP propuesta en [Martin-Bautista et al., 2013]. Como ya se mencionó en el Capítulo 5, esta característica constituye el punto de partida para que Wonder 3.0 brinde soporte a la dimensión contextual.

6.1.1. Arquitectura general

La arquitectura general del sistema se muestra en la Figura 6.1. A continuación describimos sus componentes.

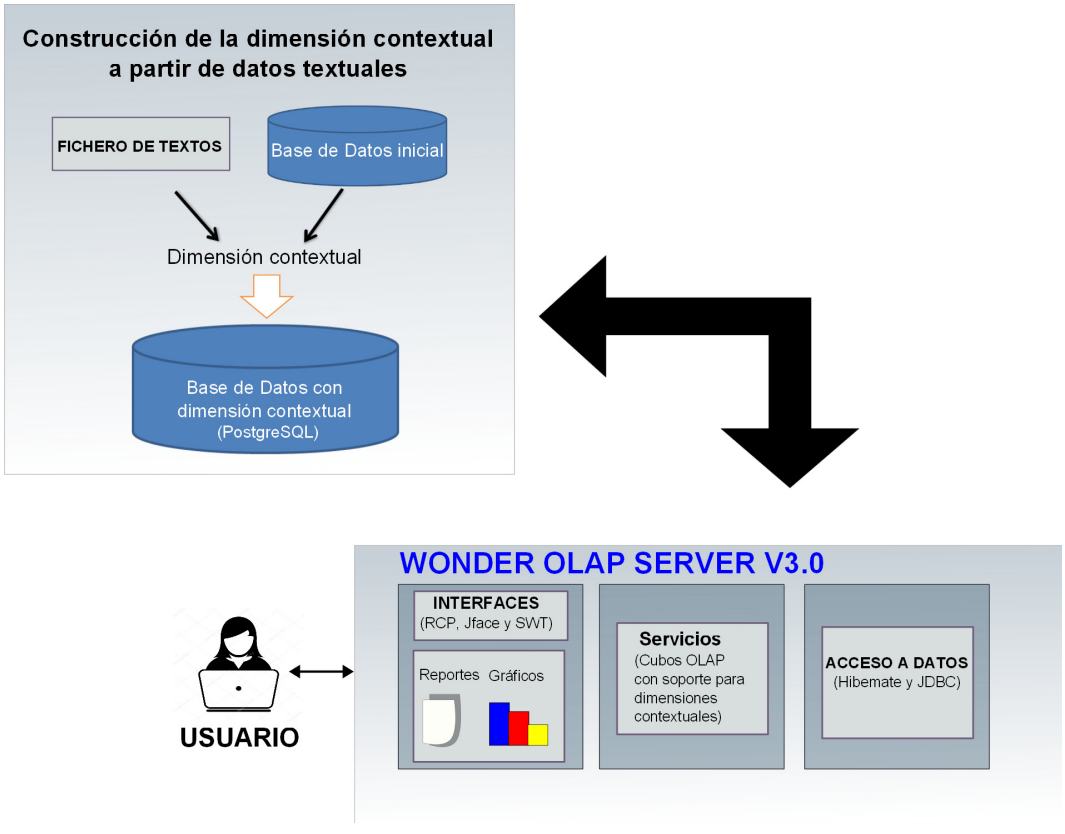


FIGURA 6.1: Arquitectura general del sistema

Construcción de la dimensión contextual: A partir de un conjunto de datos textuales, este sistema permite construir una dimensión contextual. Como se puede apreciar en la Figura 6.1 los textos pueden proceder tanto de ficheros como de bases de datos. Una vez creada la dimensión contextual, es almacenada en una base de datos en PostgreSQL. Además del texto analizado, la dimensión está compuesta por la etiqueta correspondiente al contexto que pertenece, el nivel de la jerarquía de contextos seleccionado, así como la estructura-AP inducida para cada texto. El Apéndice A, describe en detalles las principales características del sistema que permite crear la dimensión contextual.

Wonder OLAP Server: Este sistema tiene como fuente de datos la base de datos donde se almacena la dimensión contextual creada. De esta base de datos carga la información necesaria para crear los cubos de datos, y de forma especial, la información relacionada con la dimensión contextual.

Está formado por capas, que interactúan entre sí para dar respuesta a las necesidades del cliente. Mediante este sistema, se pueden construir cubos de datos con dimensiones contextuales y clásicas indistintamente. Brindando así la posibilidad de convertir en información útil para la toma de decisiones, la información implícita asociada a los datos textuales de redes sociales contenida en dichas dimensiones contextuales.

Wonder también permite realizar consultas usando las dimensiones contextuales y clásicas de forma conjunta, efectuando las operaciones OLAP usuales del modelo multidimensional (Roll-up, Drill-Down, Slice y Dice) para cualquiera de ellas. Además se destaca el uso de una jerarquía de consulta por contexto y nivel de la jerarquía de contextos que forma parte de la dimensión contextual.

6.1.2. Arquitectura de Wonder 3.0

La arquitectura de Wonder ha sido implementada tomando como base la arquitectura propuesta por Microsoft para el desarrollo de aplicaciones Cliente-Servidor en varias capas. Concretamente, esta arquitectura propone un grupo de capas lógicas en las que se puede dividir la aplicación cliente, para un mejor funcionamiento y comunicación entre todos sus componentes. En la Figura 6.2 se muestra esta arquitectura.

En la Figura 6.2 las INTERFACES son las que le permiten a los usuarios ver los cubos, los informes y los gráficos de la aplicación. La capa de los SERVICIOS es la

contenedora de todos los servicios y la capa ACCESO A DATOS es la encargada del acceso a los metadatos de los cubos OLAP, los datos de los cubos OLAP y los datos fuentes. Esta capa de acceso a datos se comunica con tres bases de datos, la base de datos fuente que es desde donde se extraen los datos para poblar los cubos, la base de datos donde se almacenan los modelos y la base de datos donde se guardan los datos de los cubos OLAP creados. Las capas pueden describirse en detalle como sigue:

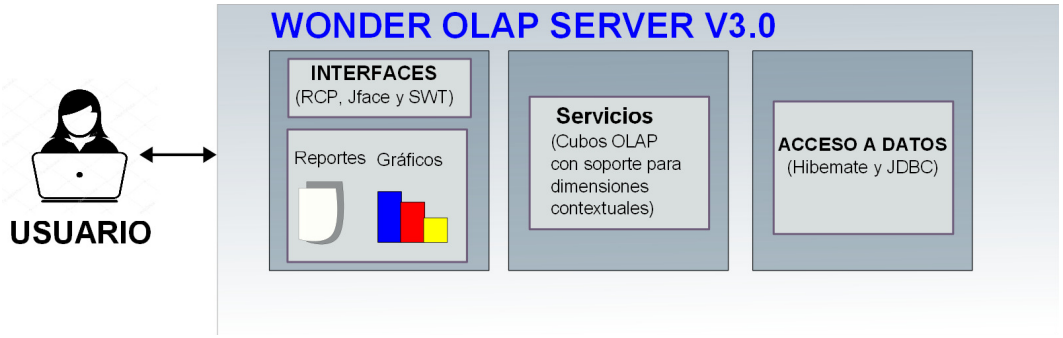


FIGURA 6.2: Arquitectura de Wonder

1. INTERFACES: esta capa de la aplicación contiene las interfaces que se muestran al usuario para que interactúe con el software. Permiten la creación de cubos OLAP para después analizar sus datos ya sea a través de gráficos o informes. Las tecnologías que se utilizaron para el diseño de esta capa son: la Plataforma de Cliente Enriquecido de Java (Rich Client Platform (RCP)) [McAffer and Lemieux, 2005], y las librerías JFace y SWT [Bull et al., 2004].

Las interfaces están separadas por vistas, la Vista OLAP que es la vista donde se muestran los cubos creados y las distintas operaciones que se pueden realizar sobre estos; la Vista de Informes que refleja los datos de los cubos y la Vista de Gráficos que es la vista donde se muestran los gráficos obtenidos. La Vista de los Gráficos tiene como limitante que muestra solo los datos bidimensionales, o sea, que si el cubo de datos es de más de dos dimensiones se le tiene que aplicar la operación Slice.

2. SERVICIOS: es la capa encargada de lograr la sincronización y la organización de las interacciones con el usuario. En ella se encuentran los servicios que encapsulan toda la lógica relacionada con un modelo multidimensional. Esta capa logra la persistencia del modelo de la aplicación mediante la capa de acceso a datos. Aquí son varios los servicios que se implementan. Todos ellos se pueden ver consultando en la ayuda el plugin *cu.jagger.wonder.service.pack.impl*.

3. ACCESO A DATOS: esta capa se encarga de la persistencia de los datos. En ella se encuentran las clases que se encargan de mantener sincronizados los datos que se muestran en la aplicación, con los datos almacenados en las bases de datos. Para cada clase del modelo existe una clase de acceso a datos. Las tecnologías que se utilizaron para su implementación son: Hibernate [Bauer and King, 2006] para la persistencia de los Metadatos OLAP y JDBC(Conectividad a bases de datos con Java) [Cecchet et al., 2004] para acceder a la fuente de datos y a los cubos OLAP. En esta capa intervienen 3 bases de datos almacenadas en PostgreSQL. Una base de datos fuente (BD Inicial Modificada) que contiene los datos fuentes, una base de datos de modelos (modelo_olap) donde se almacenan los metadatos correspondientes a los cubos OLAP y otra base de datos (Warehouse), donde se almacenan los cubos de datos construidos, ésta se comunica con la base de modelos para la construcción de sus cubos OLAP y con la base de datos fuente para cargar los datos necesarios para dichos cubos. En Warehouse puede haber cubos que contengan dimensiones clásicas y dimensiones-AP indistintamente.

Como SGBD Wonder utiliza PostgreSQL que además de ser libre es gratuito y se puede descargar libremente de su Sitio Web (http://www.postgresql.org/) para multitud de plataformas. La versión oficial actual de PostgreSQL es la 9.4.2. Está considerado como la base de datos de código abierto más avanzada del mundo. PostgreSQL proporciona un gran número de características que normalmente sólo se encontraban en las bases de datos comerciales tales como DB2 u Oracle.

6.2. Definición de un cubo contextual en Wonder OLAP Server 3.0

6.2.1. Creación del cubo de datos con dimensión-AP

En esta sección se describe el proceso de creación de dos cubo de datos con el uso de Wonder. Los cubos son creados a partir de datos de Twitter y Dreamcatchers y en cada caso el tratamiento de las dimensiones textuales será mediante la dimensión-AP.

Cubo OLAP para el caso de Twitter

En la Figura 6.3 se muestra la interfaz que permite crear las dimensiones clásicas que va a tener el cubo. En este caso, se crea una la dimensión *fecha*. De igual forma la Figura 6.4 permite crear las dimensiones-AP, las cuales están compuestas por el atributo textual y el TDA. En este caso se ha creado una dimensión-AP que facilitará el análisis del atributo *tweet*. Este TDA no es más que la estructura-AP inducida para cada tupla y contiene la semántica asociada al atributo textual. Finalmente en la Figura 6.5 podemos ver mediante la vista de propiedades las característica del cubo creado.

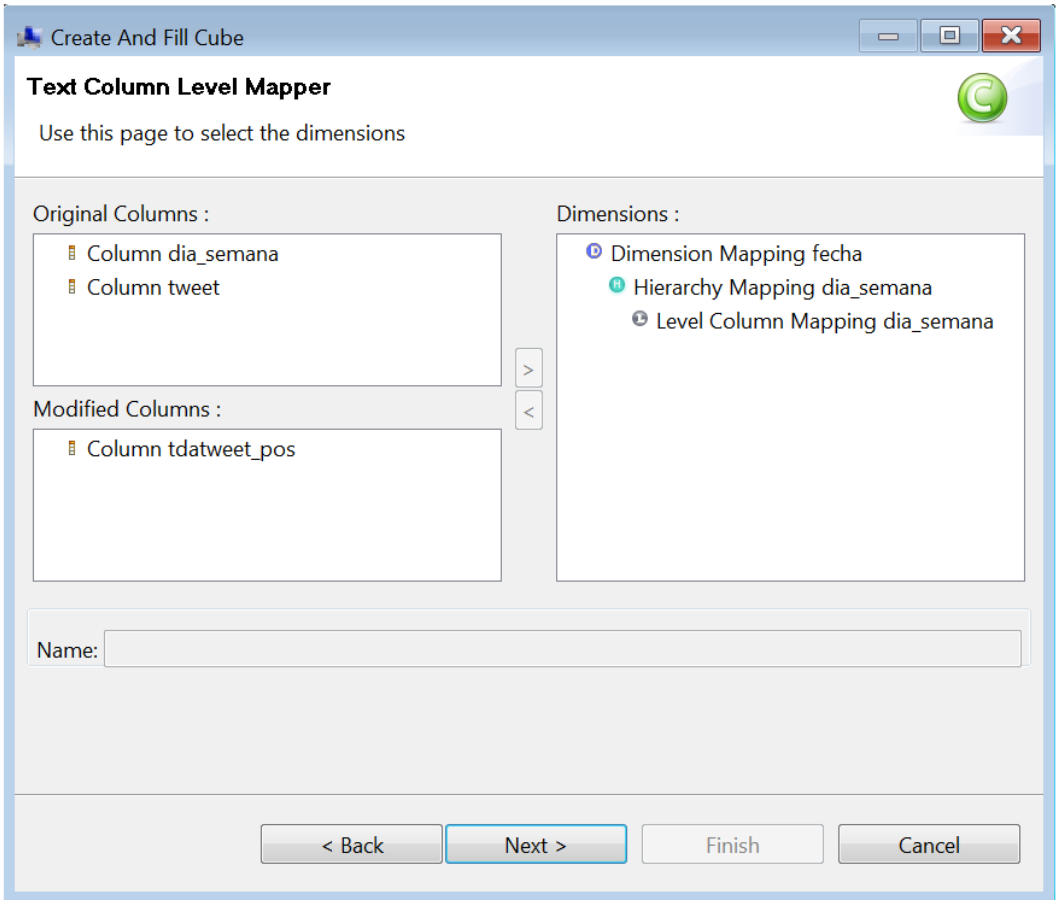


FIGURA 6.3: Definición de las dimensiones clásica

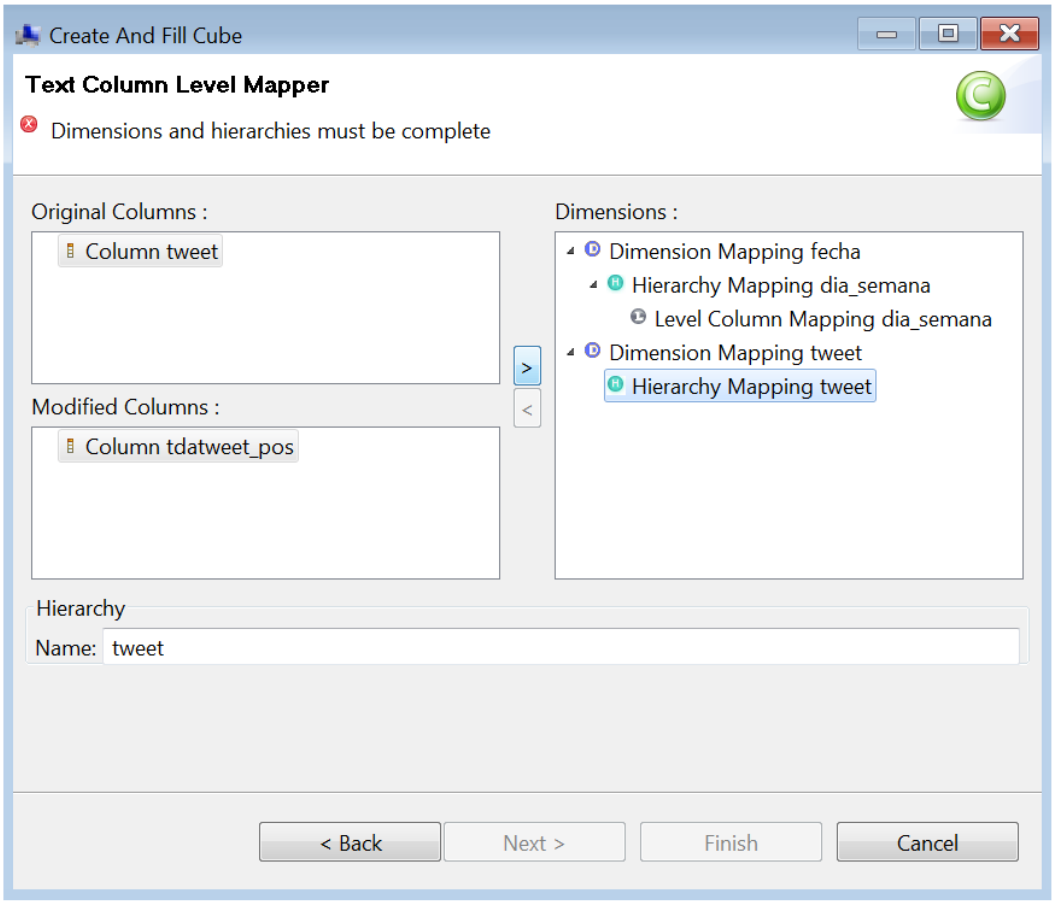


FIGURA 6.4: Definición de las dimensiones-AP

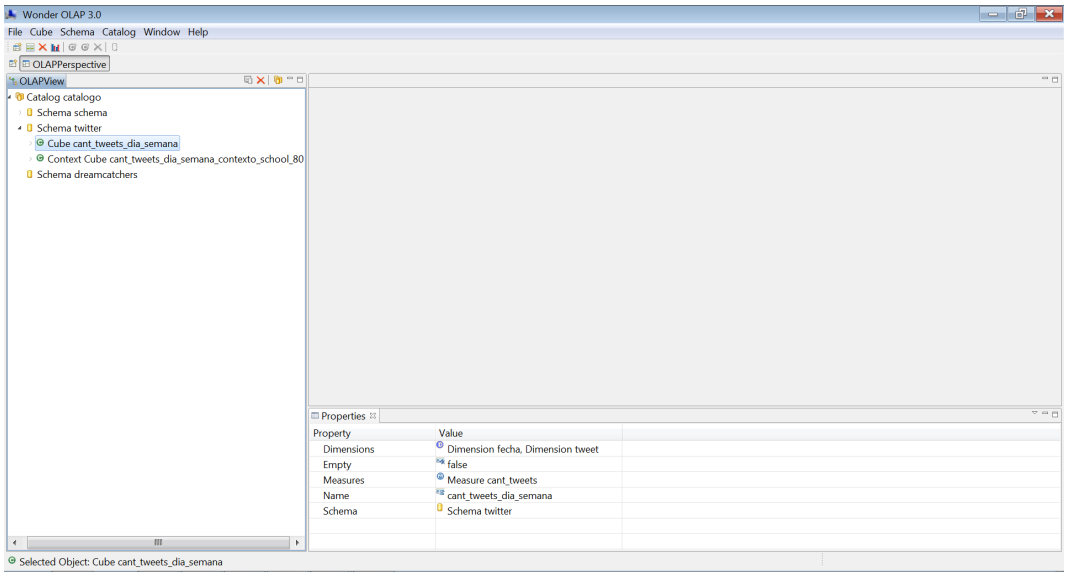


FIGURA 6.5: Vista que permite ver las propiedades del cubo creado

Cubo OLAP para el caso de Dreamcatchers

De igual forma que para Twitter, hemos creado un cubo OLAP con dimensiones-AP. La Figura 6.6 muestra la interfaz que permite crear las dimensiones clásicas que va a tener el cubo. En este caso, se crea la dimensión *lugar*. La Figura 6.7 permite crear las dimensiones-AP de la misma forma que para cubo de Twitter. En este caso la dimensión-AP contiene la semántica asociada al atributo *comentario*. Finalmente en la Figura 6.8 se pueden ver las característica del cubo creado.

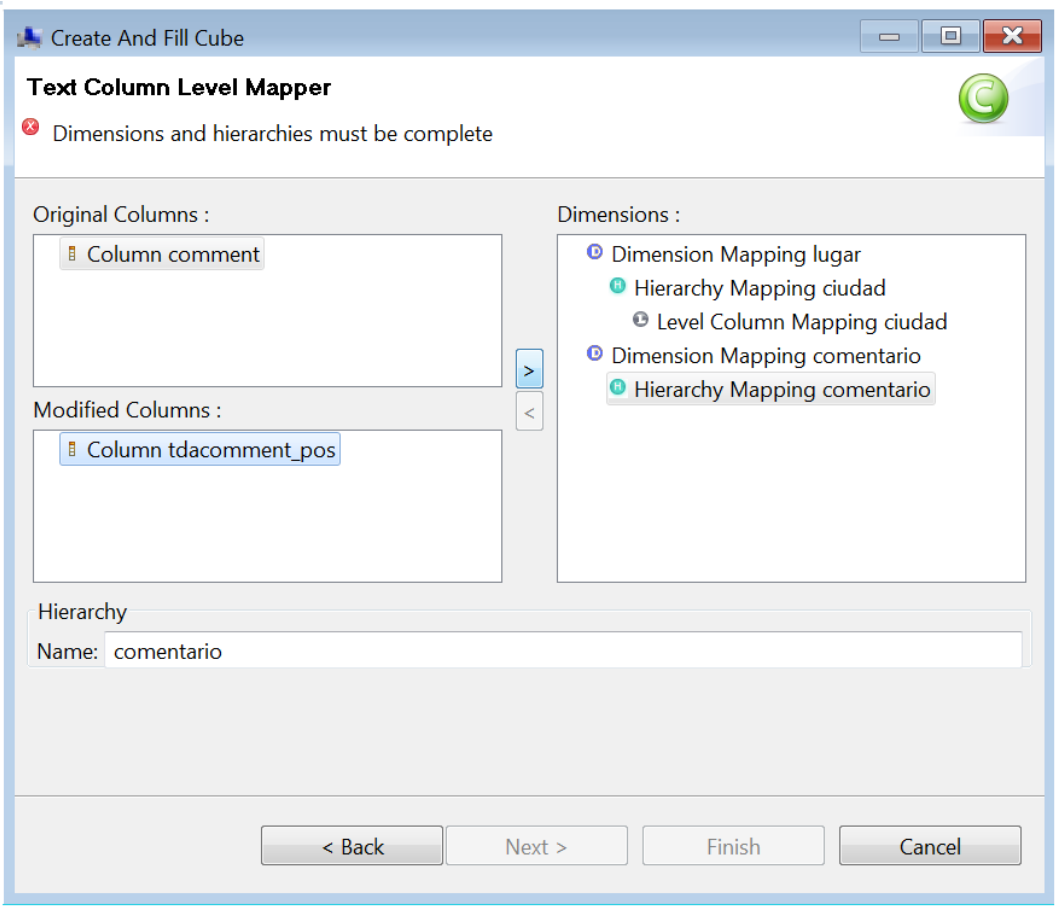


FIGURA 6.6: Definición de las dimensiones clásica

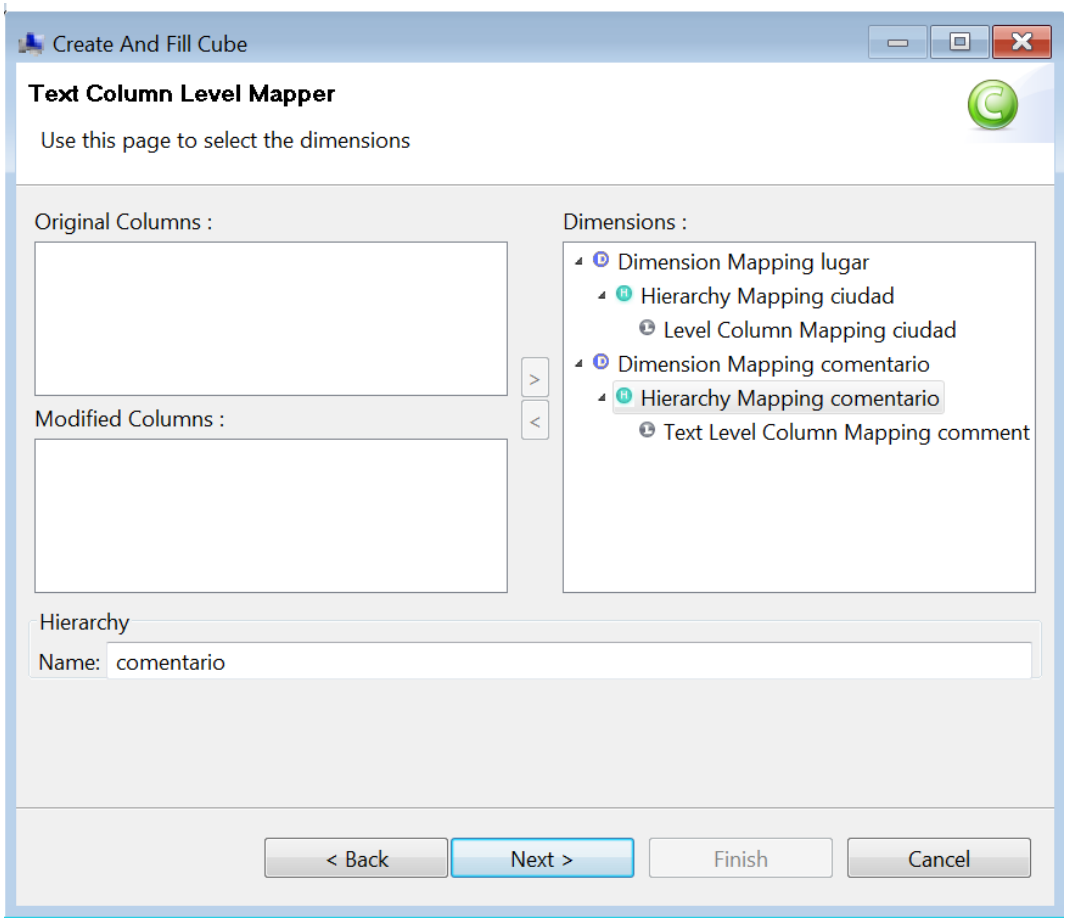


FIGURA 6.7: Definición de las dimensiones-AP

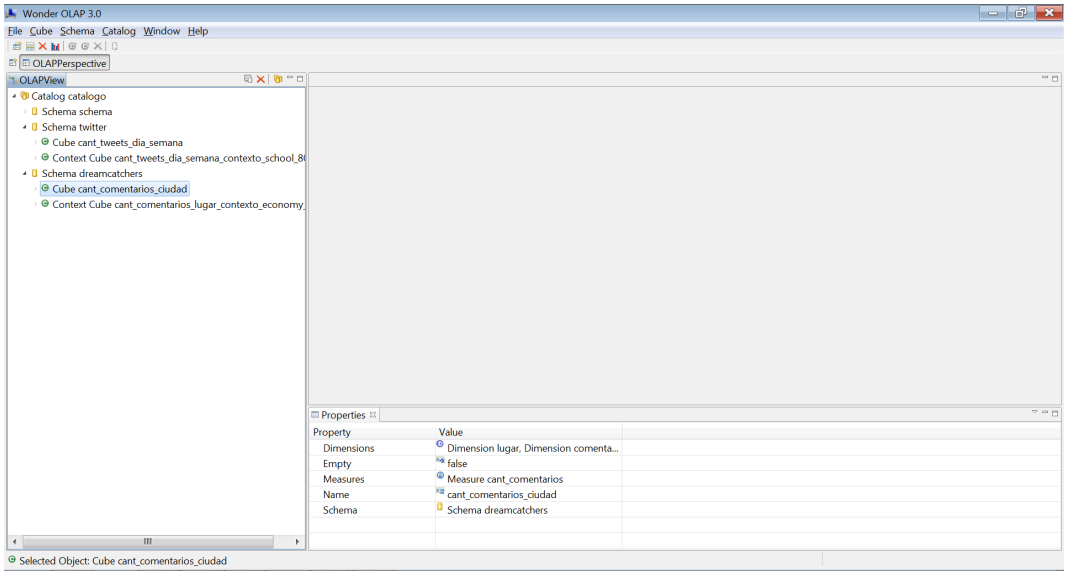


FIGURA 6.8: Vista que permite ver las propiedades del cubo creado

6.2.2. Creación del cubo de datos con dimensión contextual

En esta sección al igual que en la anterior, se describe el proceso de creación de dos cubo de datos con el uso de Wonder. En este caso, los cubos creados a partir de datos de Twitter y Dreamcatchers en lugar de tratar los atributos textuales mediante la dimensión-AP, serán tratados con la dimensión contextual. Para poder luego establecer una comparación entre los cubos con dimensiones-AP y los cubo contextuales, se han seleccionado las mismas dimensiones y los mismos hechos para cada conjunto de datos.

Cubo OLAP para el caso de Twitter

Como se puede apreciar en la Figura 6.9, crear un cubo contextual difiere a la creación de un cubo con dimensiones clásicas y con dimensiones-AP. Aquí además de los atributos que serán seleccionados como dimensiones y hechos, se debe especificar el nivel de la jerarquía de contextos que se desea analizar, así como la cantidad de grupos donde se quiere realizar el corte (en este caso 80) y por último el atributo que contiene las etiquetas asignadas a los textos.

La Figura 6.10 muestra las etiquetas presentes en el atributo seleccionado como etiqueta mediante una nube de etiquetas. Una vez seleccionado el contexto que se desea analizar, es posible continuar con el proceso de creación del cubo contextual. En este caso se ha seleccionado el contexto *SCHOOL*. La Figura 6.11 muestra la interfaz para crear la dimensión contextual, que como se puede apreciar es completamente igual que las dimensiones-AP. Finalmente en la Figura 6.12 podemos ver mediante la vista de propiedades las característica del cubo creado.

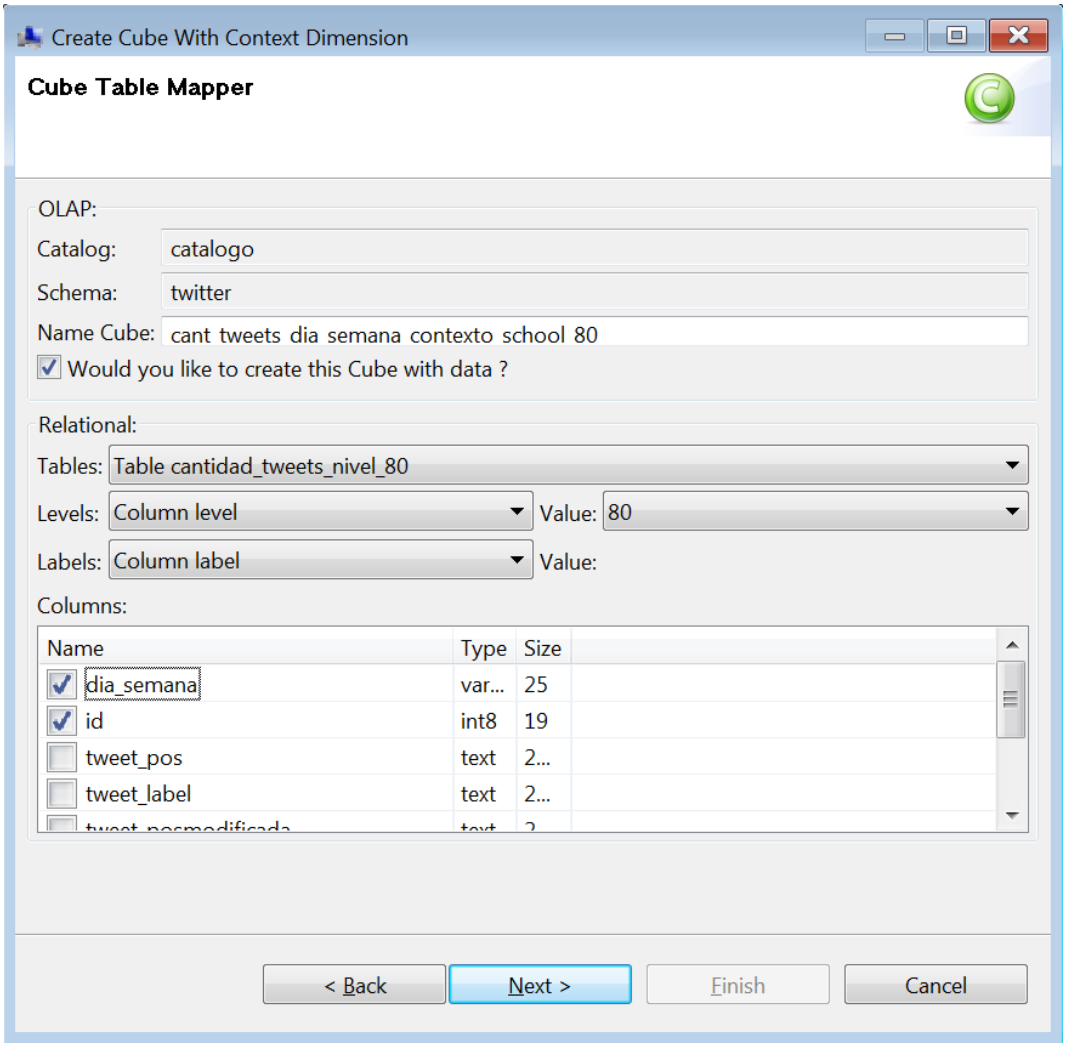


FIGURA 6.9: Selección de los atributos del cubo contextual

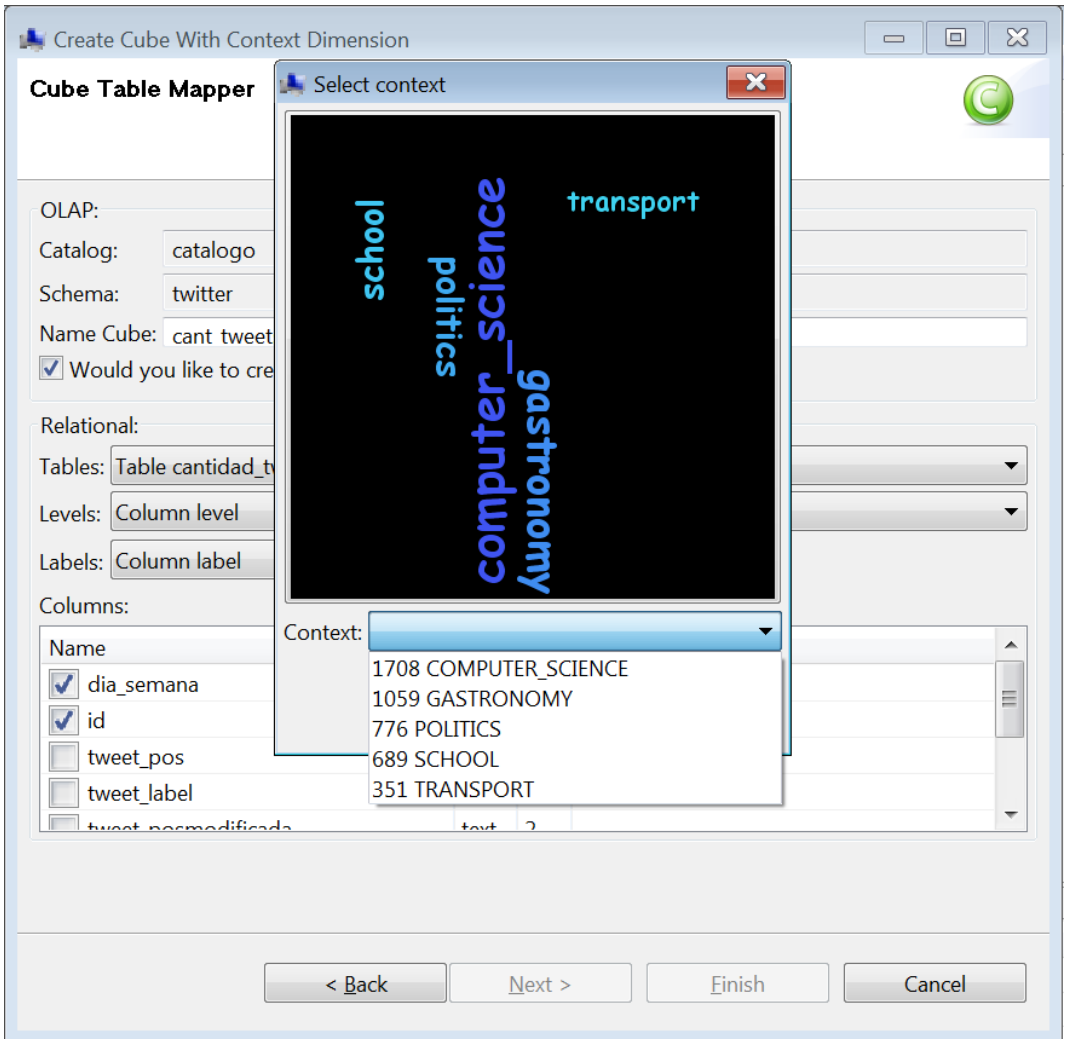


FIGURA 6.10: Nube de etiquetas con los contextos presentes en los datos seleccionados

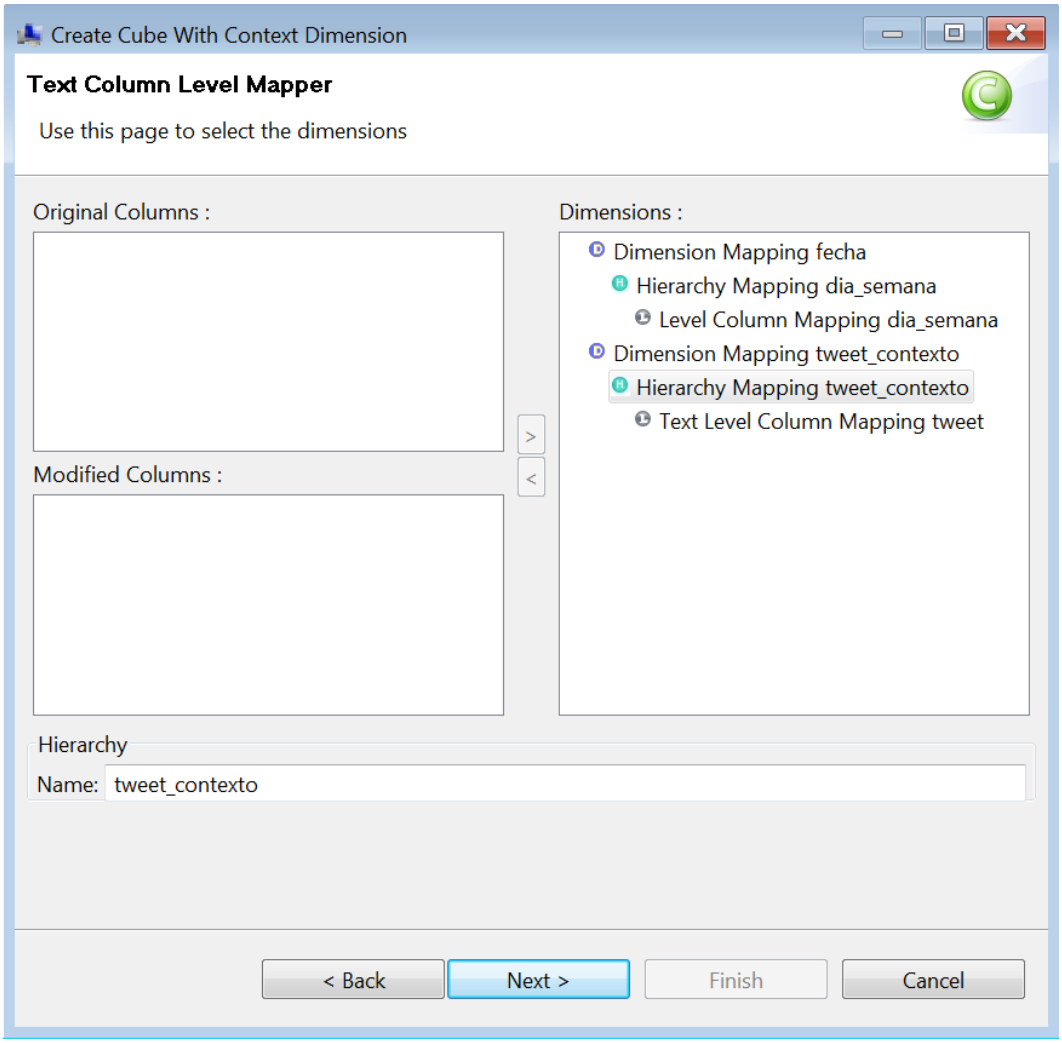


FIGURA 6.11: Definición de las dimensiones contextual

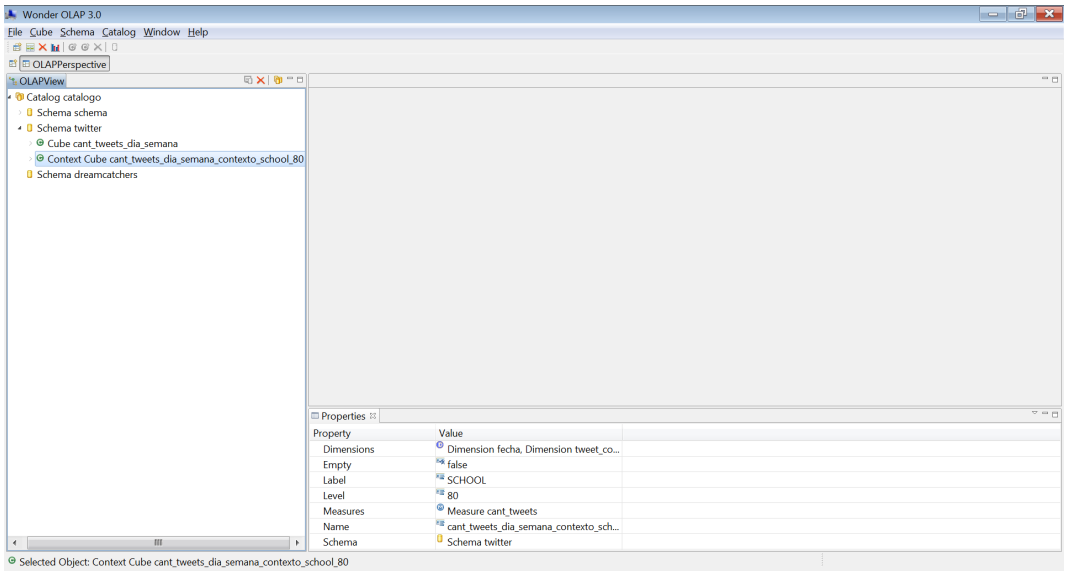


FIGURA 6.12: Vista preliminar del cubo creado

Cubo OLAP para el caso de Dreamcatchers

De igual forma que para el caso anterior, se procede a crear un cubo contextual para datos de Dreamcatchers. La Figura 6.6 muestra los contextos de los datos a analizar. En este caso, se ha seleccionado el contexto *ECONOMY*. La Figura 6.7 permite crear las dimensiones contextuales y la Figura 6.8 muestra la vista final del cubo contextual creado.

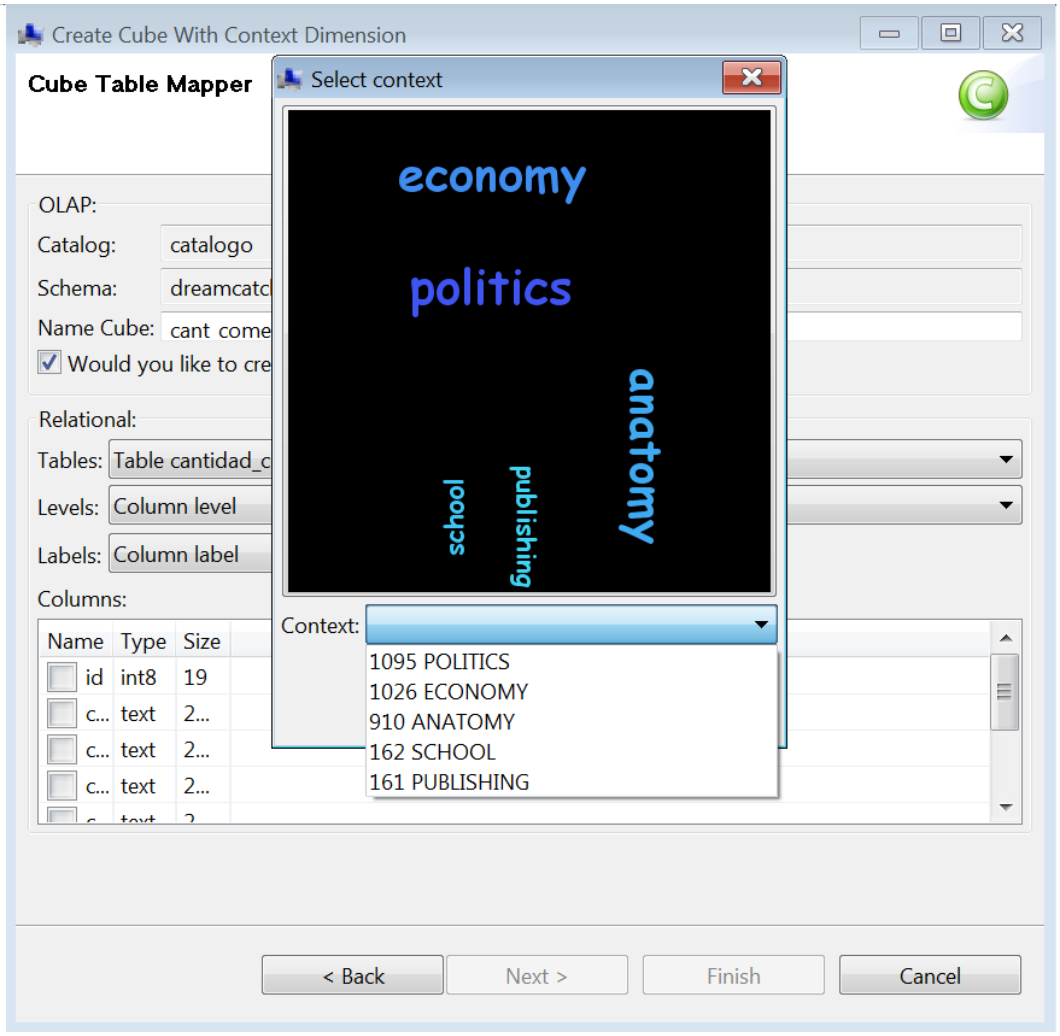


FIGURA 6.13: Nube de etiquetas con los contextos presentes en los datos seleccionados

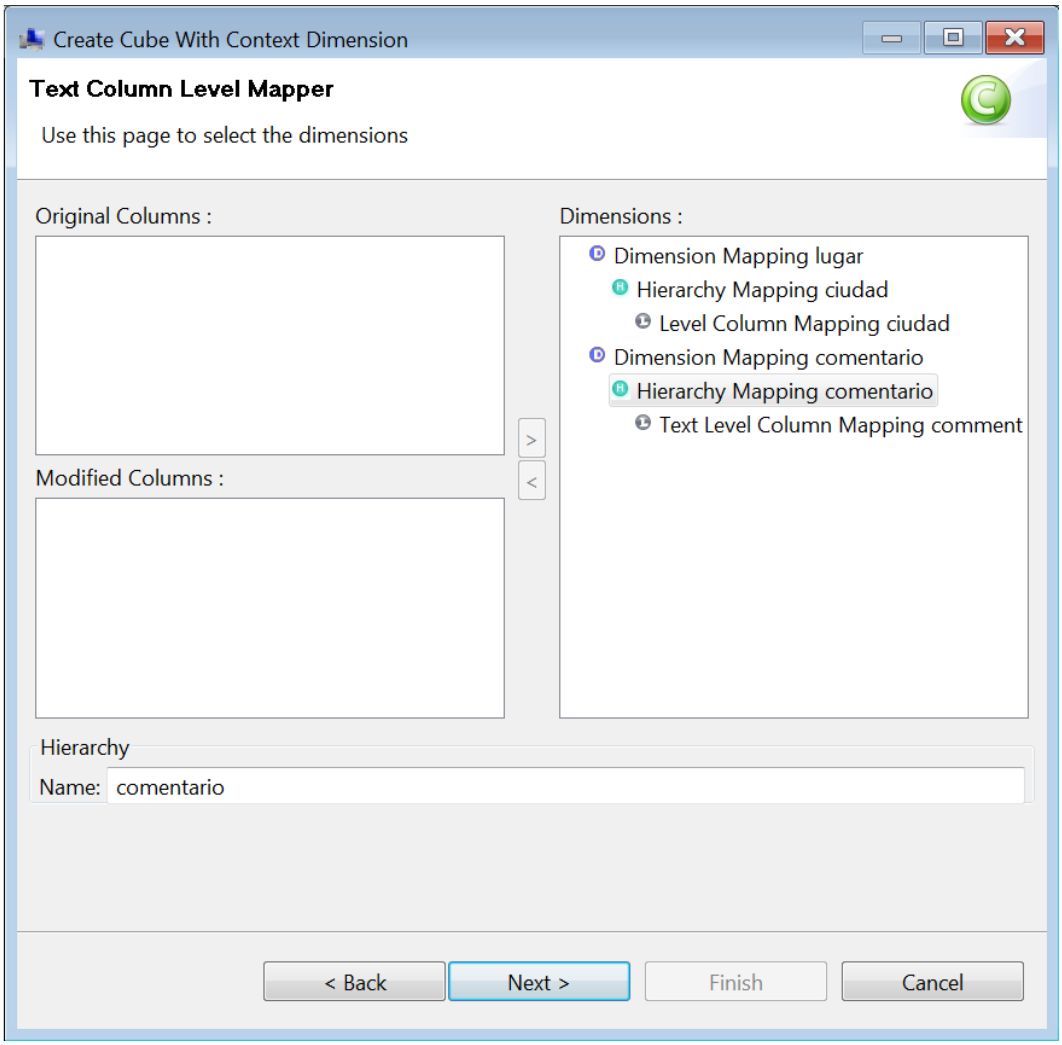


FIGURA 6.14: Definición de las dimensiones contextuales

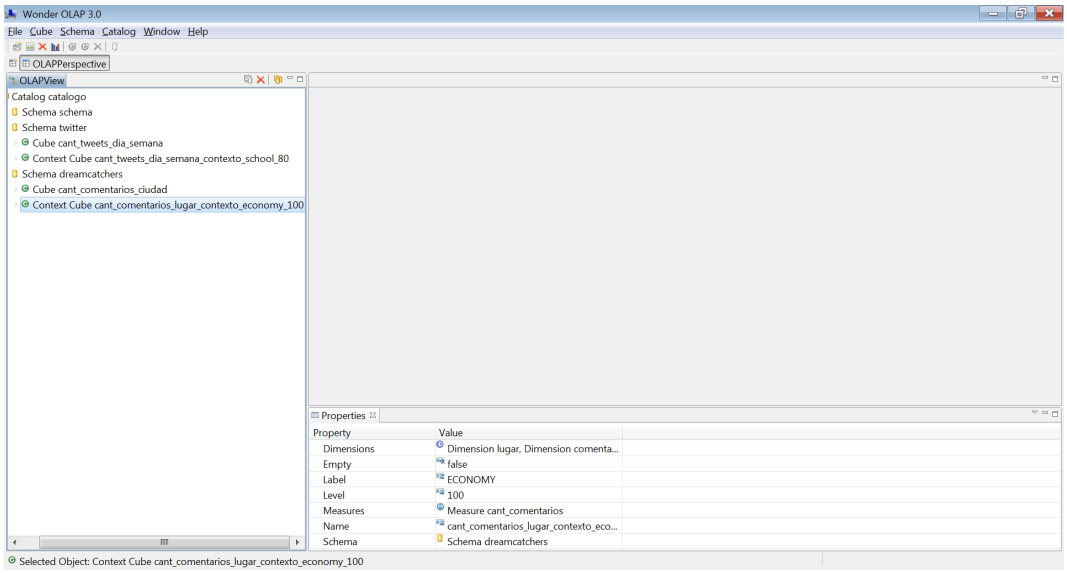


FIGURA 6.15: Vista que permite ver las propiedades del cubo creado

6.3. Análisis multidimensional de Twitter

6.3.1. Descripción de los datos a utilizar

Los datos de Twitter que serán utilizados en esta sección son exactamente los mismos que los utilizados en el ejemplo presentado en en la Sección 5.2.2. La única diferencia es que en este caso de se va a analizar el contexto *SCHOOL* al cual pertenecen 689 tweets. Se va a estudiar la cantidad de tweets por día de la semana que para determinada frase de búsqueda.

6.3.2. Ejemplos de consultas

Primero realizaremos una consulta en el cubo de datos que contiene la dimensión-AP. Luego haremos la misma consulta sobre el cubo contextual. Las Figuras 6.16 y 6.17 muestran la interfaz que permite construir la consulta para cada dimensión. En el caso de la dimensión-AP, se puede apreciar en la lista de operadores las diferentes extensiones de consulta para este tipo de dimensión. La consulta especificada incluye las frases de búsqueda *AFTERNOON*, *AFTERNOON LUNCH*, *ANTHROPOLOGY*, *ANTHROPOLOGY EXAM*, *GOOGLE*, *GOOGLE ORGANIZATION*, *LUNCH* y *ORGANIZATION*.

Por otra parte, la Figura 6.18 muestra la jerarquía de consulta del atributo *tweet*. Como se puede apreciar en este caso los textos no están organizados por contextos, por lo que se hace más engorroso el análisis. Finalmente, la Figura 6.19 muestra el gráfico resultante de la consulta especificada que permite estudiar la cantidad de tweets por día de la semana que contienen las frases de búsqueda especificadas.

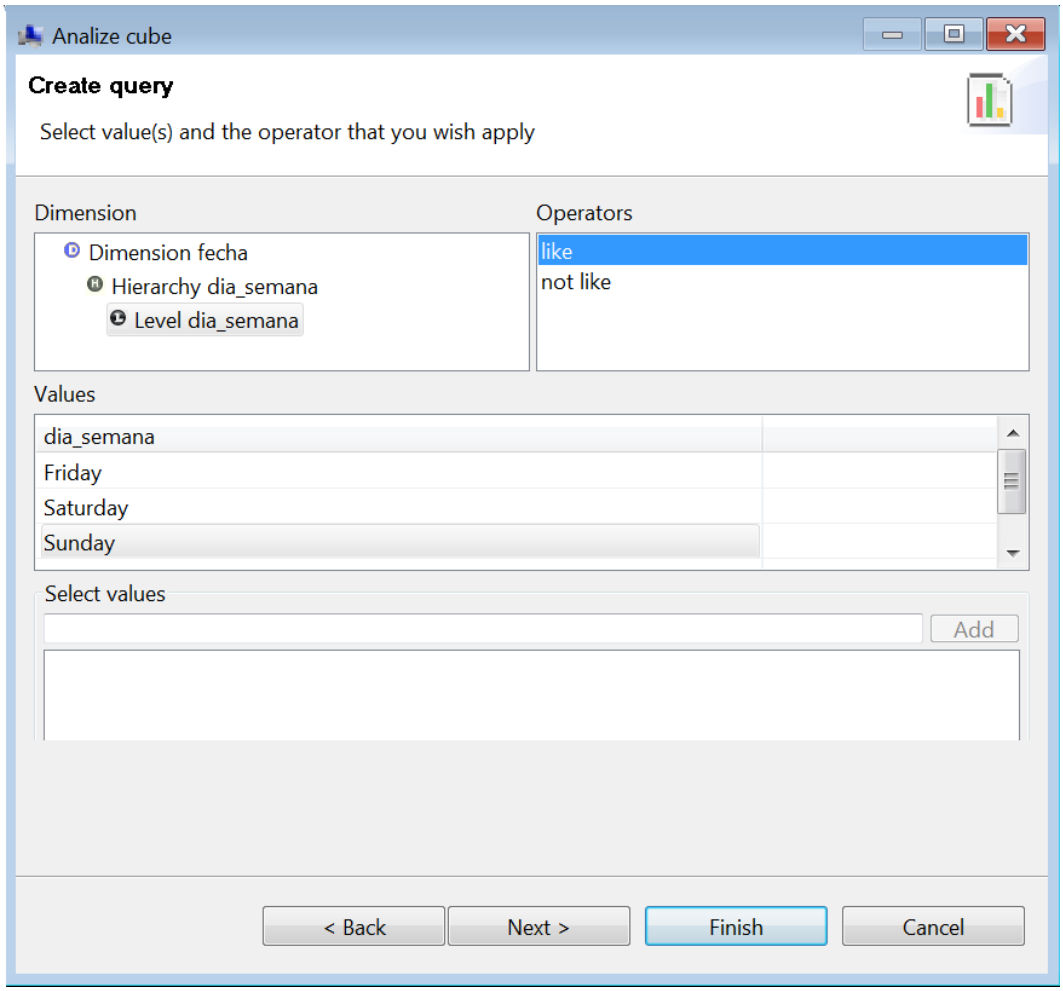


FIGURA 6.16: Vista que permite construir la consulta para una dimensión clásica

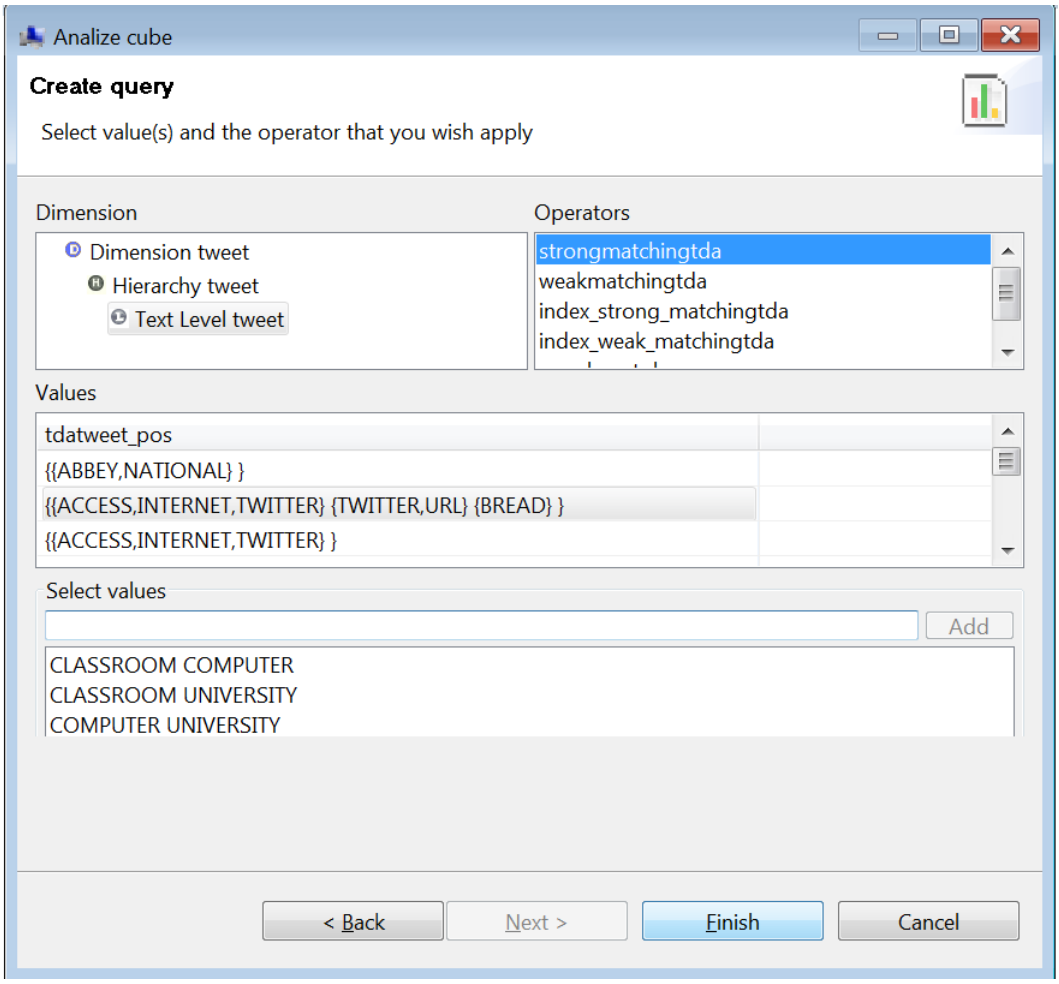


FIGURA 6.17: Vista que permite construir la consulta para una dimensión-AP

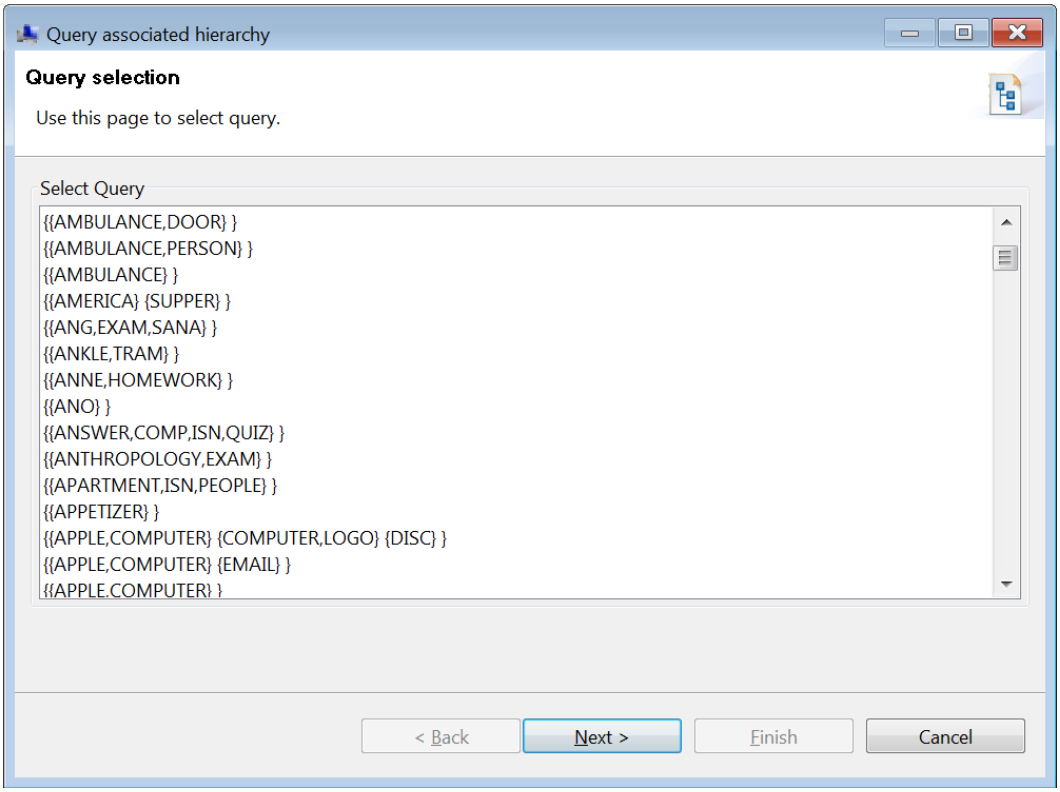


FIGURA 6.18: Vista que muestra la jerarquía de consulta

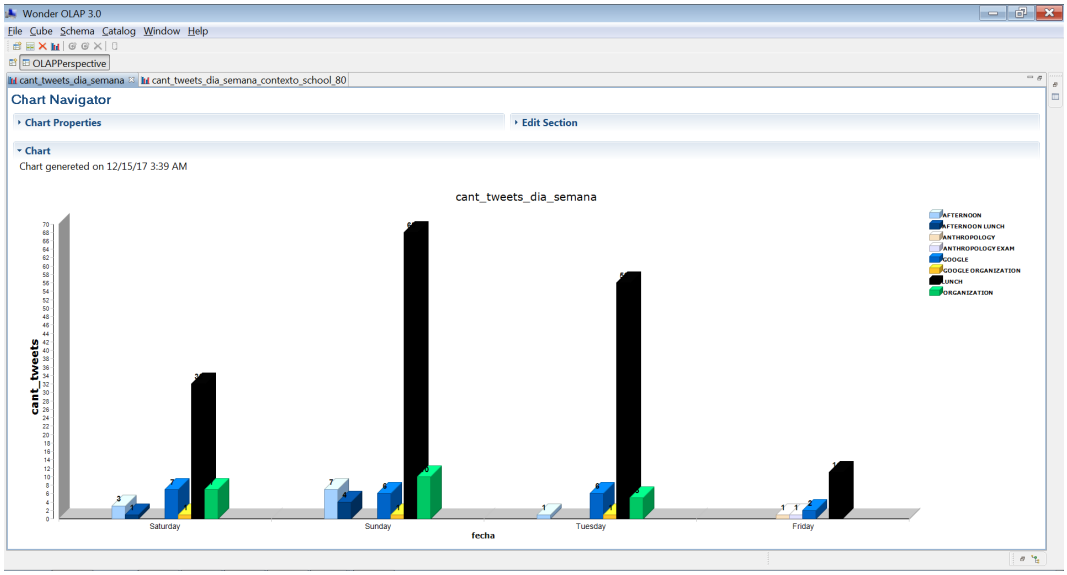


FIGURA 6.19: Gráfico resultante de la consulta realizada sobre el cubo con dimensión-AP

En la Figura 6.20 se puede apreciar la jerarquía de consulta asociada al atributo *tweet* para el caso del cubo contextual. Como es de esperar, en este caso realizar una consulta resulta más fácil, ya que los textos a analizar pertenecen al contexto

SCHOOL. Por esta razón, en la jerarquía de consulta se aprecian mayormente términos relacionados con dicho contexto. La Figura 6.21 muestra los resultados de la consulta sobre este cubo contextual mediante un gráfico de barras.

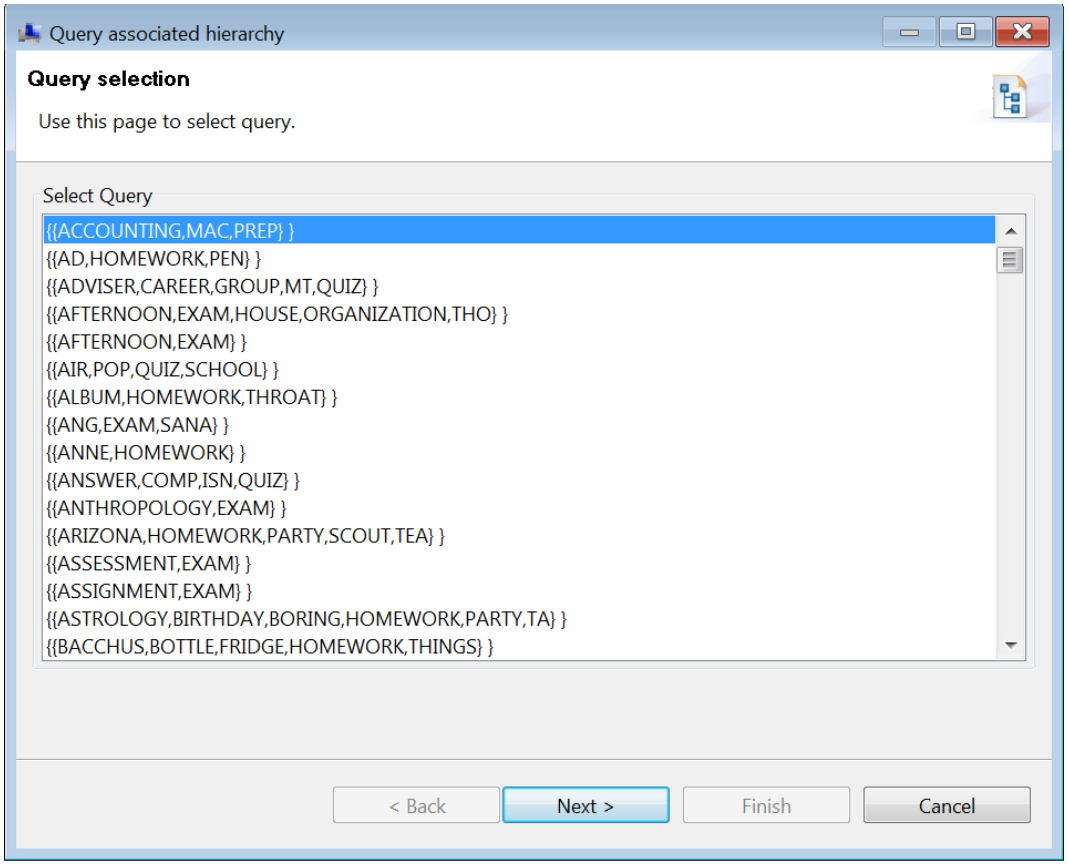


FIGURA 6.20: Vista que muestra la jerarquía de consulta

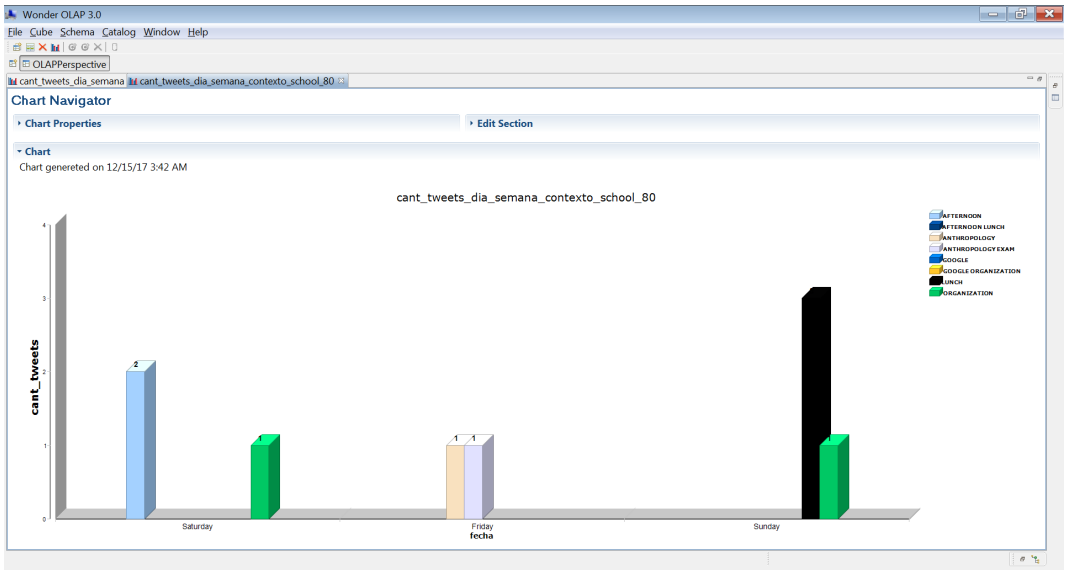


FIGURA 6.21: Gráfico resultante de la consulta realizada sobre el cubo contextual

6.4. Análisis multidimensional de Dreamcatchers

6.4.1. Descripción de los datos a utilizar

Los datos de Dreamcatchers han sido seleccionados aleatoriamente (3354 comentarios). Además contamos con el dato correspondiente a la ciudad donde se ha realizado el comentario. Se desea analizar el contexto *ECONOMY* al cual consta de 1026 comentarios. Se desea estudiar la cantidad de comentarios por ciudad para determinada frase de búsqueda.

6.4.2. Ejemplos de consultas

Al igual que en la sección anterior, primero realizaremos una consulta en el cubo de datos que contiene la dimensión-AP. Luego haremos la misma consulta sobre el cubo contextual.Las Figuras 6.22 y 6.23 muestran la intefaz que permite construir la consulta para cada dimensión. Para la dimensión clásica *ciudad* se han especificado los valores *La Habana, Camagüey, Ciudad de la Habana, Santiago de Cuba y Villa Clara*. La consulta especificada para la dimensión-AP incluye las frases de búsqueda *CORAZÓN, CORAZÓN PECHO, CUENTA, CUENTA RETIRO, ESCUELA, ESCUELA INTERNET, PECHO y RETIRO*.

Por otra parte, la Figura 6.24 muestra la jerarquía de consulta del atributo *comentario*. De igual forma que en el ejemplo de Twitter los textos no están organizados por contextos, por lo que se hace más engorroso el análisis. La Figura 6.25 muestra el gráfico resultante de la consulta especificada que permite estudiar la cantidad de comentarios por ciudad que contienen las frases de búsqueda especificadas.

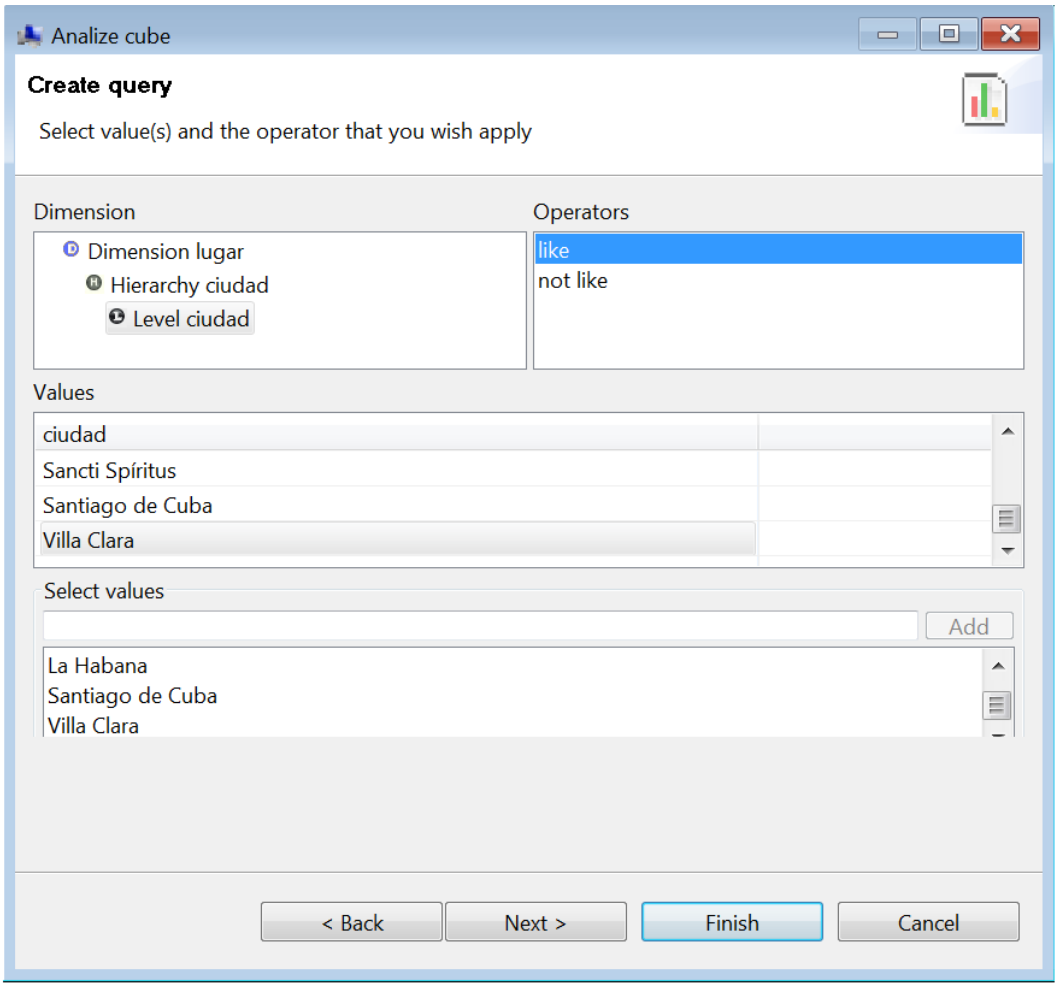


FIGURA 6.22: Vista que permite construir la consulta para una dimensión clásica

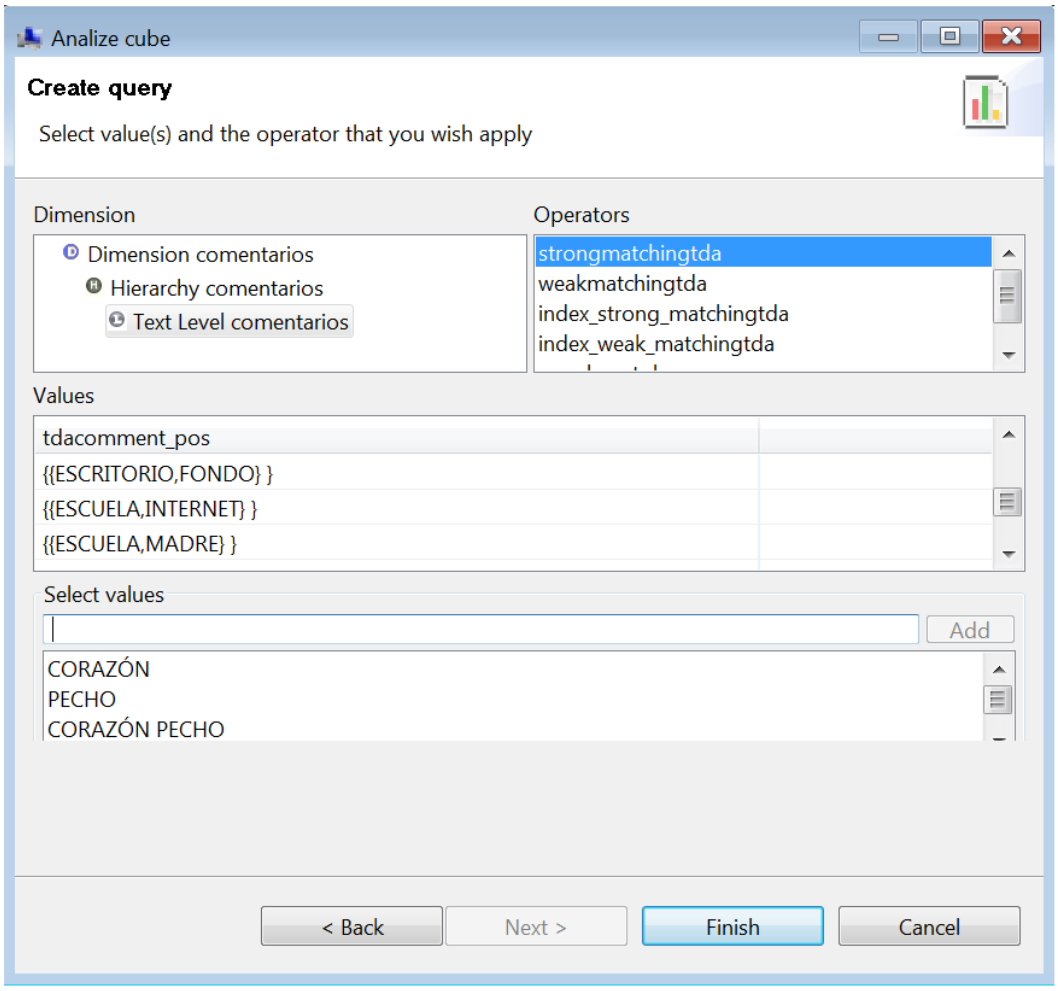


FIGURA 6.23: Vista que permite construir la consulta para una dimensión-AP

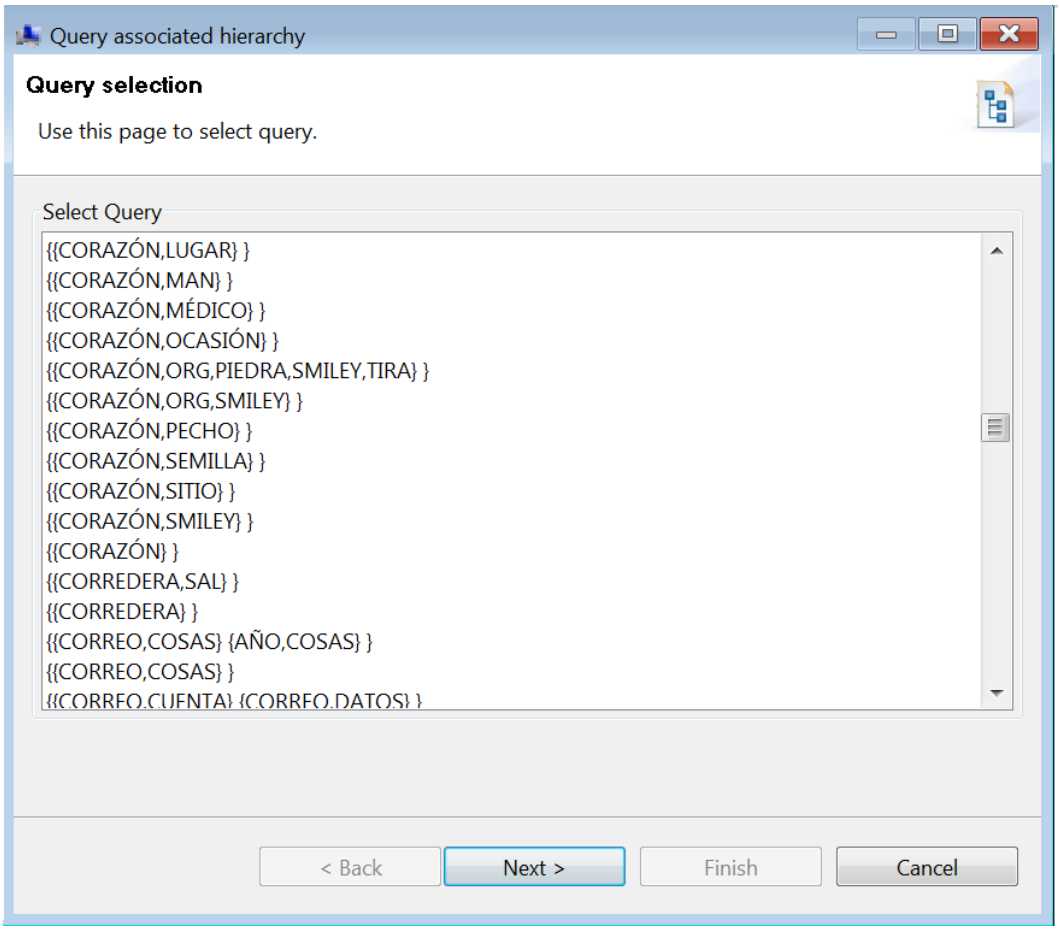


FIGURA 6.24: Vista que muestra la jerarquía de consulta

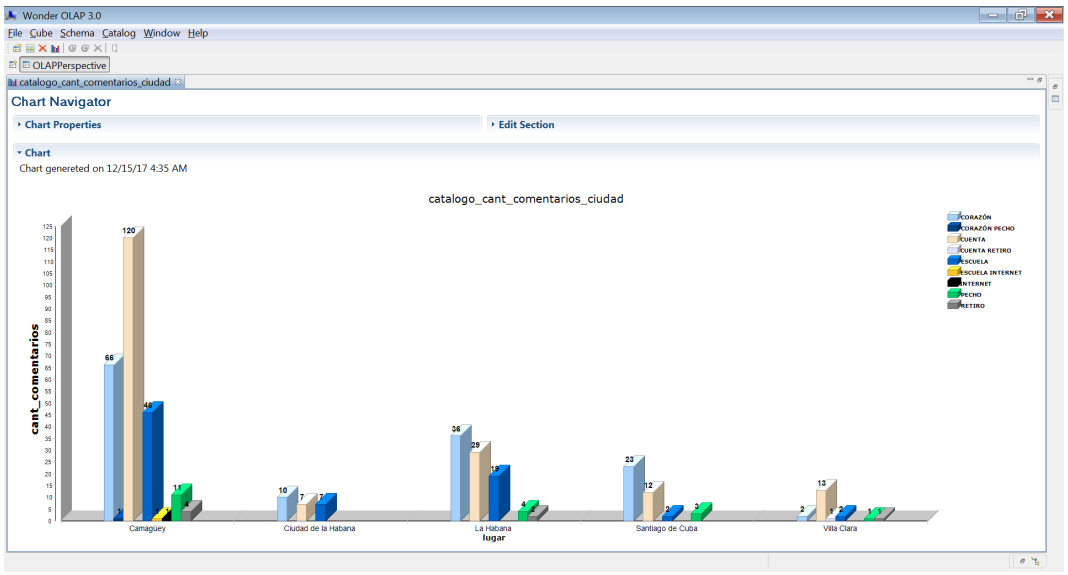


FIGURA 6.25: Gráfico resultante de la consulta realizada sobre el cubo con dimensión-AP

En la Figura 6.26 se puede apreciar la jerarquía de consulta asociada al atributo *comentario* para el caso del cubo contextual. En este caso e igual que para Twitter,

realizar una consulta resulta más fácil, ya que los textos a analizar pertenecen al contexto *ECONOMY* Figura 6.26. La Figura 6.27 muestra los resultados de la consulta sobre este cubo contextual mediante un gráfico de barras.

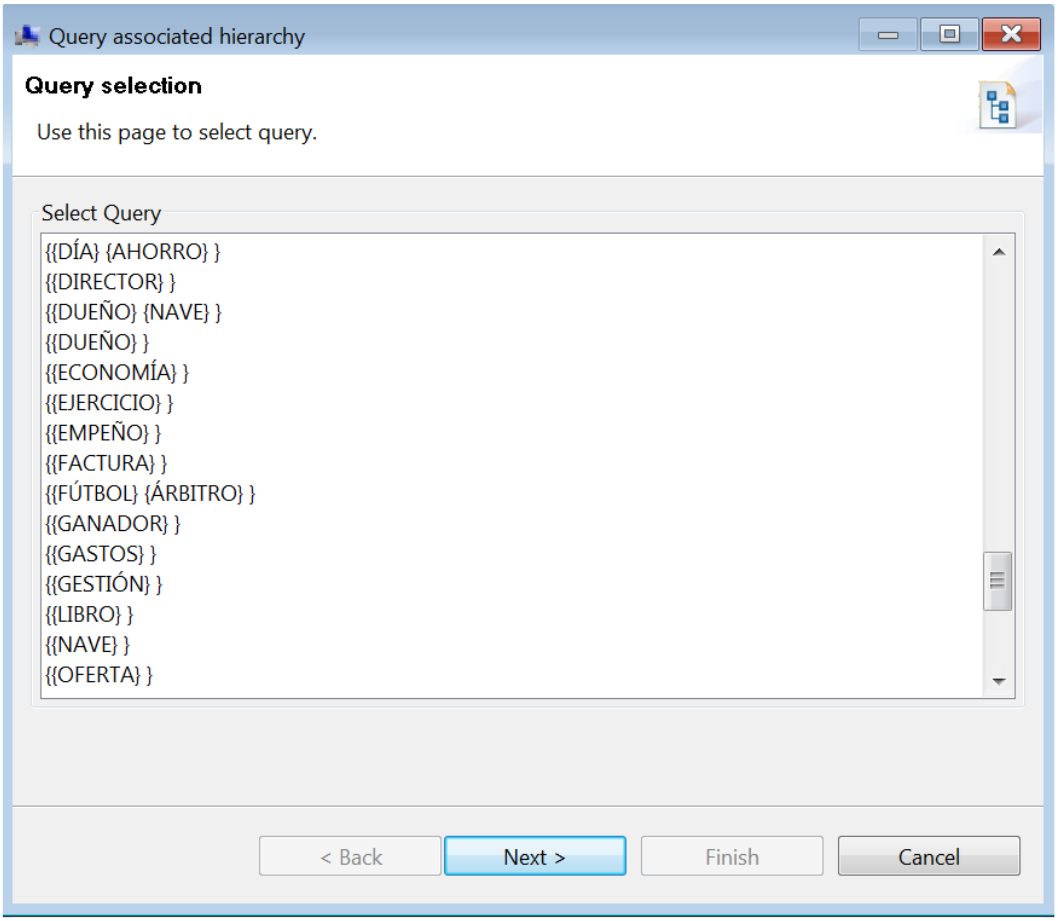


FIGURA 6.26: Vista que muestra la jerarquía de consulta

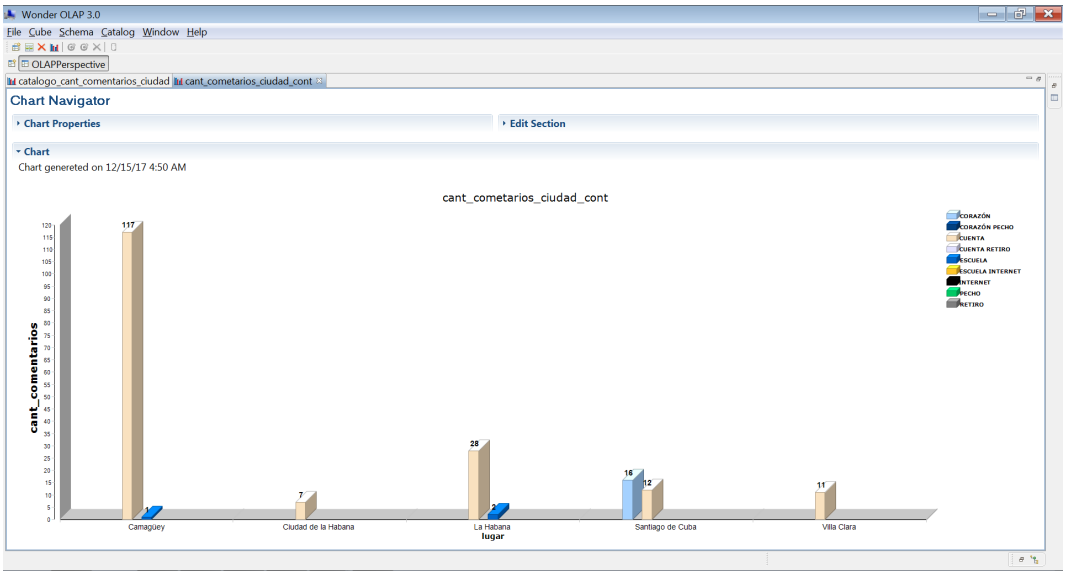


FIGURA 6.27: Gráfico resultante de la consulta realizada sobre el cubo contextual

6.5. Conclusiones

En el presente capítulo se ha llevado a cabo la explotación de la dimensión contextual en un modelo multidimensional con el fin de demostrar las facilidades que proporciona dicha dimensión para el análisis de datos textuales. Para la construcción de la dimensión contextual se han utilizado datos reales de dos redes sociales (Twitter y Dreamcatchers) y como servidor OLAP para la integración y análisis de la dimensión, Wonder 3.0.

Se ha demostrado mediante ejemplos los beneficios que brinda crear un cubo de datos que modele los textos como una dimensión contextual en lugar de tratarlos directamente como una dimensión clásica para su posterior análisis. Mediante una serie de consultas, hemos expuesto las diferentes posibilidades de consulta que ofrece Wonder 3.0 sobre una dimensión contextual, así como la salida de datos del sistema mediante gráficos y reportes.

Capítulo 7

Conclusiones y trabajos futuros

En el presente capítulo hemos resumido los objetivos alcanzados durante el desarrollo de este trabajo y que han sido expuestos en la presente memoria. Luego, se exponen ciertas líneas de investigación que en el futuro complementarán los resultados obtenidos y aquí presentados.

7.1. Conclusiones

Al comienzo de este trabajo se planteó el objetivo fundamental, el cual consiste en tratar y analizar de forma automática los datos textuales de redes sociales conjuntamente con datos tradicionales mediante un modelo multidimensional.

De forma general, para conseguir estos objetivos, se ha profundizado en el estudio de las técnicas y herramientas necesarias para desarrollar la metodología propuesta. Se han detectado de forma automática los principales contextos presentes en textos de cualquier índole. También se ha analizado la influencia que tienen los términos de sentimientos en el proceso de detección de contextos en textos de redes sociales. Además, una vez agrupados o segmentados los textos por contextos, se ha construido una dimensión contextual para facilitar el análisis de los textos de redes sociales mediante un sistema OLAP. Finalmente, se han desarrollado las herramientas de software necesarias que permiten realizar de forma automática todo lo anteriormente expuesto, así como la realización de consultas complejas sobre atributos textuales de redes sociales.

Se debe destacar, que nuestra solución no sólo resuelve las necesidades del problema de investigación, pues la metodología propuesta puede ser aplicada en otros contextos y ayudar a resolver el problema relacionado con el tratamiento y análisis

de atributos textuales. De forma más detallada, la labor realizada queda descrita a continuación:

1. *Se han detectado de forma automática los principales contextos abordados en datos textuales.*

   - Para ello, se ha definido una metodología basada en el uso de una ontología (WordNet Domains). Esto permite a los usuarios organizar los textos automáticamente en grupos (contextos) sin tener información previa, donde cada grupo constituye un conjunto de tópicos.
   - La ontología utilizada permite que la metodología pueda ser aplicada en diferentes dominios sin importar los temas abordados en los textos, ya que dicha ontología abarca casi en su totalidad las áreas del conocimiento humano. Además, el uso del recurso léxico multilingüe MCR 3.0 hace a nuestra metodología independiente del idioma.
   - El resultado final de la metodología desarrollada es una jerarquía de contextos. Esta jerarquía se obtiene como resultado del uso algoritmos de agrupamiento jerárquicos y de aplicar cortes en el dendograma construido por dichos algoritmos. La jerarquía obtenida permite realizar análisis más elaborados (al tener los textos previamente organizados por contextos o temáticas) y en diferentes niveles de abstracción.
   - Se ha experimentado con seis conjuntos de datos (tres en inglés y tres en español) y con tres algoritmos de agrupamiento jerárquico (Complete Link, Average y Ward). Para cada caso se evaluó la calidad de los algoritmos de agrupamiento cuando se aplicó la metodología propuesta basada en el uso de la ontología WordNet Domains y sin aplicarla. Con este fin se utilizó como medida el Coeficiente de Silueta, la cual es una medida no supervisada y que tiene en cuenta tanto la cohesión como la separación entre los grupos creados. En cada caso se demostró que siempre que se aplicaba la metodología los resultados obtenidos eran mejores que cuando no era aplicada.

2. *Se ha analizado la influencia que tienen los términos de sentimientos en la detección automática de contextos en textos de redes sociales.*

- Con este fin, se ha incorporado un filtro durante el preprocesamiento semántico de los textos de la metodología propuesta para la detección de contextos, el cual permite detectar y eliminar los términos de sentimientos, ya que dichos términos influyen de forma negativa en la detección de contextos. Para ello, se han utilizado los recursos léxicos SentiWordNet 3.0, SenticNet 3 y WordNet Affect.
- Al descartar todos los términos con orientación sentimental, se pueden descartar una gran cantidad de documentos los cuales no formarán parte del proceso de detección de contextos, pues en textos de redes sociales los usuarios suelen expresar su opinión sobre diversos temas. Cuando se eliminan un gran número textos, esto puede influir en el proceso de detección de contextos. Por tal motivo, además de eliminar todos los términos de sentimientos, se han establecido siete umbrales para eliminar aquellos términos cuya polaridad sea superior al umbral especificado. Esto va a permitir establecer un consenso entre el valor de la medida de calidad de grupos creados por los algoritmos de agrupamiento jerárquico (Coeficiente de Silueta) y la cantidad de textos a analizar.
- Se ha experimentado con ocho conjuntos de datos de las redes sociales Twitter y Dreamcatchers. En cada caso, los resultados cuando el filtro es aplicado mejoran considerablemente en comparación cuando no es aplicado, destacando sobremanera el recurso SenticNet 3. Finalmente, se logró establecer un equilibrio entre la cantidad de textos analizados y el Coeficiente de Silueta gracias a los umbrales de polaridad establecidos. Se debe destacar que esto último sólo es posible aplicarlo con los recursos SentiWordNet 3.0 y SenticNet 3 los cuales brindan la polaridad de los términos.

3. *Se ha construido e integrado una dimensión contextual en un modelo multidimensional, facilitando el análisis de los textos de redes sociales mediante dicha dimensión en conjunto con dimensiones clásicas.*

   - Mediante el uso de la metodología propuesta para la detección de contextos, basada en el uso de una ontología, se ha obtenido una jerarquía de contextos la cual permite agrupar los textos por los principales temas abordados. Luego para cada contexto de dicha jerarquía, se constituye

una jerarquía de consulta o dominio la cual contiene los principales tópicos relacionados al contexto en cuestión. Esta estructura formada por la jerarquía de contextos y la jerarquía de consulta asociada a cada contexto le hemos llamado *dimensión contextual*.

- Dicha dimensión es integrada en un modelo multidimensional permitiendo así el análisis de los datos textuales por contextos y tópicos de igual forma y conjuntamente con las dimensiones convencionales. Para la integración ha sido extendido el modelo multidimensional (extensiones de almacenamiento y consulta) implementado por el sistema OLAP utilizado (Wonder 3.0).
- Se ha desarrollado un ejemplo con datos reales de Twitter, y se utilizó el sistema Wonder 3.0 como servidor OLAP. Los resultados demuestran las posibilidades de consultas que brinda la dimensión contextual para el análisis multidimensional de los datos textuales de redes sociales integrados con datos clásicos.

4. *Se ha llevado a cabo la explotación de la dimensión contextual.*

   - Se han utilizado datos reales de las redes sociales Twitter y Dreamcatchers y como servidor OLAP para la integración y análisis de la dimensión, Wonder 3.0.
   - Se han explicado los beneficios que brinda crear un cubo de datos que modele los textos como una dimensión contextual en lugar de tratarlos directamente como una dimensión clásica. Mediante una serie de consultas, se ha demostrado las diferentes posibilidades de consulta que ofrece Wonder 3.0 sobre una dimensión contextual.

7.2. Trabajos futuros

El trabajo descrito en la presente memoria, constituye un punto de partida para un conjunto de tareas a desarrollar en trabajos futuros. Muchas de ellas no han sido tratadas por encontrarse fuera de los objetivos planteados, y otras han aparecido como consecuencia de las propuestas realizadas. A continuación, veremos las líneas de investigación que hemos considerado más interesantes para futuras aportaciones.

- **Extender la metodología propuesta para la Detección de Contextos en un entorno Big Data.**
  En el presente trabajo se han analizado conjuntos de datos de hasta 30000 documentos. Debido a la gran complejidad computacional de los algoritmos de agrupamiento, resulta de gran interés paralelizar dicho proceso y de esta forma mejorar el rendimiento de la metodología.

- **Realizar un estudio de la metodología en casos de Streaming.**
  En esta línea los trabajos futuros estarán orientados a analizar el comportamiento del enfoque propuesto a partir de datos textuales obtenidos en tiempo real. Lo cual añade la dificultad de tener que asignar un contexto a un texto.

- **Aplicar la metodología para la Detección de Contextos en entornos concretos como el hospitalario, turístico, etc.**
  Desde este punto de vista, si analizamos textos pertenecientes a un dominio específico contaríamos con un vocabulario más restringido, lo cual podría mejorar el resultado de la propuesta desde el punto de vista semántico. Además, si a esto le sumamos la inclusión de una ontología de dominio, esto mejoraría considerablemente el resultado final.

Apéndice A

Sistema para la construcción de la dimensión contextual

A.1. Descripción general del sistema

A partir de un conjunto de datos textuales, este sistema permite construir una dimensión contextual. Los textos pueden proceder tanto de ficheros como de bases de datos. Una vez creada la dimensión contextual, es almacenada en una base de datos en PostgreSQL. Además del texto analizado la dimensión está compuesta por la etiqueta correspondiente al contexto que pertenece, el nivel de la jerarquía de contextos seleccionado, así como la estructura-AP inducida para cada texto.

A.2. Requisitos funcionales del sistema

A continuación se describen los principales funcionalidades del sistema:

1. Iniciar sesión en el sistema.
2. Gestionar usuarios del sistema.
3. Crear un proyecto.
4. Crear una instancia de un proyecto.
5. Ejecutar una instancia.
6. Determinar la cantidad de grupos para la cual el algoritmo de agrupamiento utilizado brinda el mejor resultado.
7. Crear jerarquía de contextos.

8. Etiquetar cada contexto con la etiqueta de WordNet Domains más representativa entre los textos del contexto.

9. Crear jerarquía de consulta para cada contexto de una jerarquía de contextos.

10. Guardar la dimensión contextual creada.

A.3. Arquitectura del sistema

La Figura A.1 muestra la arquitectura del sistema web para la construcción automática de la dimensión contextual. Es un sistema web creado bajo mediante el patrón de arquitectura Modelo-Vista-Controlador (MVC). Se han utilizado los siguientes frameworks:

- **Spring Boot**, **Spring MVC**, **Spring data-mongo** y **Spring Security** para crear de forma rápida y segura una aplicación web MVC en Java con acceso a datos en MongoDB,
- **JQuery UI** como framework para crear las intarfaces de usuario y
- **MongoDB** para almacenar los metadatos del sistema.

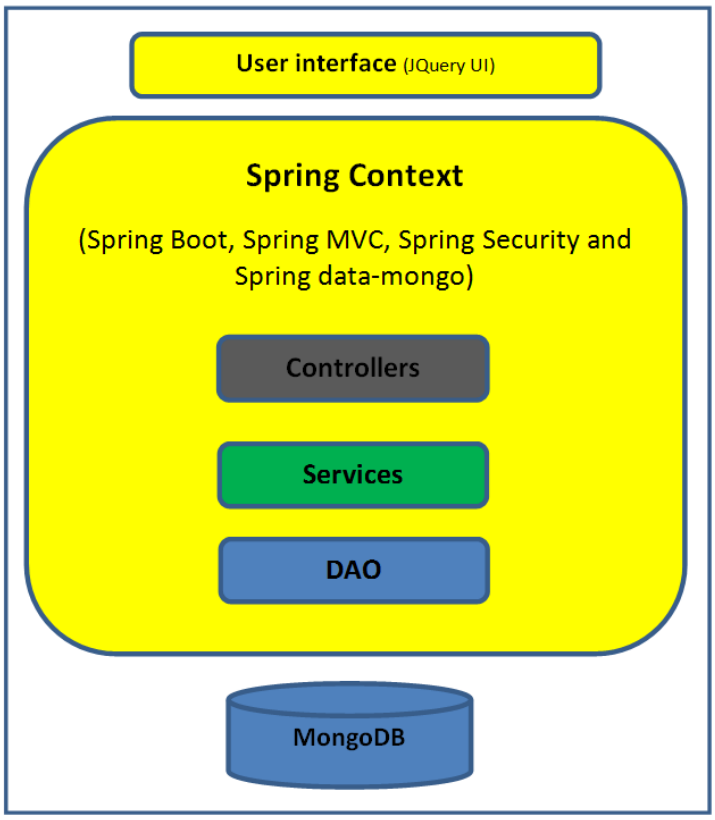


FIGURA A.1: Arquitectura del sistema web para la construcción automática de la dimensión contextual.

A.4. Modelo de datos

En la Figura A.2 se muestra el modelo de datos del sistema web para la construcción automática de la dimensión contextual.

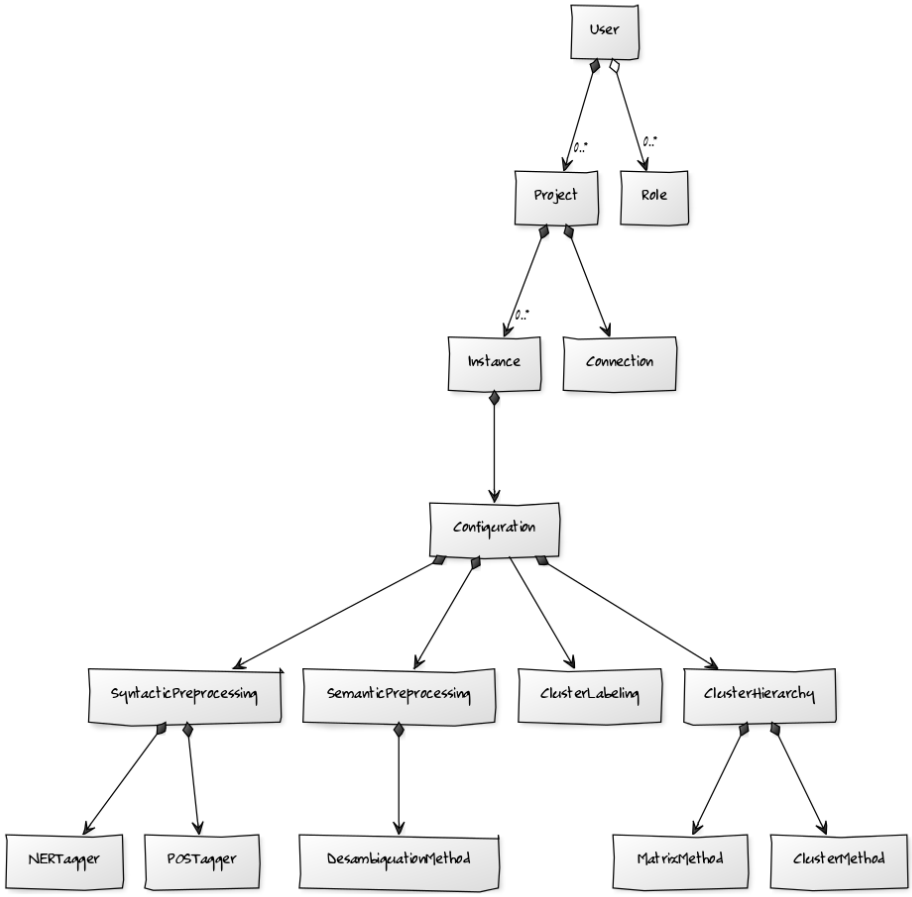


FIGURA A.2: Modelo del sistema web para la construcción automática de la dimensión contextual.

Apéndice B

Herramientas utilizadas

A continuación se explicarán cada una de las herramientas utilizadas por la metodología para la detección automática de contextos propuesta en el presente trabajo.

B.1. Stanford Part-of-Speech Tagger

Permite a partir de un texto asignar categorías gramaticales a cada palabra, como sustantivo, adjetivo, verbo, etc., incluye modelos para los idiomas Árabe, Chino, Francés, Español, y Alemán, además del Inglés donde es considerado actualmente el etiquetador gramatical con mayor precisión [Toutanova et al., 2003].

B.2. Stanford Named Entity Recognition Tagger

Esta herramienta permite etiquetar aquellas palabras de un texto que representan nombres o cosas como personas, compañías, lugares, etc., a partir de varios modelos proporcionados en la distribución. Para el presente trabajo se ha utilizado el modelo basado en tres clases (*Persona, Organización y Lugar*) para los idiomas Inglés y Español [Finkel et al., 2005]. Ambas herramientas han sido desarrolladas por *The Stanford Natural Language Processing Group* de la Universidad de Stanford, implementadas en Java y bajo la licencia GNU General Public License.

B.3. Multilingual Central Repository 3.0

MCR 3.0 está basado en WordNet 3.0 e integra WordNets de cinco idiomas, entre ellos Inglés y Español los cuales son de interés para el presente trabajo, haciendo la metodología independiente del idioma. Además MCR 3.0 integra recursos léxicos como: WordNet Domains [Magnini and Cavaglia, 2000], una nueva versión

de Base Concepts, Top Ontology [Álvez et al., 2008], y la ontología AdimenSUMO [Pease et al., 2002].

B.3.1. WordNet Domains

Es un recurso léxico creado de forma semiautomática para dotar a WordNet con etiquetas de dominios. Cada sentido (synset) de WordNet es anotado con al menos una etiqueta de dominio de WordNet Domains, las cuales están organizadas jerárquicamente [Magnini and Cavaglia, 2000].

WordNet Domains permite reducir el grado de polisemia de las palabras, ya que puede agrupar en un dominio a sentidos que contienen un mismo término [Magnini et al., 2002]. Dada estas características, consideramos que WordNet Domains se ajusta perfectamente al interés del presente trabajo para homogeneizar la representación de los textos y crear una jerarquía semántica de tópicos.

B.4. BabelNet

BabelNet es un diccionario enciclopédico multilingüe, con una cobertura lexicográfica y enciclopédica de términos, y una red semántica que conecta conceptos y entidades en una gran red de relaciones semánticas. contiene cerca de 14 millones de entradas, llamadas sentido (BabelNel synsets). Cada sentido de BabelNet representa un significado y contiene todos los sinónimos de ese significado en varios idiomas

BabelNet brinda soporte para 271 idiomas y se obtiene de la integración automática de 15 recursos, entre los cuales se encuentra WordNet y Wikipedia. Además, contiene traducciones obtenidas a partir de sentencias anotadas por los sentidos.

Bibliografía

[Agirre et al., 2012] Agirre, A. G., Laparra, E., and Rigau, G. (2012). Multilingual central repository version 3.0. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of International Conference on Very Large Data Bases (VLDB' 94)*, Santiago, Chile.

[Allan et al., 1998a] Allan, J., Carbonell, J., and Doddington, G. (1998a). Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

[Allan et al., 1998b] Allan, J., Papka, R., and Lavrenko, V. (1998b). On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 37–45, New York, NY, USA. ACM.

[Álvez et al., 2008] Álvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A., and Rigau, G. (2008). Complete and consistent annotation of wordnet using the top concept ontology. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[Andrea and Fabrizio, 2006] Andrea, E. and Fabrizio, S. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, pages 417–422.

[Avila et al., 2011] Avila, E., Miranda, M., and Bautista, M. (2011). *Datawarehousing Con Procesamiento de Datos Textuales*. EAE.

[Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

[Banea et al., 2008] Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii.

[Bauer and King, 2006] Bauer, C. and King, G. (2006). *Java Persistence with Hibernate*. Manning Publications Co., Greenwich, CT, USA.

[Bringay et al., 2011] Bringay, S., Béchet, N., Bouillot, F., Poncelet, P., Roche, M., and Teisseire, M. (2011). Towards an on-line analysis of tweets processing. In *Proceedings of the 22nd International Conference on Database and Expert Systems Applications - Volume Part II*, DEXA'11, pages 154–161, Berlin, Heidelberg. Springer-Verlag.

[Bull et al., 2004] Bull, R. I., Best, C., and Storey, M.-A. (2004). Advanced widgets for eclipse. In *Proceedings of the 2004 OOPSLA Workshop on Eclipse Technology eXchange*, eclipse '04, pages 6–11, New York, NY, USA. ACM.

[Cambria et al., 2014] Cambria, E., Olsher, D., and Rajagopal, D. (2014). Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *AAAI Conference on Artificial Intelligence*.

[Cecchet et al., 2004] Cecchet, E., Marguerite, J., and Zwaenepoel, W. (2004). C-jdbc: Flexible database clustering middleware. In *In Proceedings of the USENIX 2004 Annual Technical Conference*, pages 9–18.

[Chung-Hong, 2012] Chung-Hong, L. (2012). Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams. *Expert Systems with Applications*, 39(18):pp. 13338–13356.

[Codd et al., 1993] Codd, E. F., Codd, S. B., and Salley, C. T. (1993). Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. E. F. Codd and Associates.

[Deshmukh et al., 2013] Deshmukh, D., Kamble, S., and Dandekar, P. (2013). Survey on hierarchical document clustering techniques fihc & f2 ihc. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(7):pp. 157–161.

[Duan and Zeng, 2013] Duan, J. and Zeng, J. (2013). Web objectionable text content detection using topic modeling technique. *Expert Systems with Applications*, 40:6094–6104.

[Finkel et al., 2005] Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Galhardas et al., 2001] Galhardas, H., Florescu, D., Shasha, D., Simon, E., and Saita, C. (2001). Improving data cleaning quality using a data lineage facility. In Theodoratos, D., Hammer, J., Jeusfeld, M. A., and Staudt, M., editors, *Proceedings of the 3rd Intl. Workshop on Design and Management of Data Warehouses, DMDW'2001, Interlaken, Switzerland, June 4, 2001*, volume 39 of *CEUR Workshop Proceedings*, page 3. CEUR-WS.org.

[Gao et al., 2013] Gao, N., Gao, L., He, Y., Wang, H., and Sun, Q. (2013). Topic detection based on group average hierarchical clustering. In *International Conference on Advanced Cloud and Big Data (CBD, 2013)*, pages 88–92. IEEE.

[Guille et al., 2013] Guille, A., Hacid, H., Favre, C., and Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *SIGMOD Record*, 42(2).

[Kruskal and Wallis, 1952] Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, pages 583–621.

[Lesk, 1986] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

[Lewis et al., 2004] Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

[Lin and He, 2009] Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *18th ACM Conference on Information and Knowledge Management 8CIKM09)*, pages 375–384, New York, NY, USA. ACM.

[Liu et al., 2013] Liu, X., Tang, K., Hancock, J., Han, J., Song, M., Xu, R., and Pokorny, B. (2013). A text cube approach to human, social and cultural behavior in the twitter stream. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'13, pages 321–330, Berlin, Heidelberg. Springer-Verlag.

[Magnini and Cavaglia, 2000] Magnini, B. and Cavaglia, G. (2000). Integrating subject field codes into wordnet. In *LREC*. European Language Resources Association.

[Magnini et al., 2002] Magnini, B., Strapparava, C., Pezzulo, G., and Gliozzo, A. M. (2002). The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Martin et al., 2013] Martin, C., Corney, D., and Goker, A. (2013). Mining newsworthy topics from social media. In *BCS SGAI Workshop on Social Media Analysis*, pages 35–46, Cambridge, UK.

[Martin-Bautista et al., 2008] Martin-Bautista, M. J., Martínez-Folgoso, S., and Vila, M. A. (2008). A new semantic representation for short texts. In *Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery (DaWaK' 2008)*, Lecture Notes in Computer Science (LNCS), Turin, Italy. Springer-Verlag.

[Martin-Bautista et al., 2015] Martin-Bautista, M. J., Martínez-Folgoso, S., and Vila, M. A. (2015). A new approach for representing and querying textual attributes in databases. *International Journal of Intelligent Systems*, pages pp. 1021–1045.

[Martín-Bautista et al., 2010] Martín-Bautista, M. J., Molina, C., Tejeda, E., and Vila, M. A. (2010). Using textual dimensions in data warehousing processes. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications - 13th International Conference, IPMU 2010, Dortmund, Germany, June 28 - July 2, 2010. Proceedings, Part II*, pages 158–167.

[Martin-Bautista et al., 2013] Martin-Bautista, M. J., Molina, C., Tejeda-Avila, E., and Vila, M. A. (2013). A new multidimensional model with text dimensions: definition and implementation. *International Journal of Computational Intelligence Systems*, 6(1):pp. 137–155.

[Martin-Bautista et al., 2006] Martin-Bautista, M. J., Prados, M., Vila, M. A., and Martínez-Folgoso, S. (2006). A knowledge representation for short texts based on frequent itemsets. In *Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU' 2006)*, Paris, France.

[Martinez-Romo and Araujo, 2013] Martinez-Romo, J. and Araujo, L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40:2992–3000.

[McAffer and Lemieux, 2005] McAffer, J. and Lemieux, J.-M. (2005). *Eclipse Rich Client Platform: Designing, Coding, and Packaging Java(TM) Applications*. Addison-Wesley Professional.

[Moya et al., 2011] Moya, L. G., Kudama, S., Cabo, M. J. A., and Llavori, R. B. (2011). Integrating web feed opinions into a corporate data warehouse. pages 20–27. cited By 0.

[Navigli and Ponzetto, 2012] Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

[Otte and Rousseau, 2002] Otte, E. and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6):pp. 441–453.

[Park et al., 2013] Park, D., Yu, J., Park, J. S., and Kim, M. S. (2013). Netcube: a comprehensive network traffic analysis model based on multidimensional olap data cube. *International Journal of Network Management*, 23(2):pp. 101–118.

[Pease et al., 2002] Pease, A., Niles, I., and Li, J. (2002). The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, page 2002.

[Pennacchiotti and Gurumurthy, 2011] Pennacchiotti, M. and Gurumurthy, S. (2011). Investigating topic models for social media user recommendation. In *20th International Conference Companion on World Wide Web*, pages 101–102, New York, NY, USA. ACM.

[Pérez et al., 2008] Pérez, J. M., Berlanga, R., Aramburu, M. J., and Pedersen, T. B. (2008). Towards a data warehouse contextualized with web opinions. In *e-Business Engineering, 2008. ICEBE '08. IEEE International Conference on*, pages 697–702.

[RaghavaRao et al., 2012] RaghavaRao, N., Sravankumar, K., and Madhu, P. (2012). A survey on document clustering with hierarchical methods and similarity measures. *International Journal of Engineering Research & Technology (IJERT)*, 1(7).

[Rehman et al., 2013] Rehman, N. U., Weiler, A., and Scholl, M. H. (2013). Olaping social media: The case of twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 1139–1146, New York, NY, USA. ACM.

[Rousseeuw, 1987] Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):pp. 53–65.

[Salton and McGill, 1983] Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, USA.

[Skarmeta et al., 2000a] Skarmeta, A. G., Bensaid, A., and Tazi, N. (2000a). Data mining for text categorization with semi-supervised agglomerative hierarchical clustering. *International Journal of Intelligent Systems*, 15(7):633–646.

[Skarmeta et al., 2000b] Skarmeta, A. G., Bensaid, A., and Tazi, N. (2000b). Data mining for text categorization with semi-supervised agglomerative hierarchical clustering. *International Journal of Intelligent Systems*, 15(7):pp. 633–646.

[Toutanova et al., 2003] Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Valitutti, 2004] Valitutti, R. (2004). Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

[Voorhees, 1986] Voorhees, E. M. (1986). Implementing agglomerative hierarchical clustering for use in information retrieval. Technical Report TR86–765, Cornell University, Ithaca, NY.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

[Willett, 1988] Willett, P. (1988). Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*, 24(5):pp. 577–597.

[Wu et al., 2011] Wu, J., Gao, W., Zhang, B., Liu, J., and Li, C. (2011). Cluster based detection and analysis of internet topics. In *4th International Symposium on Computational Intelligence and Design, ISCID 2011*, volume 2, pages 371–374.

[Xiaohui et al., 2013] Xiaohui, H., Xiaofeng, Z., Yunming, Y., Shengchun, D., and Xutao, L. (2013). A topic detection approach through hierarchical clustering on concept graph. *Applied Mathematics & Information Sciences*, 7(6):pp. 2285–2295.

[Young-Woo and Sycara, 2004] Young-Woo, S. and Sycara, K. (2004). Text clustering for topic detection. Technical Report CMU-RI-TR-04-03, Robotics Institute, Pittsburgh, PA.

[Zhang et al., 2009] Zhang, D., Zhai, C., and Han, J. (2009). Topic cube: Topic modeling for olap on multidimensional text databases. In *SDM*, volume 9, pages 1124–1135. SIAM.

[Zhao et al., 2011a] Zhao, P., Li, X., Xin, D., and Han, J. (2011a). Graph cube: On warehousing and olap multidimensional networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 853–864, New York, NY, USA. ACM.

[Zhao et al., 2011b] Zhao, W. X., Weng, J., He, J., Lim, E. P., and Yan, H. (2011b). Comparing twitter and traditional media using topic models. In *33rd European conference on advances in information retrieval (ECIR11)*, pages 338–349. Berlin, Heidelberg: Springer-Verlag.

[Zhao and Karypis, 2002] Zhao, Y. and Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages pp. 515–524, New York, NY, USA. ACM.

[Zhao and Karypis, 2004] Zhao, Y. and Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Mach. Learn.*, 55(3):311–331.

[Zheng and Li, 2011] Zheng, L. and Li, T. (2011). Semi-supervised hierarchical clustering. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, pages 982–991, Washington, DC, USA. IEEE Computer Society.