UNIVERSIDAD
DE GRANADA

TEXT AND OPINION MINING TECHNIQUES IN SOCIAL MEDIA ENVIRONMENTS

DOCTORAL DISSERTATION
presented to obtain the
DOCTOR OF PHILOSOPHY DEGREE
in the
INFORMATION AND COMMUNICATION TECHNOLOGY PROGRAM
by

José Ángel Díaz García

Ph.D. Advisors

María José Martín Bautista & María Dolores Ruiz Jiménez

DEPARTMENT OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

Granada, Octubre 2023

Editor: Universidad de Granada. Tesis Doctorales
Autor: José Ángel Díaz García
ISBN: 978-84-1195-094-7
URI: https://hdl.handle.net/10481/85705

This Ph.D. dissertation titled "*Text and Opinion mining techniques in social media environments*", which is presented by José Ángel Díaz García to obtaining the Ph.D. degree, has been carried out within the Official Ph.D. Program "*Information and Communication Technologies*" of the Department of Computer Science and Artificial Intelligence of the University of Granada, and under the guidance of professors María José Martín Bautista and María Dolores Ruiz Jiménez.

The Ph.D. student and his Ph.D. advisors guarantee, by signing this doctoral thesis, that the work has been carried out by the Ph.D. student under the direction of Ph.D. advisors, and as far as our knowledge is concerned, the rights of authorship have been respected.

Granada, Octubre de 2023.

The Ph.D. student:	The Ph.D. advisor:	The Ph.D. advisor:
Sgd.: José Ángel Díaz García	Sgd.: María José Martín Bautista	Sgd.: María Dolores Ruiz Jiménez

Funding

This doctoral thesis has been supported mainly by the Spanish Ministry of Education, Culture and Sport (FPU18/00150). The project is also partially supported by the COPKIT Project, through the European Union's Horizon 2020 Research and Innovation Programme, under Grant 786687, by the Spanish Ministry for Economy and Competitiveness through a project under Grant TIN2015-64776-C3-1-R, the Andalusian government and the FEDER operative program under the project BigDataMed (P18-RT-2947 and B-TIC-145-UGR18) and grant PLEC2021-007681 funded by MCIN / AEI / 10.13039 / 501100011033 and by the European Union NextGenerationEU / PRTR. Also, the research is part of DESINFOSCAN project, founded by Ministerio de Ciencia e Innovacion and by the European Union NextGenerationEU (Grant TED2021-1289402B-C21).

A mi pareja, familia y amigos.

Agradecimientos

Quiero comenzar expresando mi más profundo agradecimiento a mis directoras de tesis, María José Martín y María Dolores Ruiz, por su apoyo, dedicación y sabiduría a lo largo de todo el proceso. Su guía y mentoría han sido fundamentales en el éxito de mi investigación doctoral.

También quiero agradecer a mi pareja, familia y amigos por estar a mi lado en los momentos de debilidad y por brindarme su amor incondicional en todo momento. Sin su apoyo emocional y su ánimo constante, este logro no habría sido posible.

No quiero dejar pasar esta oportunidad sin agradecer al Programa de Formación de Profesorado Universitario (FPU) del Ministerio de Ciencia y Educación del Gobierno de España, por su apoyo financiero durante mi programa de doctorado. Gracias a sus fondos he podido realizar esta investigación y culminar con éxito mi formación como investigador. Este programa ha sido clave en mi crecimiento académico y profesional, permitiéndome no solo desarrollar mis habilidades y conocimientos, sino también aplicarlos en un contexto real.

Por último, pero no menos importante, quiero reconocer y agradecer a todos los miembros de mi laboratorio UGRITAI. Juntos hemos construido un proyecto competitivo y vanguardista que ha dado como resultado tesis doctorales como la mía. Han sido largas horas de trabajo y dedicación, pero el resultado ha sido una experiencia inolvidable y altamente satisfactoria. Gracias a todos por su compromiso y contribución a mi formación como investigador.

Table of Contents

List of Abbreviations

KDD Knowledge Discovery in Databases.

KDT Knowledge Discovery in Texts.

LDA Latent Dirichlet Allocation.

COVID Coronavirus Disease.

NLP Natural Language Processing.

IR Information Retrieval.

KNN KNearest Neighbours.

AR Association Rules.

GAR Generalized Association Rules.

ARM Association Rules Mining.

CBOW Continuous Bag of Words.

BERT Bidirectional Encoder Representations from Transformers.

LSTM Long Short-Term Memory.

CNN Convolutional Neural Network.

RNN Recurrent Neural Network.

TSNE T-distributed Stochastic Neighbour Embedding.

TF Term Frequency.

TF-IDF Term Frequency–Inverse Document Frequency.

NNLM Feedforward Neural Net Language Model.

NOFACE NOise Filtering According Credibility and Expertise.

SRTD Scalable and Robust Truth Discovery.

FAV Favourite.

RT ReTweet.

URL Uniform Resource Locator.

CSV Comma-Separated Values.

MR MapReduce.

RDD Resilient Distributed Datasets.

API Application Programming Interface.

AWS Amazon Web Services.

Abstract

Social networks have assumed a crucial role in our lives, becoming a daily means of communication and information. This emergence has led to substantial advancements and improvements in various aspects of our daily routines. Social networks witness a massive influx of data every day, and when processed effectively, this data can confer competitive advantages to businesses and aid in the mitigation of significant issues for the society such as the proliferation of misinformation.

This thesis focuses on the design and development of solutions specifically tailored to handle unstructured data from social networks, with a primary emphasis on opinion mining. The research has yielded promising results, including the introduction of unsupervised opinion mining techniques for large-scale sentiment analysis. Additionally, novel metrics and algorithms have been proposed to effectively combat misinformation, leveraging user-generated content and experience.

The outcomes of this research have made substantial contributions to the field of opinion mining in social networks. The findings showcase significant progress in the analysis of unstructured data, coupled with effective strategies to counter the dissemination of misinformation and the study of opinion holders (users). These solutions provide robust and efficient means to comprehend the opinions and sentiments expressed within social networks, thereby presenting profound implications for both businesses and society as a whole.

Resumen

Las redes sociales han asumido un papel crucial en nuestras vidas, convirtiéndose en un medio cotidiano de comunicación e información. Esta aparición ha propiciado avances y mejoras sustanciales en diversos aspectos de nuestras rutinas diarias. En las redes sociales diariamente se genera una masiva e ingente cantidad de datos que cuando se procesan de forma eficaz, pueden converger en ventajas competitivas para las empresas o para la sociedad, ayudando a mitigar problemas importantes como la proliferación de la desinformación.

Esta tesis se centra en el diseño y desarrollo de soluciones específicamente adaptadas para tratar datos no estructurados procedentes de redes sociales, con especial énfasis en la minería de opinión. La investigación ha finalizado arrojando resultados muy relevantes como la introducción de técnicas no supervisadas de minería de opinión para el análisis de sentimientos a gran escala. Además, se han propuesto nuevas métricas y algoritmos para combatir eficazmente la desinformación, aprovechando la experiencia y el contenido generado por el usuario.

Los resultados de esta tesis han contribuido sustancialmente al campo de la minería de opinión en redes sociales y muestran avances significativos en el análisis de datos no estructurados, junto con estrategias eficaces para contrarrestar la difusión de desinformación y el estudio de los usuarios que generan estas opiniones. Estas soluciones proporcionan medios sólidos y eficaces para comprender las opiniones y sentimientos expresados en las redes sociales, así como su credibilidad lo que tiene profundas implicaciones tanto para las empresas como para la sociedad en su conjunto.

Chapter I

Ph.D. Dissertation

1 Introduction

In the current digital era, social media has acquired unprecedented relevance in our everyday lives. These platforms have revolutionized the way we communicate, share information, and connect with people from around the world. Since their emergence, social media has become an integral part of our social interactions, and their influence spans from personal to professional spheres.

Social media has democratized the dissemination of information and provided individuals with the ability to express themselves and participate in global conversations. Millions of people use platforms like Facebook, Twitter, Instagram, and LinkedIn daily to share ideas, opinions, experiences, and multimedia content. However, the exponential growth of user-generated content has presented a new challenge: the efficient processing of the vast amount of unstructured information generated on these platforms.

Unstructured content on social media encompasses a wide range of data, including posts, comments, images, videos, and links. As more and more people engage in social media, there is an explosion of content, making its management and analysis increasingly difficult. The lack of structure and organization in this context means that processing and extracting relevant information from it pose a significant challenge.

The importance of addressing this issue lies in the potential value that can be gained from analyzing and understanding the unstructured content of social media. This data can contain valuable insights, trends, and information about various aspects of society, such as public opinions, consumer preferences, current events, and social behaviors. However, to fully harness this wealth of information, it is crucial to develop efficient and effective solutions for processing and analyzing the unstructured content of social media. This is where data mining and natural language processing (NLP) techniques come into play, enabling us to effectively manage this vast amount of information.

Data mining constitutes a pivotal component of the Knowledge Discovery in Databases (KDD) process, offering diverse avenues for unveiling concealed insights from vast datasets. When data mining is directed towards unstructured text data, we are dealing with text mining. Text mining can be seen as a subset of data mining, that is primarily dedicated to refining, organizing, and adapting unstructured text data for further analysis, rendering it more intelligible and compatible with various machine-learning techniques. Within the realm of text mining, we can discern two principal domains: one pertains to unsupervised techniques, while the other involves supervised methods, contingent upon the nature of the input data. An unsupervised system does not need tagged datasets, while supervised datasets need tagged datasets. Both techniques are useful and depend on the nature of the data and the problem that we are facing to choose one or the other.

In the context of social media, the vast and rapidly generated data make it a challenging task to tag and evaluate every piece of information. Therefore, our research focuses on the necessity of dealing with massive amounts of data using unsupervised techniques. The thesis will center on text and opinion mining techniques, employing methodologies such as association rules, clustering, and unsupervised word embeddings like Word2Vec. These methods will enable the extraction of valuable insights, sentiments, and hidden relationships from the immense volume of social media data in an unsupervised manner.

2 Objectives

In light of the core concepts discussed above, the primary objective of this study is to design and develop text and opinion mining techniques that aim to enhance the processing of massive volumes of unstructured text derived from social media platforms. The exponential growth of user-generated content in social media has resulted in an immense challenge in effectively managing and extracting valuable insights from this unstructured data. Therefore, the central focus of this research is guided by the hypothesis: **The development and implementation of novel methods and algorithms for the automatic analysis and extraction of meaningful information from social media environments will lead us to a better understanding of this data deriving useful hidden insights**. In line with this hypothesis, we can further break down our specific objectives as follows:

- **(O1) Study the state of the art of text and opinion mining:** This objective focuses on conducting a comprehensive review of the current state of the art in Data Mining techniques employed for opinion mining and text mining. Special attention will be given to techniques that address the challenges of handling large volumes of data in an unsupervised way. By examining the existing techniques, this research aims to identify gaps, limitations, and potential areas for improvement, setting the foundation for developing novel techniques.
- **(O2) Design and development of text and opinion mining techniques:** This objective focuses on designing and implementing text and opinion mining techniques with a special focus on the need to obtain techniques that can process the large volumes of unlabelled data that are so rapidly generated in social networks in a generalizable, efficient way and without the need for large-scale, time-consuming and costly data processing. For this, we have focused on two techniques, association rules and word embeddings, so this objective in turn can be broken down into two under the following motivations:
  - **(O2.1)** Using association rules: Association rules provide a powerful tool for uncovering interesting relationships and patterns in large datasets without the need for explicit labels. By leveraging association rule mining, the thesis seeks to identify meaningful connections between different entities and sentiments in social media content
  - **(O2.2)** Using word embbedings: Word embeddings offer an efficient and effective way to represent words as dense vector representations in a continuous space. By utilizing word embeddings, the thesis aims to enhance opinion mining by capturing semantic relationships between words. These embeddings provide a compact and meaningful representation of words, enabling more accurate and computationally efficient knowledge extraction from social media data.
- **(O3) Experimental design and validation of the system in real problems**: In this objective, the developed opinion mining systems and text mining techniques will be rigorously tested, evaluated, and validated using real-world problems and datasets. An experimental design will be established to assess the system's performance, scalability, and usability. The research will involve conducting experiments on diverse social media datasets, representative of different domains and contexts such us political opinion forecasting or disinformation analysis on COVID datasets. The validation process will provide valuable insights into the system's robustness and generalizability, strengths, limitations, and areas for further improvement, ensuring its effectiveness for addressing real-world challenges in opinion mining from social media data.

3 Thesis development methodology

The work conducted in this thesis falls within the realm of computational science, which combines the principles of the scientific method with specific stages aligned with data science and computational techniques. By adopting this approach, we have harnessed the power of rigorous experimentation and analysis while leveraging data-driven methodologies to address the challenges of opinion mining in social networks. The stages of the methodology are:

- **Problem definition and hypothesis formulation**: The first step in the methodology is to clearly define the problem to be addressed in the research. This involves understanding the specific objectives of the study, such as designing efficient opinion mining techniques for large-scale unlabelled data in social networks.
- **Literature review and background research**: Before proceeding with the unsupervised techniques, a comprehensive literature review is conducted to understand the existing state-of-the-art methods in opinion mining and its applications by means of association rules and word embeddings. This step helps in identifying relevant techniques and gaining insights into the challenges and opportunities in the field.
- **Data collection**: The acquisition of extensive and high-quality datasets is indispensable in the context of NLP and social media mining, especially for testing and advancing subsequent stages of the research. In this regard, the data collection phase plays a vital role in ensuring the availability of relevant data derived from real-world use cases for subsequent analysis.
- **Experimentation and observations compilation**: In this stage of our methodology, we rigorously tested the developed techniques by compiling and analyzing results related to performance, time, memory usage, and scalability. To ensure the reliability and generalizability of our findings, we conducted statistical tests and hypothesis contrasts between multiple executions with different random initializations. This comprehensive approach allowed us to make informed comparisons and validate the effectiveness of the techniques in handling real-world data scenarios.
- **Interpretation**: The results obtained from the developed techniques are interpreted in the context of the research objectives and contrasted with the initial hypotheses. The findings are discussed, and conclusions are drawn, highlighting the contributions of the proposed text and opinion mining techniques.
- **Thesis writing**: In the final stage of our methodology we consolidate all the information and insights obtained throughout the research process. This stage involves the systematic extraction and organization of results, observations, and conclusions derived from the data science techniques and experimentation.

4 Theoretical framework

This section provides a concise overview of the theoretical background driving our thesis. It covers key topics such as text mining, opinion mining, association rules, Word Embeddings. Additionally, we focus on addressing the critical issue of disinformation, where fake or low-quality opinions can impact the reliability of social media content. By understanding these theoretical aspects, we lay the foundation for effective understanding of the research carried out.

4.1 Data & Text mining

The KDD process, as introduced by Fayyad et al. [FPSS96], provides a comprehensive framework for extracting valuable insights from data. Central to the KDD process is data mining, which involves the application of automated techniques to transform data into actionable knowledge. When dealing with unstructured text data, such as user-generated comments, this process is commonly referred to as text mining.

Within the realm of text mining, we can align the process with a variation known as Knowledge Discovery in Texts (KDT). In our specific context, we have adapted the traditional objectives of both KDT and KDD to cater to the task of mining textual content from social media environments. Our approach involved a series of well-defined steps that guided our processes and conclusions, ensuring a systematic and rigorous exploration of the data:

- Problem understanding: Define the research question or problem to be addressed in the context of social media text analysis. With a particular focus on identifying specific goals and objectives of the analysis.
- Data understanding: Collect and explore the social media text data, understanding its characteristics and format. Paying special attention to the types of texts, user interactions, and potential limitations or biases in the data.
- Data preparation: Preprocess and clean the social media text data, including tasks such as removing irrelevant content, normalizing text, handling noise, and resolving duplicates. Transform the unstructured text into a structured format suitable for analysis depending on the subsequent stages.
- Data mining: Apply appropriate data mining techniques, such as sentiment analysis, topic modeling, or network analysis to extract patterns, relationships, and insights from the data.
- Visualization and evaluation: Assess the quality and usefulness of the mined patterns and insights. Evaluate the performance of the applied data mining techniques. Validate the discovered knowledge using visualization tools.
- Interpretation and presentation: Interpret the mined patterns and insights in the context of the research question or problem.

Selecting an appropriate algorithm during the data mining step is a pivotal stage in the Knowledge Discovery in Texts (KDT) process. This decision determines how well the model can capture the inherent characteristics of the input data. This selection depends on the nature of the data, its volume, and the specific information needs. This stage is framed with the machine learning techniques,

and the different machine learning algorithms can be broadly classified into two main categories: Supervised and unsupervised techniques.

- **Supervised techniques:** A type of machine learning where the algorithm learns from a labeled dataset, meaning that the input data is accompanied by corresponding output labels or target values. The goal of supervised learning is to learn a mapping from input data to output labels, enabling the model to make predictions on new, unseen data. Within this category, we can find classification algorithms for assigning data into predefined classes (i.e. classes or labels) and regression algorithms for predicting continuous values (i.e. stock prices, or house prices).
- **Unsupervised techniques:** A type of machine learning where the algorithm is trained on unlabeled data, in other words, the target variable is not predefined in advance. The primary objective of unsupervised learning is to uncover patterns, structures, and relationships within the data with the absence of explicit guidance or without relying on prior information. Framed with unsupervised techniques, the most widespread techniques are:
  - Association Rule Mining: Association rule mining is used to discover interesting associations and correlations between items in large datasets. It is commonly applied in market basket analysis to identify frequently co-occurring items in shopping transactions.
  - Clustering: Clustering algorithms are based in grouping similar data points into clusters depending on their feature similarities. This technique finds extensive application in various domains, with notable use cases including customer segmentation and anomaly detection. In customer segmentation, clustering helps identify distinct groups of customers who share similar behaviors or characteristics, enabling businesses to tailor their marketing strategies and offerings accordingly. In the context of anomaly detection, clustering aids in identifying unusual patterns or outliers within datasets, which can be crucial for fraud detection, network security, and quality control, among other applications.
  - Word embeddings and autoencoders: An autoencoder is a type of neural network used for unsupervised learning and dimensionality reduction. It consists of an encoder that compresses the input data into a lower-dimensional representation and a decoder that reconstructs the original data from the compressed representation. The autoencoder tries to minimize the reconstruction error between the input and the output. On the other hand, Word2Vec algorithms are trained to maximize the likelihood of predicting the surrounding words or context words given the target word. That allows us to obtain semantical and contextual relations between words without the necessity of tagged datasets.

Text mining techniques encompass a diverse set of methods and algorithms aimed at extracting valuable patterns and insights from extensive datasets. Through the utilization of NLP techniques, text mining enables the identification of concealed relationships, sentiments, and trends that might not be readily apparent. Some of the most commonly employed techniques in text mining for social media text are:

- Text classification: Text classification [KJMH+19] techniques enable the categorization of social media posts, comments, or messages into predefined categories or topics. By training machine learning models on labeled data it becomes possible to automatically classify new, unlabeled text based on patterns and features derived from the labeled data. This facilitates the organization and analysis of social media content, for example discern between spam reviews and organic reviews [FKR+20].

- Text clustering: Text clustering [AZ12] shares similarities with text classification, but the key distinction lies in the availability of labeled datasets. In text clustering, we lack labeled data, and thus the categorization of the text is based solely on the inherent features and characteristics present in the text.

- Topic modeling: Topic modeling techniques [CU21], such as Latent Dirichlet Allocation (LDA), aim to identify latent topics within a collection of social media texts. These methods uncover hidden thematic structures and group related pieces of text together. By employing topic modeling, analysts can identify the prevalent subjects or discussions in social media conversations, providing insights into the interests [JK21] and concerns of users.

- Network analysis: Network analysis techniques [SGAO14] focus on the relationships between social media users, as well as their interactions and information flow. By constructing networks based on user mentions, hashtags, or retweets, analysts can uncover influential users [FJUA18], communities, and information diffusion patterns. Network analysis helps to uncover hidden connections and the flow of information within social media platforms.

- Sentiment analysis: Sentiment analysis, [KJ20] also known as opinion mining, focuses on determining the sentiment or emotion expressed in social media texts. By employing NLP techniques, sentiment analysis algorithms can assess whether a piece of text conveys positive, negative, or neutral sentiment. This allows tracking of public opinion [KRA20], sentiment trends, and the identification of influential topics or events. In our thesis, we distinguish between the polarity of an opinion and the opinion's components themselves. Given the significant impact of this technique on our thesis, we will thoroughly analyze both the opinion mining technique and the approach we have adopted in Section 4.2.

4.2 Opinion mining

Opinion mining is a widely employed technique in the realm of social media analysis. Its primary objective is to determine the sentiment associated with a particular piece of information, such as categorizing opinions about a restaurant as positive, negative, or neutral. This area is often referred to as sentiment analysis. While sentiment analysis and opinion analysis are sometimes considered synonymous in literature, we, along with other authors, believe that sentiment analysis constitutes just one aspect of opinion mining, not its entirety.

In [LZ12] Liu defines an opinion as a quintuple, comprising components like the **entity** being discussed, the source of the opinion or **opinion holder**, the **aspect** or evaluation of the entity, the **sentiment** of this evaluation, and the **time** of expression. Opinions are dynamic and can evolve over time. In our research, we draw parallels between these opinion components and various problems and solutions that text and opinion mining techniques can address, i.e for a published tweet:

- Entity: About what is the published tweet, for example, a brand. Can be accomplished using techniques like NER or topic analysis.

- Opinion holder: We can detect bots and credible users by employing credibility modeling techniques to analyze the user publishing the tweet.

- Aspect: Understanding what is being evaluated in a tweet using NER or event detection techniques.

- Sentiment: We can publish a tweet positive, negative or neutral. Here is where sentiment analysis techniques play a crucial role. In addition to sentiment analysis, emotion analysis techniques can be employed to detect emotions expressed in the text, such as happiness, sadness, anger, etc.
- Time: Date and time the tweet was published. Here, time series analysis can be utilized i.e to predict product assimilation based on the date and time a tweet was published.

Hence, while sentiment analysis is a part of opinion mining, the latter is a broader technique that encompasses various aspects beyond just sentiments. Within opinion mining, different approaches to get opinion insights are used. Some of the most widespread are:

- **Lexicon-based approaches**: These methods use sentiment lexicons or dictionaries to determine the sentiment scores of words in the text, aggregating them to obtain the overall sentiment of the document. Additionally, lexicon-based systems can be employed to extract entities and aspects related to opinions.
- **Machine Learning algorithms**: Supervised machine learning algorithms like Support Vector Machines (SVM) and Decision Trees can be trained on labeled datasets to classify opinions into sentiment categories or ratings. Unsupervised techniques such as clustering are used to group unlabeled opinions based on similarities. Both supervised and unsupervised techniques can also be used to assess the credibility of opinion holders, identifying non-credible sources.
- **Deep Learning techniques**: Deep learning models, such as Convolutional Neural Networks (CNNs) and Transformers (e.g., BERT), have shown great success in opinion mining. These models can capture complex patterns and contextual information in the text, leading to more accurate systems. These systems are being applied to solve problems such as the credibility of opinion holders and opinions, sentiment and emotion identification or relation extraction between entities and aspects.
- **Hybrid approaches**: Hybrid methods combine multiple techniques, such as lexicon-based and machine learning approaches, to leverage their respective strengths and improve overall performance.

In our thesis, we primarily focused on sentiment analysis and discerning credible opinion holders. This aspect will be further explored in Section 4.4.1 due to its significance.

4.3 Association rules

Association rules belong to the Data Mining field and have been used and studied for a long time. One of the first references to them dates back to 1993 [AIS93]. They are used to obtain relevant information from large transactional databases. A transactional database could be, for example, a shopping basket database, where the items would be the products, or a text database, as in our case, where the items are the words, or more specifically, the entities represented by the words. In a more formal way, let t=$\{A,B,C\}$ be a transaction of three items *(A, B and C)*, and any combination of these items forms an itemset. In this case, all the possible itemsets are: $\{A,B,C\}$, $\{A,B\}$, $\{B,C\}$, $\{A,C\}$, $\{A\}$,$\{B\}$ *and* $\{C\}$. According to this, an association rule would be represented in the form $X \rightarrow Y$, where X is an itemset that represents the antecedent, and Y an itemset called consequent where $(X \cap Y = \phi)$. As a result, we can conclude that consequent itemsets have a co-occurrence relation

with antecedent itemsets. Therefore, association rules can be used as a method for extracting hidden relationships among items or elements within transactional databases, data warehouses or other types of data storage. The classical way of measuring the goodness of association rules regarding a given problem is using three measures: support, confidence and lift, which are defined as follows:

- Support of an itemset. It is represented as *supp (X)*, and is the proportion of transactions containing itemset X out of the total number of transactions of the dataset (D). Support is defined by equation (I.5).

$$supp(X) = \frac{|t \in D : X \subseteq t|}{|D|} \tag{I.1}$$

- Support of an association rule. It is represented as *supp(X → Y)* and is the total amount of transactions containing both itemsets X and Y, as defined in the following equation:

$$supp(X \rightarrow Y) = supp(X \cup Y) \tag{I.2}$$

- Confidence of an association rule. It is represented as *conf (X→Y)* and represents the proportion of transactions containing itemset X which also contains Y. The equation is:

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{I.3}$$

- Lift. It is a useful measure to assess independence between itemsets of an association rule. The measure *lift (X→Y)* represents the degree to which X is frequent when Y is present or vice versa. Lift is a very interesting measure as it relates the antecedent and the consequent through the concept of independence. A value of 1 indicates that the appearance of the consequent and the antecedent in the same rule is independent, therefore, the rule has no effect. On the other hand, lift values greater than 1 indicate a dependence between antecedent and consequent that will make the rule perfect for predicting the consequent in future datasets. Negative values indicate that the presence of one item has a negative effect on the presence of another. Lift is defined mathematically in the following way:

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)} \tag{I.4}$$

Since association rules demonstrated their great potential to obtain hidden co-occurrence relationships within transactional databases, they have been increasingly applied in different fields. Among other fields, one in which association rules have attracted a lot of interest is Text Mining. One of the first papers which addresses the problem of text mining with association rules, is the paper presented by Martin-Bautista et al. [MBSSV04]. In this work, textual transactions are defined, on which fuzzy association rules are applied. Textual transactions are necessary in order to be able to apply association rules on text, opening the possibility of applying association rules to text mining problems. In this field, text entities (opinions, tweets,...) are handled as transactions in which each of the words is an item. In this way, we can obtain relationships and metrics about co-occurrences in large text databases. Technically, we could define a text transaction as:

Definition 1. *Text transaction: Let W be a set of items (words in our context). A text transaction is defined as a subset of words, i.e. a word will be present or not in a transaction.*

For example, in a Twitter database, in which each tweet is a transaction, a text transaction will be composed of each of the terms that appear in that tweet. So the items will be the words. The structure will be stored in a term matrix in which the terms that appear will be labelled with 1 and those that are not present as 0. For example for the transactional database $D = \{t1, t2\}$ being $t1 = (just, like, emails, requested, congress)$ and $t2 = (just, anyone, knows, use, delete, keys)$ the representation of text transactions is shown in Table I.1.

Table I.1: Example of a term matrix in a database with two textual transactions.

Transaction\Item	***anyone***	***congress***	***delete***	***emails***	***just***	***keys***	***knows***	***like***	***requested***	***use***
t1	0	1	0	1	1	0	0	1	1	0
t2	1	0	1	0	1	1	1	0	0	1

We can see how in a real problem this matrix will be very sparse. This is a problem that is often mitigated with different data cleaning processes. For example, in some cases, terms that are similar or represent a closely related entity are exchanged for this entity so that the term matrix is less sparse.

Transactions are an essential part of the association rule mining extraction process and without them we would not be able to mine frequent patterns. Association rules cannot be applied on raw text, so with this internal representation of text as textual transactions, association rules can be applied to almost any textual data problem. The transactions can be represented as a matrix where each cell contains 1 or 0 if the term is present or not. However, it is interesting in text mining to use the absolute frequency of terms (TF) [Rob04] or by employing its Term Frequency - Inverse Document Frequency (TF-IDF), or even its fuzzy membership value, which gives to this representation a great versatility still to be exploited.

Before finishing this section it is necessary to mention that the internal binary or fuzzy matrix representation is not the only possible way to represent the textual transactions. In some studies a standard representation is used in which the terms and texts present in social networks are used to form a new transactional database. For example, let us imagine a database of tweets. By means of the text mining procedure we can obtain feelings about each of the tweets and build a transactional element called sentiment, which could take the values sentiment=positive, sentiment=negative, sentiment=neutral. Another transactional element could be names, which would take the output produced by a Named Entity Recognition over the tweets. The Named Entity Recognition (NER) process is an automatic process that identifies entities and assings them a category within their grammatical or lexical category [RCE+11]. One of the most famous systems, proposed by Stanford University [MSB+14], can be used to obtain the names of people, places, etc. mentioned in a given document. With these textual categories located and tagged, we can create transactions. For example, names = (Katie) or names = (Biden,Trump). In this case a concrete example would be $D = \{t1, t2\}$ being $t1 = (sentiment = positive, names = Trump)$ and $t2 = (sentiment = neutral, names = Biden)$. On these two transactions we could perfectly apply association rules to find which feelings have more co-occurrence with certain names. Having any of these internal representations for transactions, it is possible to apply any association rules mining approach.

4.3.1 Apriori algorithm

The most widespread approach for mining association rules is based on the downward-closure property of support and consists of two stages. To be considered frequent, the itemset has to exceed

the minimum support threshold. In the second stage, association rules are obtained using the confidence or other assessment measure. To obtain the frequent itemsets, the algorithms, based on a minimum support value, will generate all the possible combinations of itemsets and will check if they are frequent or not. In each iteration, all the possible different itemsets that can be formed by combining those of the previous iteration are generated, so the itemsets will grow in size.

Within this category we find most of the algorithms for obtaining association rules, such as Apriori, proposed by Agrawal and Srikant [AS+94]. Apriori is based on the premise that if an itemset is frequent, then all its subsets are also frequent. Meaning that when we find one of unfrequent itemsets, all other itemsets containing this one do not need to be analyzed. Thus, we can prune the search tree avoiding checks and increasing efficiency. Although the most widely used algorithm is the Apriori algorithm, it is not the only one. We can find others such as Apriori-TID proposed by , FP-Growth proposed by Han et al. [HPY00] and Eclat [OZP+97].

4.3.2 Apriori-TID

The Apriori TID algorithm, introduced in [WLP+09], is a memory-efficient variation of the original Apriori algorithm designed to handle large transactional databases. The key improvement lies in the use of a novel data structure called TID-list, which stores transaction identifiers for each item in the database. This enables efficient candidate generation and pruning during the mining process. The algorithm begins by scanning the database to identify frequent 1-itemsets and then proceeds to iteratively generate candidate itemsets of increasing sizes. At each iteration, the TID-lists are leveraged to determine the support count of candidate itemsets, and those that fail to meet the minimum support threshold are pruned. By employing these memory-saving techniques, the Apriori-TID algorithm allows for faster and more scalable mining of frequent itemsets from extensive datasets.

4.3.3 Eclat algorithm

Eclat [OZP+97] is an improved version of the Apriori algorithm, which improves the execution time of the algorithm. The main difference with the Apriori algorithm is that the Eclat algorithm performs a vertical search, similar to the depth-first search of a graph, as opposed to the breadth-first search performed by the Apriori algorithm. The basic idea is to compute intersections between items. To do this, a list of items is created, in which the items are related to the transactions in which they appear. With this list, the algorithm can compute the support value of a candidate itemset and avoid generating subsets that will not reach the support threshold. In this way, it can reduce the computation time in obtaining rules, but the management of this list implies a higher memory consumption.

4.3.4 FP-Growth algorithm

The FP-Growth algorithm [HPY00] was proposed in 2000, as a solution to the memory problems generated by typical methods such as Apriori, seen above. It is a very efficient algorithm and is widely used in problems and solutions that could be framed under the name of big Data. FP-Growth creates a compressed model of the original database using a data structure called FP-tree, which is made up of two essential elements:

- Transaction network: Thanks to this network, the entire database can be abbreviated. At each node, an itemset and its support are described and calculated by following the path from the root to the node in question.
- Header table: This is a table of lists of items. That is, for each item, a list is created that links nodes of the network where it appears. Once the tree is constructed, using a recursive approach based on divide and conquer, frequent itemsets are extracted. To do this, first the support of each of the items that appear in the header table is obtained. Second, for each of the items that exceed the minimum support, the following steps are carried out:

  1. The section of the tree where the item appears is extracted by readjusting the support values of the items that appear in that section.
  2. Considering the extracted section, a new FP-tree is created.
  3. The itemsets that exceed the minimum support of this last FP-tree are extracted.

Thus, FP-Growth requires less memory than Apriori. Additionally, the divide and conquer principle makes FP-Growth attractive in big data environments. It is worth noting that one of the reasons why FP-Growth is more efficient is that it is not exhaustive, i.e. it does not get all possible rules. Apriori, on the other hand, does [FBRMB16].

4.4 Word Embeddings

Word embeddings have emerged as a fundamental and widely used technique in NLP due to their ability to capture semantic relationships between words [LLCS15], [LG14]. In NLP tasks, such as sentiment analysis, machine translation, and information retrieval, understanding the contextual meaning of words is crucial for accurate and meaningful results. Traditional methods, like one-hot encoding, represent words as sparse and high-dimensional vectors, which lack semantic information and are computationally expensive.

Word embeddings, on the other hand, offer a dense and low-dimensional representation of words in a continuous vector space. This representation allows for semantic relationships to be encoded in the vector distances and directions, enabling mathematical operations between word vectors to reveal meaningful relationships. For instance, simple arithmetic operations like *king* - *man* + *woman* can yield the vector representation for *queen* demonstrating the ability of word embeddings to capture associations between words like gender. Two prominent models for generating word embeddings are Word2Vec [MCCD13, MSC+13] and FastText [BGJM16].

Word2Vec uses neural networks to learn word embeddings by predicting context words based on a given word (Skip-Gram) or predicting a word based on its surrounding context (Continuous Bag of Words, CBOW). Both models are based on the model Feedforward Neural Net Language Model (NNLM) proposed in [BDV00]. The NNLM consists of an input layer, projection layer, hidden layer, and output layer. The input layer employs a vector of 1-of-v coding, for N previous words to a given word present in the corpus (where v is the number of distinct words in the corpus). The projection layer takes the input and projects it onto the hidden layer. While the projection process is relatively straightforward, the challenge arises when transitioning to the hidden layer.

In Continuous Bag of Words (CBOW), the non-linear hidden layer is not present, and the projection layer is shared for all words in the vocabulary, this allows Word2Vec to be more efficient than NNLM. Unlike NNLM which considers only the N previous words, CBOW considers both the previous and the future words, allowing the model to learn words based on their context. This makes

CBOW particularly useful for capturing the meaning of a word in its given context. On the other hand, Skip-Gram is another model within Word2Vec that takes a different approach. Instead of predicting the context words given a specific word, Skip-Gram predicts the target word based on its surrounding context words. This model allows for more fine-grained word relationships to be captured, as it focuses on learning the associations between a word and its context words. In Figure 1, retrieved from the original research proposed by Mikolov [MCCD13], CBOW and Skip-Gram models are shown.

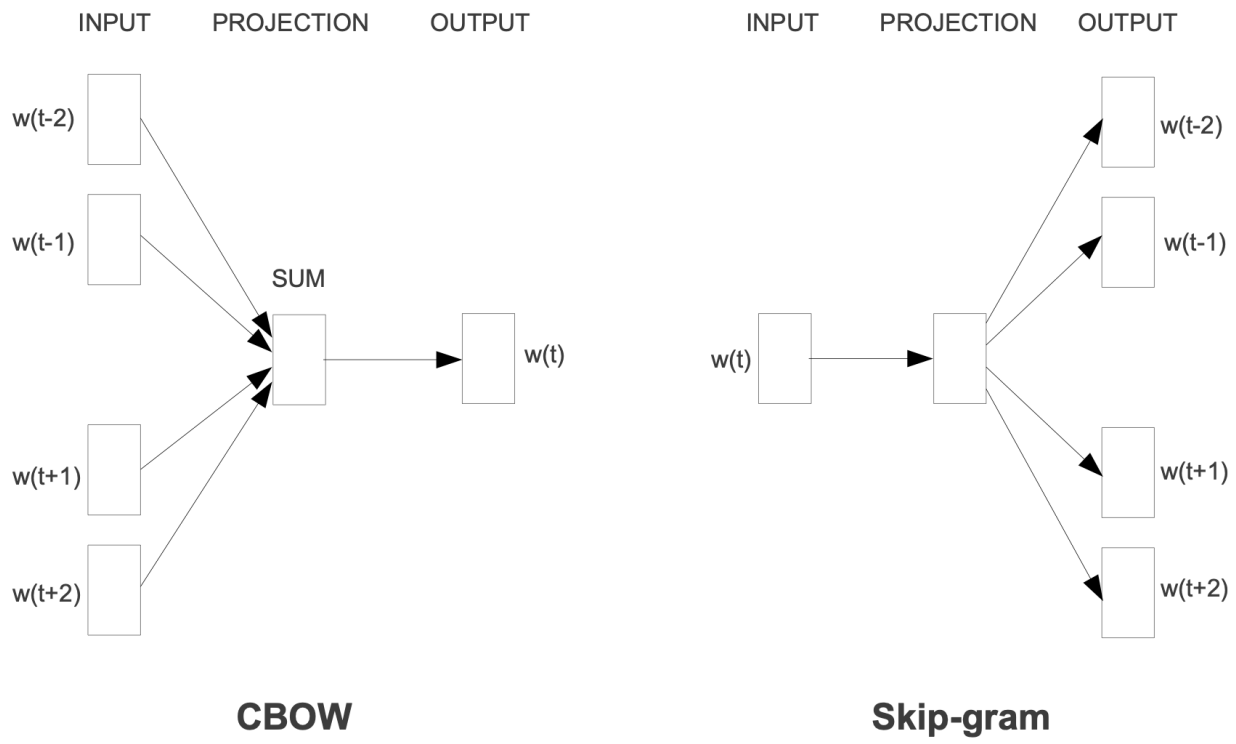

Figure 1: CBOW and Skip-Gram models

In contrast, FastText extends Word2Vec by considering subword information through character n-grams. For example for the word *matter*, and *n=3* we would have - *ma, mat, att, tte, ter, er-* and the final representation would be the sum of the vectors associated with each n-gram, which is particularly useful for handling out-of-vocabulary words and improving generalization.

Word embeddings have various advantages, including dimensionality reduction, capturing word relationships, and enabling efficient computation in downstream NLP tasks. They have transformed the field of NLP by providing effective and semantically meaningful representations of words, enhancing the performance of various language-based applications. Throughout this thesis, word embeddings play a pivotal role in sentiment analysis, entity recognition, and aspect-based opinion mining, contributing to a deeper understanding of social media data.

Before concluding this section, it is essential to highlight the groundbreaking Bidirectional Encoder Representations from Transformers (BERT) model. Proposed by Devlin et al. in [DCLT18], BERT has revolutionized the field of NLP. While both Word2Vec, FastText, and BERT can encode a text corpus into dense vectors, they differ significantly in their approach and capabilities.

Word embeddings are typically treated as word-level language models, where each word is represented independently of the sentence context. In contrast, BERT operates at the sentence level, comprehending the entire context and meaning dynamically. This dynamic nature enables BERT to learn contextualized word representations, allowing a word to have different vector representations based on the specific context of the sentence.

Furthermore, the training strategies of word embeddings and BERT are quite distinct. Word embeddings are commonly trained on extensive corpora using unsupervised learning, with the objective of predicting word contexts or targets. On the contrary, BERT is pre-trained using a masked language modeling task and a next-sentence prediction task. This pre-training equips BERT with a powerful ability to capture bidirectional context, resulting in a deeper understanding of language.

The application of these models in NLP tasks also varies significantly. Word embeddings are directly applied to specific tasks with minimal adaptation. In contrast, BERT requires fine-tuning on task-specific data to achieve optimal performance. In any case, both techniques can be used to support downstream tasks, such as those related to opinion mining.

4.4.1 Credibility and disinformation

Social networks have undoubtedly achieved remarkable success, but they have also become a target for nefarious purposes. One of the most significant challenges we encounter while navigating through social networks is the abundance of fake opinions or fake news, which are deliberately spread to misinform society for various motives. In our thesis, which centers on opinion mining, we must first address the critical issue of content credibility. But, what are credible information and disinformation?

In our field of application, credible and relevant information is understood as information that may contain valuable content in a certain context. For example, in a health topic, relevant information would be that opinion issued by a doctor about prevention measures for a certain disease. That is, in the case of Twitter, for the proposed case of health, we will be interested in keeping those opinions (tweets) and opinion holders (twitter users) of relevance to medicine, discarding those samples (tweets) that are not related to this sector.

As for misinformation or disinformation, there are more and more papers that provide a description or new characteristics of this concept [KA21], [AKI19]. Specifically, in [AKI19] they provide a very interesting vision of misinformation seen in different ways such as willful misinformation, fictional discussions, and non-verifiable information or news. In any case, it is untruthful information that is disseminated for various purposes, such as to negatively influence political issues. In these cases where there is a clear intention to disseminate false content, we will speak of disinformation. Therefore, the difference between misinformation and disinformation lies in the intentionality or purpose. In the scope of this thesis, we will focus on experience-driven misinformation reduction, since a person on Twitter may share false content, because of their beliefs without actually knowing whether that is false to a greater or lesser extent.

Being able to discern what is true or relevant in social networks and what is not (misleading or fake opinions, news and users), has taken up a great amount of literature in recent years [SSW$^+$17], [OHA$^+$19], with artificial intelligence systems assuming a great importance in the process. The process of eliminating misinformation on social networks is, by its nature, closely linked to the processes of dimensionality reduction and instance selection, as both seek to eliminate data that is not interesting for subsequent data mining processes. This process can be guided by statistical methods such as features or instance selection algorithms [OLCOMTK10], or by objective credibility

data in the case of misinformation detection.

Credibility has been a subject of exploration across various methods, with the majority focusing on utilizing network features to calculate a threshold that distinguishes credible from non-credible entities. The initial step in dealing with credibility involves employing text mining techniques to compute these metrics. Alternatively, supervised techniques can also be applied to learn what makes a user credible. We will delve deeper into this topic in Section 4.5.3

4.5 State of the Art

To achieve our objectives and make significant contributions to the field, our thesis began with a thorough analysis of the current state of the art. We have placed a strong emphasis on three key areas closely related to our objectives: sentiment analysis in social media, with a special focus on techniques based on association rules, the credibility of content or opinion holders (users), and the emerging topic of addressing misinformation. While some of our findings have already been published in high-impact journals [DGRMB23, DGRMB22], we deemed it essential to consolidate the most critical research from each area in this document. This approach allowed us to identify ongoing efforts and determine how we can enhance the existing knowledge. To ensure clarity and organization, we have categorized the section into subsections, each focusing on a specific topic of study.

4.5.1 Sentiment analysis in social media

In the domain of sentiment classification for textual entities, supervised methods have garnered significant attention. Recent papers such as [KYK+19, MSSS20] have explored various directed approaches, including decision trees for sentiment classification of tweets.

In parallel, deep learning techniques have been widely employed in the field of text mining [TF20, MK19, ZWL18, DSRO20]. While deep learning solutions yield promising results, they suffer from two main drawbacks. Firstly, they operate as black boxes, lacking interpretability in their outputs. Secondly, these methods demand considerable effort in collecting pre-classified data to tailor them for specific use cases. This presents a notable limitation for real-world applications.

Recognizing the need for more interpretable and unsupervised approaches, we aim to fill this research gap. Our objective is to develop methods that not only provide accurate sentiment classification but also offer insights and explanations for the decisions made. By focusing on unsupervised techniques, we aim to reduce the reliance on labeled data, making our solutions more adaptable and applicable across diverse domains. Moreover, the interpretability of our proposed methods will enable users to gain a deeper understanding of the opinion miming and sentiment analysis process, instilling greater confidence in the results and facilitating better decision-making.

4.5.2 Association rules and sentiment analysis in social media

Upon analyzing existing approaches in opinion mining, we identified a potential gap that can be filled by association rules, offering both interpretability and non-guided techniques. Association rules have demonstrated success in various social media mining systems, making them a promising avenue for our research.

In the realm of social networks, association rules have been applied effectively, as demonstrated in

papers like the one proposed by Erlandsson et al. [EBBJ16]. This study utilizes association rules to identify influencers on Facebook, showcasing the power of this technique for summarizing information and identifying patterns in social networks. Similarly, association rules have been used for information summarization on Twitter, aiming to identify influential users [ADEZ18] and relevant tweets [PNH18]. However, these studies encountered limitations in handling large volumes of data, hampering their applicability to real-world scenarios with higher data and variable volumes. To overcome these limitations, certain works have embraced the realm of big data, providing promising solutions. For instance, Adedoyin-Olowe et al. [AOGD$^+$16] and Fernandez-Basso et al. [FBFAMBR19] employed association rules and frequent itemset extraction, respectively, in streaming data to detect events in sports and politics.

Considering the accomplishments of the aforementioned papers, our attention turns to developing a new opinion mining system based on association rules. Our focus is twofold: first, to create a system capable of summarizing vast amounts of data, in line with research advocating the potential of association rules in summarization [PNH18]; and second, to design a system capable of handling substantial volumes of data effectively.

One of the first attempts to apply association rules to sentiment analysis is the work of Yuan et al. [YOXS13]. This work proposes a new measure to discriminate frequent terms with ambiguous sentiment orientations, facilitating subsequent sentiment analysis. A related study by Dehkharghani et al. [DMJS14] uses association rules to establish links between co-occurring terms in tweets, subsequently classifying them based on the sentiments associated with the obtained rules.

Shifting focus to sentiment analysis, two methods employ a mixed approach of association rules and sentiment analysis to uncover patterns on Twitter. Mamgain et al. [MPM16] and Bing et al. [BCO14] adopt this approach by first conducting sentiment analysis to associate sentiments with each item and then applying the Apriori algorithm to obtain patterns. Mamgain et al. use their model to assist students in choosing the best college in India, while Bing et al. apply it for stock market prediction. While the strength of using both tools from a mixed approach is evident, these proposals suffer from limitations due to the limited number of tweets employed and their specific domain of application.

In recent trends, we observe an inclination towards extended association rules (see Section 2.2). For instance, Hahsler and Karpienko [HK17] propose a matrix-based visualization using a hierarchical simplification of the items forming association rules. Other approaches, like those proposed by Cagliero and Fiori [CF12, CF13], explore dynamic association rules and generalization of rules according to taxonomies such as places, time, or context on Twitter data. These extended association rule techniques offer promising possibilities for content propagation and evolution analysis over time.

Based on the insights gained from this comprehensive review, we recognize that association rules possess significant summarization potential, particularly in big data environments. We also acknowledge the necessity for adaptable systems in the dynamic realm of social media and information needs. The potential of extended association rules further strengthens our objectives, leading us to develop a system capable of generalizing association rules by sentiments, providing comprehensive summarization of big data datasets, and extracting sentiments across multiple topics without the need for training.

4.5.3 Credibility

In social media environments, dealing with challenges like misinformation, noise, and vast amounts of data is a constant concern. Addressing these factors can be viewed as an instance selection problem [OLCOMTK10, CS14], guided by intrinsic social network factors such as user engagement and expertise in specific topics [CSP11]. This area can be divided into two main branches: content filtering and user filtering based on credibility.

Castillo et al. [CMP11] have addresed the problem of credibility on Twitter. Their research is one of the most comprehensive and attempts to classify contents (tweets) based on whether they are credible or not. To evaluate and create the model, they use a large number of indicators that are closely linked to the analysis that a human would do to study the credibility of a tweet, such as whether an account is verified, whether the user has enough followers or whether the tweet uses appropriate hashtags, to name some of the features taken into consideration by the classification system. Concerning the user, it also obtains information but in a very simplistic way: for example, if it has biography or if it is empty. At this point we found that we could do a important contribution, taking all the potential of biographies to guide the process of filtering using word embeddigs.

Kang et al. [KOH12] propose three different methods for obtaining credibility ratings. The first method analyzes the social graph of Twitter, looking at ratios between concepts like retweets and number of followers. The second method focuses on content, while the third is a hybrid model considering both graphs and content, with the graph-based model performing better. Additionally, there is a credibility-oriented dimension for event-related content. Hassan [Has18] uses text mining techniques on event-related tweets, guided by the frequency of terms in different topics and evaluated using various classifiers.

In the domain of user credibility, we found approaches like Cognos and CredSaT. Cognos [GSB+12] provides a web solution for searching experts in specific topics using Twitter lists, where users add others related to certain subjects. CredSaT [ASWCZ19], a big data solution, considers content and timestamps to rank expert and influential users. It also includes a semantic analysis layer with sentiment analysis to enrich the corpus of experts. Finally, it is necessary to mention the papers proposed by Alrubaian et al. [AAQA+16, AAHA18]. These papers also deal with the analysis of credibility on Twitter in a very exhaustive way through 3 modules. These modules deal with content credibility, reputation and expertise. Notably, the works proposed by Alrubaian et al. [AAQA+16, AAHA18] comprehensively tackle credibility analysis on Twitter through three modules focusing on content credibility, reputation, and expertise.

Among the areas related to credibility, we discovered that it is more efficient to filter users based on their credibility, as this automatically filters their content, saving computational time and improving accuracy. Thus, our focus is on identifying the most relevant users for a given analysis. To effectively assess user credibility, we recognize the need for multiple metrics, considering the complexity of determining which users are genuinely interesting. Our focus revolves around expertise, engagement, and credibility as essential factors in this process. Leveraging biographies, which professionals often use to share information about their expertise, serves as a valuable resource for our framework, distinct from approaches solely based on content as seen in prior works [AAQA+16, AAHA18]. This emphasis on biography-driven credibility assessment contributes to the precision and effectiveness of our research.

4.5.4 Misinformation detection

Detecting non-credible users is crucial in reducing misinformation and fake news, as such users are likely to share unreliable information. By identifying and addressing non-credible users, we can also tackle the spread of misinformation effectively. In this section, we aim to compile cutting-edge papers on this topic and align our research with the opportunities they present.

Supervised approaches dominate the field of misinformation and fake news reduction. For instance, in [OA20], Ozbay and Alatas employ 23 different classification algorithms on a labeled fake news dataset from the political scene, achieving promising results. Similarly, in [CPM+19], authors explore various classification methods, including traditional decision trees and neural networks, with impressive outcomes. In the context of classification, [BJTC21] proposes a classifier based on K nearest neighbors and bio-inspired algorithms to detect worthless information, particularly in email spam, using different distance metrics.

Deep learning approaches have also been widely utilized to detect fake news and rumors. Early works, such as [MGM+16], present architectures like Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) layers and Gated Recurrent Unit (GRU) layers, which significantly improve model performance. However, these models require retraining for different domains due to the need for labeled datasets. On the other hand, various deep learning-based papers, such as [MSADLG18, Kal18, MFE+19], train classifiers on pre-labeled fake and non-fake databases. While these approaches achieve high accuracy, they suffer from low generalizability and interpretability since they rely on pre-tagged datasets that might not be representative of dynamic social media environments.

In addition to classification through deep learning, several novel approaches have emerged for detecting misinformation. One such proposal, as described in [KAGE21], leverages concepts like novelty and emotions to guide the detection process. The foundation of this approach lies in the observation that misinformation tends to be emotionally charged to facilitate its spread. To achieve this, the authors combine powerful techniques such as BERT, LSTM, and feed-forward networks in their model. [NKV21] introduces a unique combination of neural network topologies for fake news detection. They employ Convolutional Neural Networks (CNNs) to extract fake news features, and subsequently, they use Recurrent Neural Networks (RNNs) to capture the sequential dependencies between terms. The resulting output is then utilized for fake news classification. In [KCQJ20], authors focus on early detection of the spread of fake news by analyzing Twitter conversations surrounding such content. They employ neural networks to effectively identify and address the propagation of fake news in its initial stages. In [LW20] Liu et al. introduces a unique approach by concatenating user-related features and user-generated text. This combination is used as input for a rumor classification layer, where word embeddings play a significant role in the analysis.

We aim to offer solutions that do not require extensive training and are interpretable. Instead of relying solely on deep learning, we leverage word embeddings in an unsupervised manner to reduce irrelevant data and address non-credible information. We emphasize a fusion of user-based features (e.g., number of favorites or retweets) and text-based features (e.g., experience-related words in biographies) to enhance the accuracy of our systems, aligning with the findings in [LW20]. By taking this approach, we can provide society with practical and interpretable systems to combat misinformation and non-credible information effectively.

5 Contributions

In this section, we present a comprehensive summary and discussion of the main contributions of the thesis, aligning them with the objectives outlined in Section 2. The following notable contributions have been made, with publications in high-impact international journals:

- Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2023). A survey on the use of association rules mining techniques in textual social media. Artificial Intelligence Review, 56(2), 1175-1200.

  This publication presents a comprehensive survey on the utilization of association rules mining techniques in textual social media. The study explores and analyzes existing methods, providing valuable insights into the applicability and effectiveness of these techniques in social media text analysis.

- Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2020). Non-query-based pattern mining and sentiment analysis for massive microblogging online texts. IEEE Access, 8, 78166-78182.

  In this work, we proposed new techniques based on association rules for non-query-based pattern mining and sentiment analysis that are applied to massive microblogging online texts. The research demonstrates innovative approaches to efficiently analyze sentiment and patterns in large-scale data, contributing to improved understanding of user sentiments in microblogging platforms in an unsupervised way.

- Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2022). NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise. Expert Systems with Applications, 208, 118063.

  By means of word embeddings, the NOFACE framework offers a novel solution for filtering irrelevant content and opinions in social media based on credibility and expertise. By introducing this new approach, the research enhances the reliability and quality of information disseminated through social media environments.

- Jose A. Diaz-Garcia, Karel Gutiérrez-Batista, Carlos Fernandez Basso, M. Dolores Ruiz & Maria J. Martin-Bautista. (2023) A flexible big data system for credibility-based filtering of social media information according to expertise. Submitted to: International Journal of Computational Intelligence Systems.

  The flexible big data system proposed in this publication leverages credibility-based filtering to enhance the processing of social media information, particularly with regard to expertise. This contribution provides a scalable and efficient solution for managing vast amounts of social media opinions while considering the credibility and expertise of the opinion sources.

These contributions significantly advance the state-of-the-art in opinion mining. They align with the objectives outlined in Section 2 and represent substantial progress in addressing the challenges posed by large-scale unlabelled data in social media environments.

In next section, we focus on the main contributions and results of our experimentation, as the theoretical implications have been discussed in Section 4. By highlighting the progress made and the impact of our research, we aim to underscore the significance of our research.

5.1 A survey on the use of association rules mining techniques in textual social media

The first objective (**O1**) of our thesis is to conduct an in-depth study of the state-of-the-art techniques in text and opinion mining, with a particular emphasis on approaches that can be effectively applied in an unsupervised manner to tackle the inherent challenges of social media environments. To accomplish this, we embarked on an extensive process of gathering, reading, and analyzing papers and relevant information on these techniques. The insights and findings from our comprehensive literature review were compiled in the related works section of our thesis (Section 4.5) and were also published in our research papers [DGRMB20, DGFBRMB20, DGRMB22].

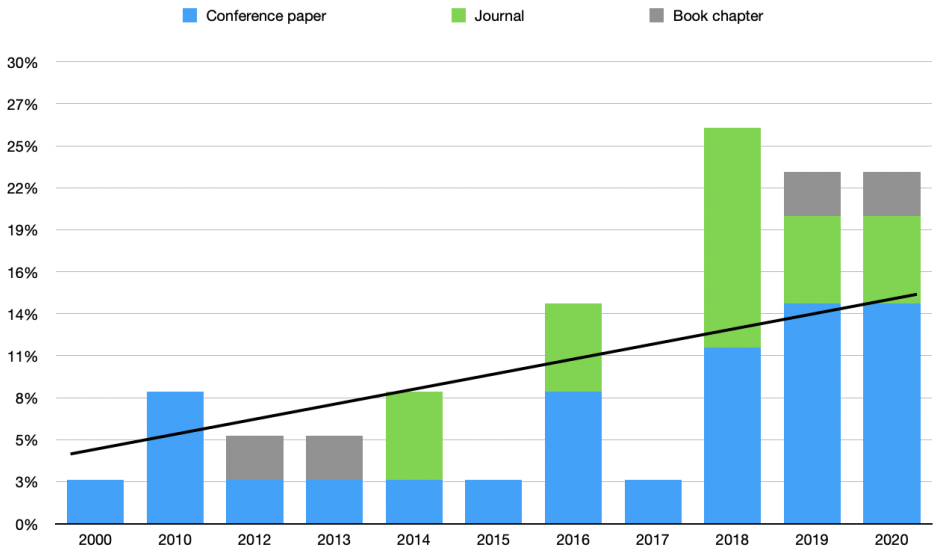

Figure 2: Papers according to their type and year of publication in social media mining using association rules

During our investigation, we observed a notable trend in the application of association rules in the analysis of social media text, which has gained increasing significance in recent years (Figure 2). Recognizing the relevance and potential impact of this research direction, we decided to delve deeper into this area by conducting a dedicated and thorough survey on the utilization of association rules mining techniques for texts coming from social media environments. As far as we know, this is the first survey that addresses social media mining with association rules. The paper also focuses on current challenges that are being addressed by the association rules and that open up promising avenues for future research.

The survey was conducted using the Systematic Literature Review methodology [BB06], which provided a structured and rigorous approach to gathering and analyzing relevant research papers. Our research questions were carefully formulated to align with the main objectives of our thesis, ensuring that the survey's focus was directed towards obtaining valuable insights and addressing specific aspects related to the state-of-the-art techniques in text and opinion mining. Namely, our research questions were:

- **RQ1**: What tasks are currently being solved with association rules in user-generated text?
- **RQ2**: What areas or fields of application have been addressed with association rules in

user-generated text?

- **RQ3**: What are the current trends and future problems to be faced by association rules in social media mining?

While addressing these questions, we acquired invaluable insights that significantly informed our approach to our second objective (**O2**). Through our systematic literature review, we gained a comprehensive understanding of the challenges and advancements in opinion and text mining within social network environments, as well as the current trends being addressed by the academic and scientific community in this field. This insightful exploration guided our proposals and provided valuable insights into potential problem areas and future research directions.

The results and conclusions of the paper can be effectively summarized and aligned with the responses to our research questions as follows:

- **RQ1**: Association rules have been prominently applied in various tasks, including summarization, event and topic detection, sentiment analysis, forecasting, and collaborative social systems. While there may be other applications, these are the most relevant ones, attracting a substantial number of articles in the literature.
- **RQ2**: Association rules have been effectively applied in diverse fields, encompassing academia, crime detection, healthcare, insurance, influence analysis, leisure activities, natural disaster management, politics, sports analytics, transportation, and trend analysis.
- **RQ3**: The future prospects for association rules include exploring streaming association rules, temporal association rules that consider time dependencies, and extended association rules incorporating social flags or graph-based structures to model complex relationships in social networks.

Our survey provided valuable insights into the applications and potential progress of association rules in the realm of textual social media mining. These findings serve as a solid foundation for our subsequent proposals and research direction.

The comprehensive survey results have been published in (Chapter II, Section 1):

> Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2023). A survey on the use of association rules mining techniques in textual social media. *Artificial Intelligence Review*, 56(2), 1175-1200.

5.2 Non-query-based pattern mining and sentiment analysis for massive microblogging online texts

After conducting an exhaustive review of pattern mining in text from social networks, we observed that association rules were prominently applied in fields such as influence studies, politics, and trend analysis. Additionally, we identified extended association rules by social flags as a significant area that requires attention and development. To address these challenges and specific fields of application, we concentrated our efforts on providing a solution based on association rules. By doing so, we can effectively achieve our second objective (**O2 and O2.1**) and contribute to the advancement of the research in this domain.

Therefore, our main focus is to make a significant contribution to the state of the art by introducing an innovative technique that utilizes extended association rules by social flags to tackle various challenges in fields like influence and trend analysis, particularly in political applications and beyond.

5.2.1 Generalized association rules for sentiment analysis

Aligned with our objective 2 (**O2**) and based on the findings from the survey [DGRMB23], we aim to unify extended association rules, particularly Generalized Association Rules (GAR), with social flags and sentiments, to develop a comprehensive and unsupervised framework for sentiment analysis in social media. By combining these elements, we created a powerful and versatile solution that can effectively analyze sentiments in social media data, taking into account the diverse vocabulary and content variability present in these big data environments.

Extended association rules encompass various types of rule extensions, such as GAR, quantitative association rules, temporal or sequence association rules, and N-gram/N-ary association rules. In social media environments, addressing the variability of content is a crucial challenge, given the vast amount of different vocabularies. Traditional association rules may not be the most efficient solution to handle this issue. Thus, to contribute an innovative technique to the state of the art in extended association rules, we focus on GAR. GAR allows for reducing the vocabulary and providing an interpretable and efficient solution by aggregating elements into groups and mining association rules accordingly.

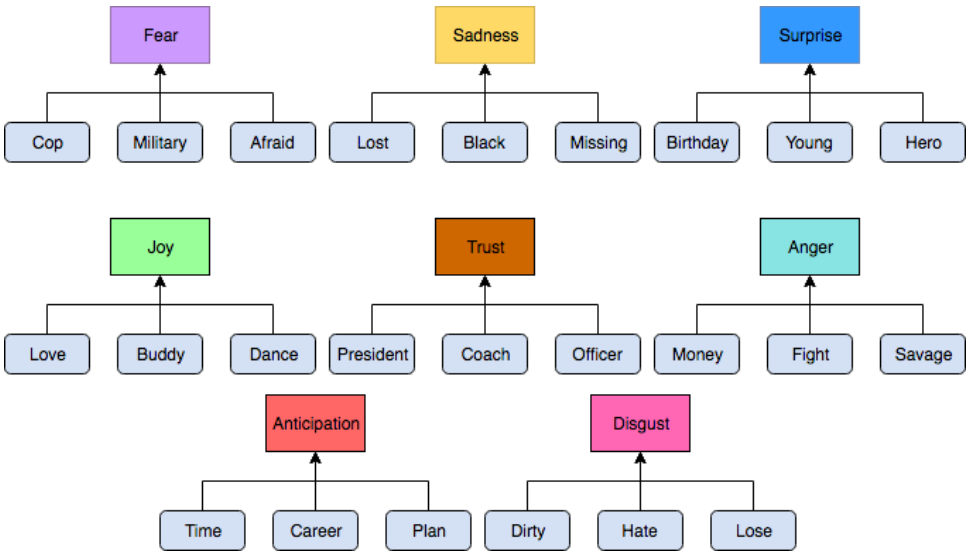

Figure 3: Generalized words based on emotions.

GAR involve studying association rules from a hierarchical perspective [SA97]. For instance, in a shopping basket, the rule *Apples, Bananas* $\rightarrow$ *Yogurt* could be generalized as *Fruit* $\rightarrow$ *Yogurt*, offering a higher level of data abstraction and revealing new relevant information. This abstraction enables us to summarize the set of rules significantly, simplifying analysis without losing valuable insights, as the original details can be easily recovered. In big data environments like social media, where vast amounts of data are prevalent, GAR becomes especially advantageous as it drastically summarizes the results, leading to improved processing time and resource utilization, resulting in stronger rules.

Our approach is based on the generalization of association rules, considering the words associated with emotions in the antecedent or consequent of discovered rules. To accomplish this, we utilize the *syuzhet package* in R programming language, which incorporates powerful sentiment dictionaries like the NRC Word-Emotion Association Lexicon by Saif M. Mohammad [MT13]. The lexicon-based approach iteratively assigns emotions to words in each tweet, creating a data structure that counts occurrences for each emotion associated with a word. Subsequently, a majority sentiment is assigned to each word. In the final stage of our sentiment analysis using association rules, we replace words with their majority emotion (see Figure 3) before mining the association rules.

The framework for obtaining generalized association rules based on sentiments was published in (Chapter II, Section 2.2):

> Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2019). Generalized association rules for sentiment analysis in twitter. In *Flexible Query Answering Systems*: 13th International Conference, FQAS 2019, Amantea, Italy, July 2–5, 2019, Proceedings 13 (pp. 166-175). Springer International Publishing.

The framework for obtaining generalized association rules by sentiments proved to be highly effective in capturing emotions and sentiment patterns within social media text data. To validate its capabilities, we conducted tests on a political dataset related to the elections between Trump and Hillary Clinton. The results demonstrated the versatility and utility of the solution, particularly in measuring the influence of certain characters on social networks. These promising outcomes motivated us to develop a more comprehensive system that integrates the core of GAR by sentiments, aiming to provide a more valuable and powerful solution. This contribution stands as one of the most significant advancements in the state of the art for our thesis, and its details will be thoroughly presented in the following section.

5.2.2 Non-query-based proposal

To enhance the utility and precision of our solution for Objective 2.1 (**O2.1**) we focused on two important conclusions that emerged from our review of the state-of-the-art in opinion mining and social media data analysis:

- Firstly, we recognized the necessity and potential of extended association rules to tackle complex social media problems, as discussed in the previous section. These rules offer greater data abstraction, enabling us to summarize and adapt to different scenarios without requiring extensive training or complex query systems.
- Secondly, we observed the predominance of highly accurate supervised techniques, often based on deep learning classifiers or SVMs, for certain tasks such as sentiment analysis. While these techniques yield impressive accuracy, they are limited in their adaptability to rapidly changing data in the dynamic world of social media and big data problems.

In the big data era, organizations and companies accumulate vast volumes of data, often without immediate knowledge of the insights it may hold. Data lakes are increasingly employed to store this non-relational data, collected without predefined queries. Many of these data lakes are sourced from social networks, capturing user-generated conversations and content related to the organization. Therefore, the demand for systems capable of handling and extracting valuable information from

this data is heightened. At this point, we define and differentiate between query-based and non-query-based systems.

Definition 2. *A query-based system: The system or framework starts with a predefined set of data limited to a specific domain, and subsequent pre-processing and data mining tasks are carried out to derive insights from this known and limited dataset.*

Definition 3. *A non-query-based system: The system or framework does not impose any filters during data collection, resulting in a massive data lake. The need for specific information emerges later and guides the data mining process through pre-processing. While data pre-processing becomes more challenging due to the sheer volume of data, non-query-based systems offer numerous possibilities for extracting inter-topic and cross-subject knowledge.*

Non-query-based systems are particularly well-suited for big data problems, especially those originating from social networks. The vast amount of rapidly generated data makes it impractical to determine pertinent queries beforehand. As a result, non-query-based systems become the optimal solution for social media applications or scenarios where the subject of interest is not fixed in advance. To better illustrate these distinctions, a comparison between query and non-query-based systems can be found in Figure 4.

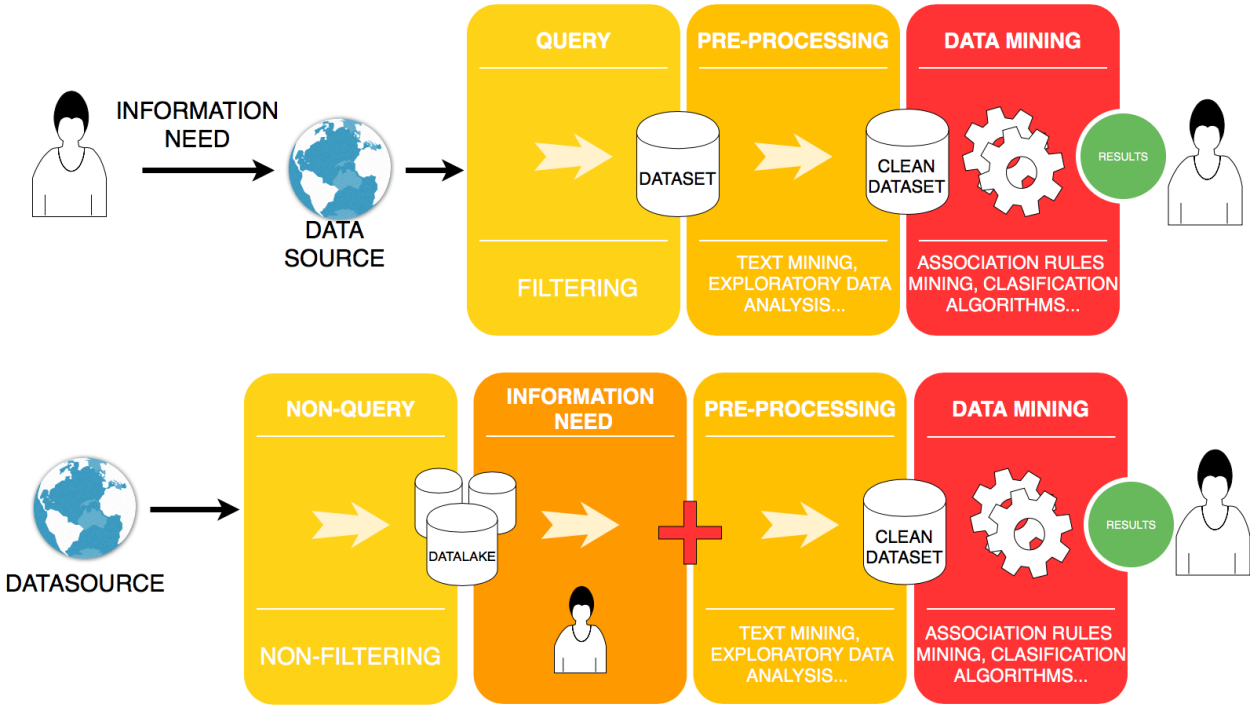

Figure 4: Comparison between query and non-query based systems

.

Considering all the aspects discussed above, the system was designed to fulfill the following requirements:

- Robustness: The system had to be robust enough to handle low-quality datasets since the data lake comprises data referencing several different topics. This ensures that the system can effectively process and extract valuable information even from diverse and noisy data sources.

- Opinion entity recognition: The system needed to provide a mechanism to distinguish between people, places, and brands to filter the dataset once the specific information of interest is defined. This filtering step is crucial in focusing the analysis on relevant entities and aspects.

- Opinion aspects summarizing a sentiment analysis: Finally, the system had to offer a way to summarize the information related to the identified entities and aspects. This includes providing sentiment analysis and valuable insights regarding the evaluated aspects.

To fulfill these requirements, the system comprises five main modules (Figure 5). The first module is responsible for pre-processing the raw data to prepare it for further analysis. The second module focuses on unsupervised named-entity recognition using dictionaries, enabling the identification of entities that receive opinions within the text.

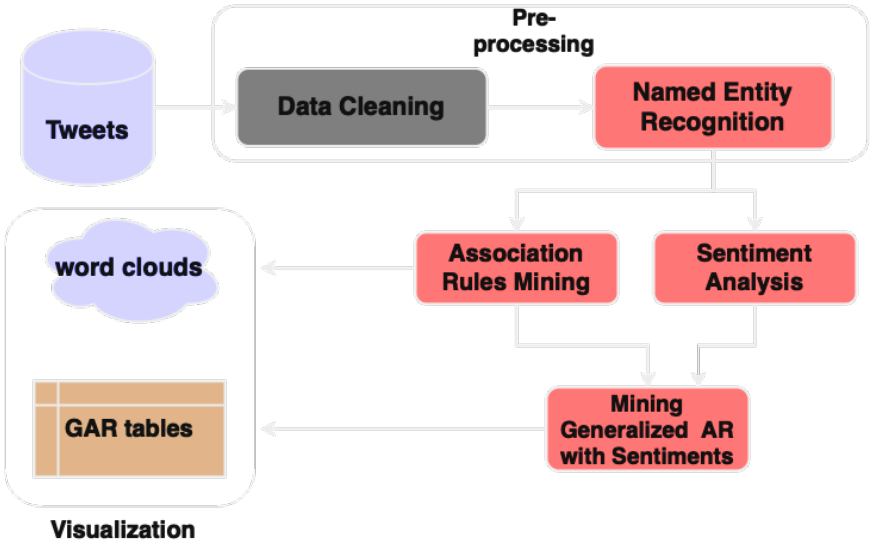

Figure 5: Methodology flow.

The three final modules form the central core of the proposed system. In the third module, we leverage association rules using either the Apriori or FP-Growth algorithm, depending on the volume of the data, to summarize the aspects. This step allows us to efficiently identify patterns and associations in the data, generating meaningful rules that capture the relationships between different entities (consequents) and aspects (antecedents). The fourth module, assign emotions to each word in the corpus leveraging a majority vote.

In the final module, we generalize the words of the consequent using the emotion located in the previous stage using the thecnique discussed in Section 2.2. By integrating these modules, the system can effectively process and analyze social media data, extracting valuable information about different entities, aspects, and their associated sentiments in an unsupervised and flexible way.

The proposed system and its application to a political forecasting use case were published in (Chapter II, Section 2.1):

Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2020). Non-query-based pattern mining and sentiment analysis for massive microblogging online texts. *IEEE Access*, 8, 78166-78182.

5.2.3 Non-query system validation in real problems

To validate our proposed non-query system for pattern mining and sentiment analysis, and to achieve our objective 3 (**O3**), we applied the system to a real use case of political forecasting, specifically related to the US elections between Donald Trump and Hillary Clinton. For this use case, we collected a random sample of tweets, applying only basic filters such as location (EEUU), language (English-speaking), and time (between January and June 2016). As the collection of tweets was done without applying any keyword filter, the sample is diverse and may contain low-quality content. The final dataset comprised 1.7 million tweets.

Since the time period corresponds to the US election campaign, we focused on obtaining insights related to Donald Trump and Hillary Clinton. Analyzing all the generated rules for proper names on Twitter during this period would be exhaustive, so we chose these two names as examples. Given that the chosen names are associated with the political world, and we have knowledge of the electoral results, our aim was to extract insights and information about these characters using our frameworks and compare them with real political events and news.

By examining the associations and sentiments surrounding Donald Trump and Hillary Clinton during the election campaign, we aim to gain valuable insights into public opinions and reactions towards these political figures on social media. By applying our system to this real data set, we were able to evaluate its effectiveness and potential contributions to sentiment analysis in social media and big data environments, ultimately validating the capabilities of our proposed approach and aligning it with real life events.

We applied our framework and obtained various visualizations over the data. Specifically, we generated traditional tables with the association rules and utilized more interpretable techniques, such as word clouds based on the rules. In Figure 6 and Figure 7, we present the results for the words present in the association rules related to each candidate. Considering that these visualizations come from a non-filtered dataset with low quality, the quality of our results is remarkably high, thanks to the power of our non-query process, which efficiently locates relevant information about the candidates and summarizes it using association rules in an interpretable and unsupervised manner. It is worth mentioning that the same study could be applied to other names as well.

If we look at the representation of Trump's rules using the word clouds (Figure 6), even a person without prior knowledge about the topic can deduce what is being discussed on Twitter and identify the main trends related to the candidate. For example, we notice the presence of words like *transgender, rape, child*, from which we can derive important trends through an inspection of the rules. Additionally, the word *Iowa* appears as relevant, a word that we previously overlooked in the manual process. Thanks to this graphic, we now see that it holds significance. Upon investigating the rules containing Iowa in the antecedent, we observe that this state played a decisive and highly contested role during the presidential elections, as polls and public opinion continuously generated information about it. In Figure 7 we found the results for Hillary Clinton, where we can easily explore other trends or patterns. In this case, we can derive conclusions, such as the importance of the relationship between the opinions of Bernie Sanders and Hillary Clinton herself. We also notice the reappearance of Iowa, as expected from our earlier study of the word cloud related to rules involving Donald Trump, since both candidates competed for votes in that state, resulting in bidirectional related rules. The word cloud representation provides a more intuitive and accessible way to identify relevant patterns and insights from the data.

he final stage of our methodological testing on the real use case of the elections involved the use of Generalized Association Rules for sentiment analysis. In Table I.3 and I.2, we present the sentiment rules obtained for each candidate. It is important to note that we have automatically condensed the

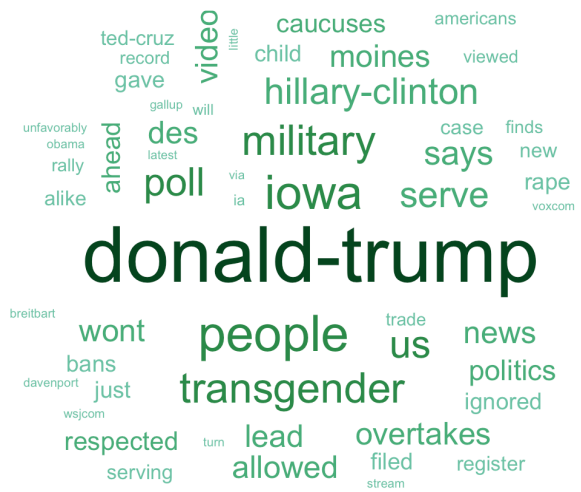

Figure 6: Word Cloud for association rules related to Donald Trump

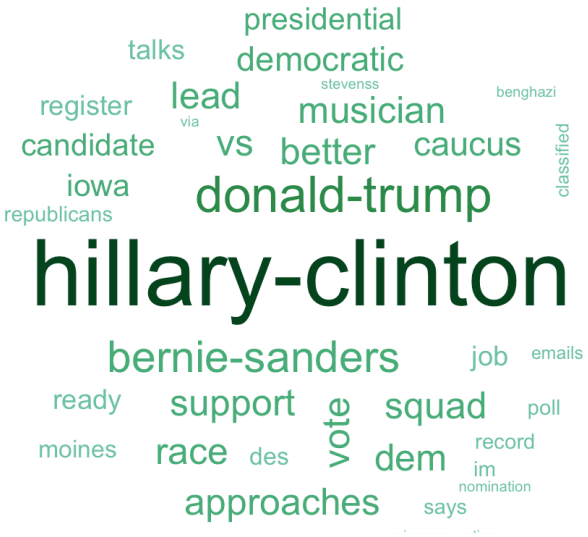

Figure 7: Word Cloud for association rules related to Hillary Clinton.

entire dataset related to each candidate into a set of 8 rules in an unsupervised manner. The support value associated with each rule serves as a metric representing the percentage of the sentiment's presence for each candidate.

This process of generating sentiment rules in an unsupervised and automatic manner is a significant contribution of our thesis, offering a new technique for mining sentiments from massive amounts of microblogging texts. These sentiment rules provide valuable insights into the sentiments expressed towards each candidate, enabling a deeper understanding of public opinion during the election period.

Antedecent	**Consequent**	**Supp**
$\{trust\}$	$\{hillary\text{-}clinton\}$	0.939688
$\{anger\}$	$\{hillary\text{-}clinton\}$	0.492217
$\{anticipation\}$	$\{hillary\text{-}clinton\}$	0.486381
$\{fear\}$	$\{hillary\text{-}clinton\}$	0.299610
$\{surprise\}$	$\{hillary\text{-}clinton\}$	0.200389
$\{joy\}$	$\{hillary\text{-}clinton\}$	0.145914
$\{sadness\}$	$\{hillary\text{-}clinton\}$	0.079766
$\{disgust\}$	$\{hillary\text{-}clinton\}$	0.077821

Table I.2: Rules based on sentiments about Hillary Clinton

In a deeper analysis of our sentimental association rules over the political forecasting problem we made a significant observation that challenged common social assumptions. Contrary to popular belief, our findings often pointed to different outcomes than society had assumed. Remarkably, as events unfolded, our findings turned out to be more accurate than the prevailing social opinions. For instance, we detected a higher number of negative sentiments against Hillary Clinton, while the general perception in society leaned towards an improbable Trump victory.

These intriguing findings prompted us to delve deeper into the nature of social media data. We realized that a considerable portion of the content on these platforms might be fake, non-credible, or

Antedecent	Consequent	Supp
{*trust*}	{*donald-trump*}	0.945927
{*anticipation*}	{*donald-trump*}	0.594113
{*surprise*}	{*donald-trump*}	0.425051
{*anger*}	{*donald-trump*}	0.345656
{*fear*}	{*donald-trump*}	0.295003
{*joy*}	{*donald-trump*}	0.226557
{*disgust*}	{*donald-trump*}	0.112936
{*sadness*}	{*donald-trump*}	0.074606

Table I.3: Rules based on sentiments about Donald Trump

irrelevant to specific topics.

Our initial approach involved applying our developed techniques to a manually labeled dataset focused on fake news. This allowed us to gain valuable insights into the characteristics and fabrication of fake news and opinions. Through this analysis, we made two significant observations. Firstly, we found that fake news are often created by modifying real news, altering small elements to create misleading information. Association rules, such as {*sexism, won*} → {*electionnight, hate*} for fake news and {*sexism, won*} → {*electionnight*} for real news, effectively captured these patterns. Secondly, we noticed that the rules associated with fake news tended to include more items, possibly due to the sensationalist elements often present in fake news content.

Surprisingly, we discovered that distinguishing between fake and real news based solely on content similarities posed challenges. Therefore, we redirected our focus towards analyzing user-related factors to address the problem of fake news and unreliable information in social networks.

The insights obtained from this first approach were compiled in (Chapter II, Section 2.3):

Díaz-García, J. A., Fernandez-Basso, C., Ruiz, M. D., & Martin-Bautista, M. J. (2020, June). Mining text patterns over fake and real tweets. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 648-660). Springer International Publishing.

5.3 NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise

In response to the findings obtained in [DGFBRMB20] and starting to achieve our second objective (**O2.2**) of developing systems based on word embeddings, we began exploring methods to address the challenge of identifying and handling non-relevant, non-credible, and misleading opinions.

By understanding the prevalence of unreliable and potentially misleading content, we sought to enhance the quality of our opinion mining techniques. Our goal became not only to analyze sentiments but also to distinguish credible information from unreliable sources. This endeavor led us to design an innovative filter based on word embeddings to better identify and filter out non-credible and disinformative opinions taking into account the user that issued the opinion or the opinion holder.

To accomplish this, we focused on leveraging user information available on social networks, such as the number of followers, retweets, likes, or the text of user biographies and descriptions. We

decided to utilize biographies because they often contain relevant information about users' professions and positions, making them valuable for our analysis. Our assumption was that professionals use social networks to disseminate content related to their work and findings, which could provide trustworthy insights on specific topics, such as medicine. The central core of our framework is to be able to know which people talking about a certain topic on Twitter really have a relationship in terms of experience and credibility with the topic under study. Modelling credibility and expertise is an arduous task, approached by other papers in a quite exhaustive and efficient way [AAQA+16]. Our framework is based on the premise that if a Twitter user is credible, his or her content is also credible and therefore the user can be used to guide the selection of the content.

We believed that using word embeddings on biografhies could enable us to identify which words were closely related to the input topic and then improve the selection of trustworthy users who could provide relevant and non-fake opinions on that topic (Algorithm 1). Thus, the initial stage of our research involved finding the most optimal word embeddings for our purpose. We conducted a comparative study between two widely used word embeddings, Word2Vec and FastText, to determine their effectiveness in enhancing our credibility filter. Table I.4 and Table I.5, contain the intervals of results in the range of minimum and maximum values obtained during the experimentation.

	Word2Vec+CBOW	**Word2Ved+Skip-Gram**
Elapsed Time	Min: 11min 7s Max: 13min 36s	Min: 13min 1s Max: 14min 22s
Words	Min: 209 Max: 228	Min: 45 Max: 64
Users located	Min: 21390 Max: 23377	Min: 7937 Max: 10044
Final dataset size	Min: 31347 Max: 34107	Min: 12281 Max: 15239

Table I.4: Minimum and maximum value for each variable in the Word2Vec experiments.

	FastText+CBOW	**FastText+Skip-Gram**
Elapsed Time	Min: 16min Max: 17min 15s	Min: 18min 20s Max: 20min 10s
Words	Min: 50 Max: 79	Min: 111 Max: 125
Users located	Min: 10975 Max: 12312	Min: 18128 Max: 20306
Final dataset size	Min: 16748 Max: 18773	Min: 26547 Max: 31611

Table I.5: Minimum and maximum value for each variable in the FastText experiments.

Upon analyzing the results, it becomes evident that FastText is more time-consuming than Word2Vec due to its n-gram decomposition. In terms of users located (i.e., users related to the topic under filtering and analysis), we observe a divergence between the models. Two models, Word2Vec+Skip-Gram and FastText+CBOW, can be immediately discarded as they offer a very restrictive behavior and find fewer words related to the domain, resulting in fewer results. This leads us to conclude that each algorithm performs better with a different inference model for their words. FastText performs better at predicting context words on a one-word basis, while Word2Vec excels at predicting one word based on several context words. Given our problem, where we have

few context words due to the limited size of Twitter texts, this difference plays a significant role. Predicting context words based on a single word (Skip-Gram) is easier for the algorithm than predicting a single word based on several words (CBOW). Although Word2Vec+CBOW obtains great results, the ratio of users found for each word is lower (102 users per word on average) compared to FastText+Skip-Gram (167 users per word on average). This indicates that the words located by FastText have a higher representation in the dataset. Consequently, the best option for our algorithm will be to use **FastText+Skip-Gram**, despite being more time-consuming, as this increase is associated with a higher match value for the selected words, their relation with the topic, and a better user selection ratio.

The comparison between models to fine-tune our final algorithm was published in (Chapter II, Section 3.2):

> Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2021, September). A Comparative Study of Word Embeddings for the Construction of a Social Media Expert Filter. In *International Conference on Flexible Query Answering Systems* (pp. 196-208). Springer International Publishing.

5.3.1 NOFACE framework

In this section, we will explore the NOFACE framework (NOise Filtering in social media According to Credibility and Expertise). This framework covers completely objective 2.2 (**O2.2**). Figure 8 provides an overview of our framework, which operates on Twitter databases and employs a series of cascading modules. NOFACE consists of three modules, along with a pre-processing module, applied in a cascading manner. This means that users who fail to pass the first filter will not proceed to the subsequent ones, resulting in reduced computation times and a more reliable solution. These modules are equipped with restrictive filters specifically designed to address the issue of deceptive or misleading information. For example, a user may falsely claim to be a doctor in their biography, misleading the algorithm into considering them an expert in the field of medicine. To counter such cases, subsequent filters are employed to evaluate and mitigate such deceptive practices. By measuring engagement and content credibility, we can identify dishonest users and exclude their content from further stages of analysis. This ensures that only trustworthy and credible sources of information proceed through the NOFACE framework, ultimately enhancing the overall quality and reliability of the insights obtained.

The framework first applies a pre-processing method. It then computes the expertise through biographical analysis, then focuses on content quality (engagement), and finally filters based on the user's credibility on the topic. The value of text mining and credibility assessment is emphasized by [KKI21], where various approaches are mentioned that support the validity of our approach. For instance, in [BAI19], the authors identify strong associations between certain social features, such as the number of followers or inclusion in lists, and more trustworthy content. This evidence motivates us to develop a system that combines user features and expertise to filter out irrelevant or untrustworthy information from social networks effectively. By doing so, NOFACE aims to enhance the reliability and quality of insights extracted from social media data.

The first filter and therefore the main filter and greatest contribution to the state-of-the-art of this approach is the **expertise** filter. This filter is aimed to locate and eliminate those users who are not really related to the topic. Algorithm 1 shows the pseudo-code of our proposal. The expertise filter uses the power of semantic relations between words located by FastText to increase the search

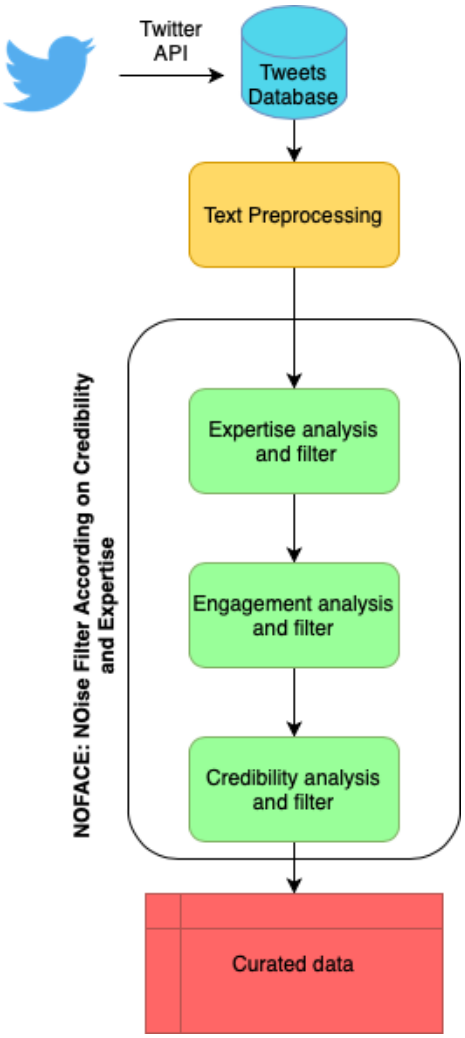

Figure 8: NOFACE framework.

space in Twitter biographies. Many works have demonstrated the power of word embeddings to expand search queries in the information retrieval process [KSK16, DMC16, RPMG16]. In our algorithm, we use this potential of word embeddings, not for retrieving documents, but for locating users that are experts on a topic.

Our algorithm works as follows. We introduce a list of words related to the topic under study, for example, about medicine, we can introduce words: *medical, doctor*. The algorithm will start to train a word embedding model on the biographies using a part of a data partition (10% of the entire dataset), and in the first iteration, it will obtain the 5 most similar words to *medical* and *doctor* among the corpus itself. The algorithm will use these 12 words, (5 more similar to medical, 5 more similar to doctor, besides doctor and medical), to find users whose biographies contain any of these terms and start creating the list of experts and topic-related users. In the next iteration, we will already have 12 words to search for their 5 similar ones, and so on. In each iteration, the algorithm checks if any user id is already present in the expert list to avoid processing it again, since its words and content are already in the search corpus, thus avoiding additional processing.

The next filter concerns **engagement**. Engagement could be defined as the capacity of a user to generate useful content that is appreciated by other users of the social network. In other words,

Algorithm 1 Expertise filter algorithm

```
Input: cleaned-dataset: Preprocessed dataset
Input: expert_words: Words from each expert
Output: Dataframe with experts in the topic
expert_set = []
finaldataframe = pd.dataframe()
split cleaned-dataset into batches
for batch in batches do
    if user_id in expert_set then
        finaldataframe.extend(batch[id=user_id])
    else
        tokenized_tweet = batch['biographies_clean']
        model = train_word2vec(tokenized_tweet)
        final_words = []
        for word in expert_words do
            if word in model then
                final_words.append(word)
            end if
        end for
        most_similar = []
        for word in final_words do
            most_similar.extend(find_5_mosts_similar(model, word))
        end for
        expert_words.extend(most_similar)
        new_experts = find_users(batch['biographies_clean'], expert_words)
        finaldataframe.extend(batch[user_id in new_experts])
        expert_set.extend(new_experts)
    end if
end for
```

it is a measure of how good is the content a user publishes on a social network. In the specific case of Twitter, interaction is usually measured in terms of RTs (Retweets) and FAVs (Favourites). A RT corresponds to a share, i.e. another user finds your content useful and shares it with their community. On the other hand, a FAV, corresponds to a 'like' on Facebook or Instagram, i.e. a way for users to indicate that they like a particular tweet. At this point, we would like to make a distinction between users who have many followers and those who have few followers. A person with many followers, consequently will also have more interaction than one with few followers, but this does not imply that their content is better. Therefore, we will define engagement as the arithmetic mean of the interaction variables: number of retweets and number of favourites, normalised by the number of followers of the user.

Mathematically for each user $u \in U$, their engagement, denoted as ϵ, is calculated with the following formula:

$$\epsilon(u) = \frac{\frac{nFavsInTopic}{nFollowers} + \frac{nRtsInTopic}{nFollowers}}{2} \tag{I.5}$$

The last filter is based on the **credibility** of the user. The user's credibility on a social network

is intimately related to his or her popularity. That is, an account becomes popular because many other accounts believe it and therefore follow it and share its content. In other words, we can model credibility for our filter, based on an arithmetical mean of the Twitter values that are related to popularity. These values are: the number of followers, the number of public lists in which the user appears, the number of retweets and the number of favourites. In the literature, other works closely related to the NOFACE framework use a standardized linear calculation of variables such as the number of followers, favourites, retweets and mentions. We have preferred to give importance to the lists, as opposed to the mentions, because the mentions are not necessary a good indicator as they can be mentions of anger or reproach, while the lists, have demonstrated in solutions like Cognos [GSB+12] offering good results. According to this, mathematically for each user $u \in U$, their credibility, denoted as ζ, is calculated with the next formula:

$$\zeta(u) = \frac{nFollowers + nLists + nRetweets + nFavs}{4} \tag{I.6}$$

To pass this last filter, the value must be above the mean of all ζ values. After applying this filter, the system will capture those user accounts related to the topic under study, whose content is usually interesting and who also have a wide popularity and credibility in social networks.

The final framework and expertise filtering algorithm were published in (Chapter II, Section 3.1):

> Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2022). NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise. *Expert Systems with Applications*, 208, 118063.

5.3.2 NOFACE validation in real-problems

To validate the proposed framework and address objective 3 (**O3**), we applied our NOFACE framework to a real-world text clustering problem involving a large COVID-19 related dataset. The dataset, a subset of the collection proposed in [Lam20], consists of 2,626,275 tweets. Within this dataset, various challenges are present, including fake opinions, noise, and irrelevant information. Our objective is to demonstrate that by applying NOFACE, we can improve subsequent data mining techniques, specifically text clustering using the K-Means algorithm.

The K-Means clustering algorithm is widely used for grouping data based on different distance metrics [LVV03]. It involves calculating distances between data points and a predetermined number of centroids, which are determined by a parameter. K-Means is known for its simplicity and efficiency [KM13]. For evaluating the performance of our approach, we employed well-established clustering evaluation techniques, namely the Silhouette coefficient and Davies-Bouldin score.

The Silhouette coefficient has been widely used in clustering problems because it determines the quality of separation and cohesion of the obtained clusters. One of the main usages of this metric is to obtain the optimal number of clusters for which a clustering algorithm shows a better performance [Rou87].

The coefficient is computed using the mean intra-cluster similarity $a(i)$ and the mean nearest-cluster similarity $b(i)$ for each sample (Eq. I.7). The overall Silhouette coefficient is the mean Silhouette coefficient of all samples. The values are in the range [-1,1], being the best results those values that are close to 1.

$$S(i) = \frac{a(i) - b(i)}{max\{a(i), b(i)\}} \tag{I.7}$$

Like the Silhouette coefficient, the Davies-Bouldin score allows the evaluation of the results of the clustering algorithms [DB79]. The Davies-Bouldin score is the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances (Eq. I.9). This measure gives a better score to clusters that are farther apart and less dispersed will have a better score. The minimum possible value is zero, with lower values indicating better clustering results, unlike the Silhouette coefficient.

$$r_k = \frac{1}{|A_k|} \sum_{x_i \in A_k} d(x_i, C_k) \tag{I.8}$$

$$DB = \frac{1}{N} \sum_{i}^{N} \sum_{j}^{N} \max_{i \neq j} \frac{r_i + r_j}{d(C_i, C_j)} \tag{I.9}$$

Where r_i and r_j are represented in Equation I.8 (intra-cluster distance), and $d(C_i, C_j)$ is the distance between the centroids C_i and C_j.

During our experimentation, we first applied our trustworthiness filter to the entire dataset, focusing on words related to medicine to ensure the selection of trustworthy content. Subsequently, we randomly selected an equal number of examples from both the filtered and non-filtered datasets. Over these datasets, we performed the clustering process and evaluated the results using different evaluation metrics. The Table I.6 presents the mean results obtained from 10 different executions.

Table I.6: Average result of the runs on the dataset filtered with our proposal and without filtering.

Metric	Filtered	Unfiltered
Silhouette Score	**0,0229**	0,00606
Davies Bouldin	**4,74957**	5,65733

For a more detailed analysis and to support the obtained results, we have conducted a statistical analysis to determine whether there are any significant differences among the obtained values for the two approaches (filtered and unfiltered). Figures 9 and 10, depict the boxplots for the Silhouette coefficient and Davies-Bouldin score respectively, regarding the two approaches. At first sight, we can see that the best results are yielded for both metrics when we use our approach.

We can see how the distribution of results, in both cases, is better when using our filtering method. The box is always wider in the case of filtering because the same accounts and content are not always selected. There is a certain random component to word embedding filtering. This causes the result to fluctuate more in the filtering case than in the non-filtering case where it always runs on similar terms. To justify that the improvement is not due to randomness, we have performed numerous runs and statistical tests. For the statistical analysis, we have used the Wilcoxon's test [Wil45] as there are only two groups (**Filtered** and **Unfiltered**). Considering the $p - values$ shown in Table I.7, we can conclude that there are significant differences between both approaches, where the approach that applies the filter offers the best results.

The contribution related to this real use-case was published in (Chapter II, Section 3.3):

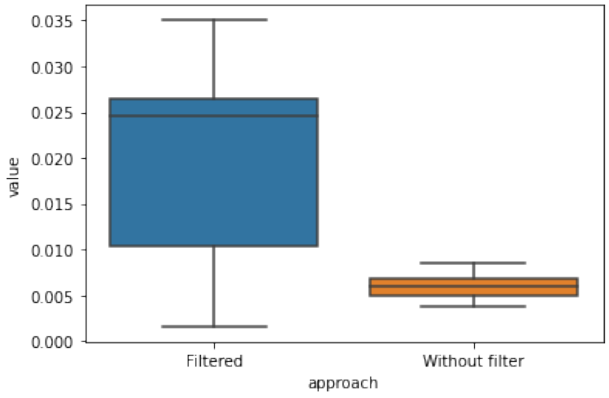

Figure 9: Boxplot of the Silhouette coefficient taking into account both approaches

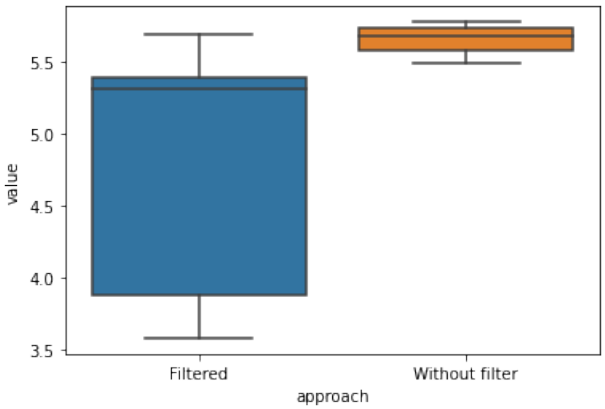

Figure 10: Boxplot of the Davies-Bouldin score taking into account both approaches

Table I.7: $p - values$ for the statistical analysis using both approaches (filtered an unfiltered) and both metrics (Silhouette and Davies-Bouldin).

Measure	**p-value**
Silhouette coefficient	0.02182
Davies-Bouldin score	0.00691

Diaz-Garcia, J. A., Fernandez-Basso, C., Gutiérrez-Batista, K., Ruiz, M. D., & Martin-Bautista, M. J. (2022, July). Improving Text Clustering Using a New Technique for Selecting Trustworthy Content in Social Networks. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 275-287). Springer International Publishing.

Finally to contrast the power of the NOFACE model between different topics, in [DGRMB22] we conducted a exhaustive research over two different use-cases one relating to COVID-19 and the other to the US elections in November 2020. The objective of our use case was (among others) to apply data mining to obtain valuable information about COVID-19 or elections. It is about obtaining, for example, clusters regarding virus containment measures, in the case of COVID-19, or clusters of tweets from independent, non-party biased sources of information in the case of Elections. To graphically compare the obtained results we have represented them by means of a t-distributed stochastic neighbour embedding (TSNE) graph [MH08]. In Figure 11 and Figure 13 the results of applying the clustering algorithm without filtering are shown and in Figure 12 and Figure 14, the results in the case of filtering using NOFACE.

One of the first things we can observe in the TSNE graph is that in the case of the NOFACE results, we have more dispersion between the clusters. This is a clear symptom that good accounts have been selected in which the features are very differentiated. In the case of not applying NOFACE, we have more overlap between clusters, and the silhouette coefficient is of a worse degree. Following with the analysis of the TSNE graph for the COVID-19 use case, in Figure 11, the blue cluster is very dispersed over the whole area of the graph, while in Figure 12 we can see that the majority cluster (in this case the purple one) is quite well defined, as also are the green, light green, red, magenta, dark blue, light blue, yellow, pink and orange ones. In this way, we can see how the application of NOFACE, has greatly improved the execution of the clustering algorithm, because in Figure 11 it is

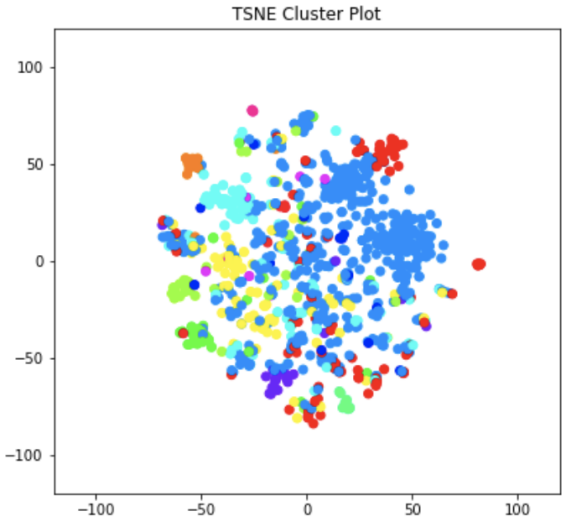

Figure 11: COVID-19 use case: TSNE plot without NOFACE

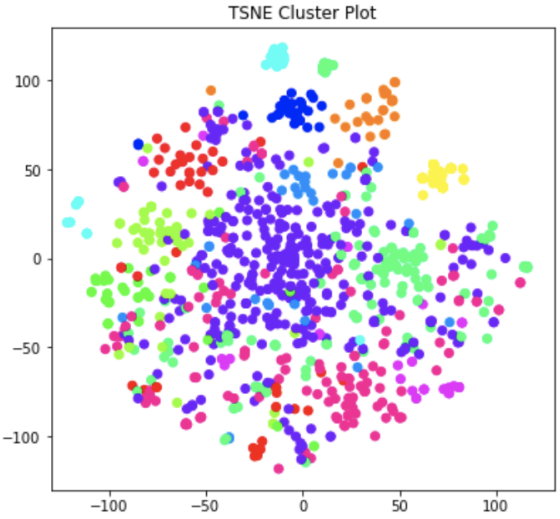

Figure 12: COVID-19 use case: TSNE plot with NOFACE

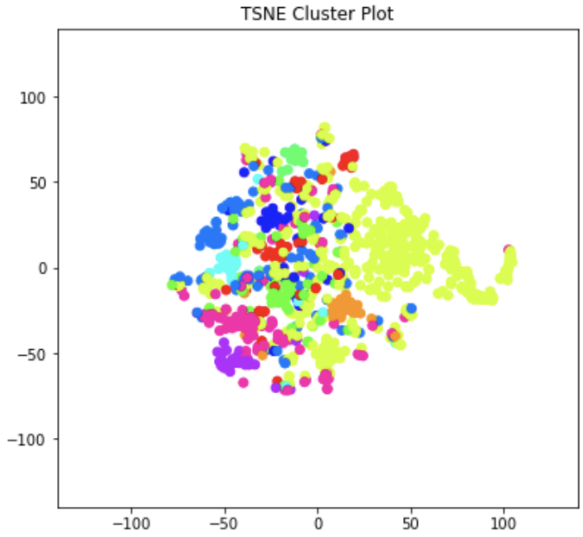

Figure 13: Elections use case: TSNE plot without NOFACE

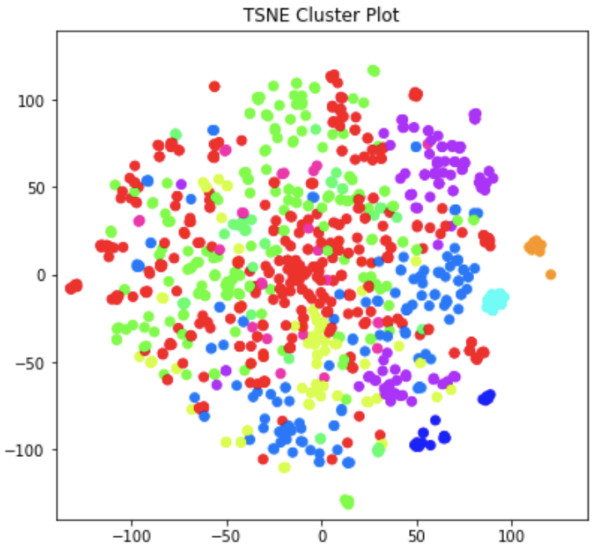

Figure 14: Elections use case: TSNE plot with NOFACE

complicated to identify more than 6 clusters (yellow, red, light blue, orange, red and magenta). This visual analysis is also corroborated by the calculation of the silhouette coefficient. Specifically, in the case of COVID-19, the silhouette coefficient is 0.0095 in the case of using the NOFACE framework and 0.0069 in the case of not using it.

5.4 A flexible big data system for credibility-based filtering of social media information according to expertise

In the final stage of the thesis, we tackled one of the primary challenges posed by data from social media environments: the massive volume and rapid generation of data. This objective aligns with

O2, where we sought to devise innovative techniques to handle such vast amounts of unstructured information effectively. To address this issue, we proposed an extension of our expertise filter, which is an integral part of NOFACE, utilizing the big data Spark framework [ZXW+16].

Spark, a powerful distributed data processing engine, provided the necessary tools to efficiently process and analyze large-scale datasets. By employing Spark's primitives and functions, we could effectively distribute the credibility filter's computation across multiple nodes, harnessing the power of parallel processing. The design of the distributed algorithm for the credibility filter involved leveraging essential Spark functions:

- *Map*: Applies a transformation function to each element of Resilient Distributed Datasets (RDD) and returns a transformed RDD.
- *FlatMap*: Similar to Map, but each input item can be mapped to 0 or more output items.
- *Reduce*: Aggregates the elements of the dataset using an aggregation function.

The algorithm has two main steps (see Algorithm 2). The first one consists of loading the dataset and filtering the experts from each dataset using the FlatMap function. The second step consists of aggregating all these data using a reduce function and grouping them by each found expert.

Algorithm 2 Main Spark procedure for expertise filter algorithm

Input: *Data:* RDD transactions: $\{t_1, \ldots, t_n\}$
Input: *similarity_threshold:* Similarity between words
Output: *Global_expert_set*: Expert discover in each $FlatMap$
Distributive computing in q chunks of transactions: $\{S_1, \ldots, S_q\}$
    $\{< Expert\_1 > \ldots < Expert\_m >\} \leftarrow S_i$.**FlatMap**(FindExpert())
    $FinalDataframe \leftarrow$ **ReduceByKey** $(getInfo())$
End distributive computation
return $FinalDataframe$

Algorithm 4 SimilarFinalWords

Input: *expert_words:* Words from each expert
Input: *similarity_threshold:* $minSim \in (0, 1]$
Output: $KeyValuePair$: Topic and Expert
while *wordinexpert_words* **do**
    if *wordinmodel* **then**
        $final\_words.append(word)$
    end if
end while
# Data frame creation with words most similar to each expert word according to the similarity threshold
while $wordinfinal\_words$ **do**
    if $word.similarity > minSim$ **then**
        $most\_similar.append(find\_most\_similar(model, word))$
    end if
end while
return *most_similar*

Algorithm 3 FindExpert

1: **Input:** *datainput:* Tweet data and topics
2: **Input:** *similarity_threshold:* $minSim \in (0, 1]$
3: **Output:** *KeyValuePair:* Topic and Expert
4: **while** $ExpertList \neq null$ **do**
5: $Expert \leftarrow ExpertList.pop()$
6: **if** $ExpertID \in Global\_expert\_set$ **then**
7: # For each expert, we add all their tweets to the final data frame
8: $Expert_{final} \leftarrow < Expert_{id}, tweets >$
9: $output.append(Expert_{final})$
10: **else**
11: # Finding new experts by processing the rest of the content
12: $tokenized\_tweet = data[`clean\_bios']$
13: $model = train(tokenized\_tweet)$
14: # Adding the most similar words in the model
15: $most\_similar = SimilarFinalWords()$
16: $expert\_words.extend(most\_similar)$
17: # Finding users having any of the words in their biographies
18: $output.append(< Expert_{id}, [expert\_words, clean\_bios, tweets] >)$
19: **end if**
20: **end while**
21: **return** *Output*

In the first stage of the algorithm, it needs the users and tweets with the information processed by the *FindExpert()* function. For this purpose, a *FlatMap* function is used (see line 5 of the Algorithm 2 and the first column of Figure 15).

The *FindExpert* function is described in Algorithm 3, which filters the *expert list* and returns the important information for each of them. The function starts by training a word embedding model on the biographies of the data partition found in the first iteration. To do this, the five most similar words to those that are passed as parameters (e.g. democratic, republican) are found. Algorithm 3 is responsible for aggregating similar words between the experts described in Algorithm 4. The algorithm will use these 12 words (5 most similar to democratic, five most similar to republican, in addition to democratic and republican) to find users whose biographies contain any of these terms and start creating the list of experts that will be stored in the distributed variable *Global_expert_set* (line 3 of Algorithm 2). In the next iteration, these 12 words will be used to search for their five similar ones, and so on.

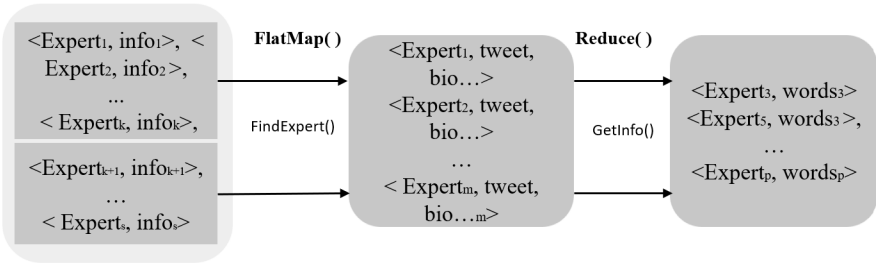

Figure 15: Workflow of big data process in Spark

The FindExpert function will return a list of peers containing the expert and its expert words and some other information. To aggregate and obtain all the information generated in a distributed manner, a *Reduce* function (see line 6 of Algorithm 2 and Figure 15) is used. As a result, a set of words associated with each expert is obtained.

The incorporation of Spark into our expertise filter allowed us to significantly reduce processing time and enhance scalability, enabling the system to handle massive datasets with ease. This extension was a critical step in ensuring that our techniques remained effective and practical in real-world scenarios, where data volumes continue to grow rapidly. We applied our algorithm to datasets coming from Twitter, comprising 5 million tweets and featuring 3 different use cases and input words cases, to validate the versatility and performance of the model. In Figure 16, the differences in terms of execution times are illustrated.

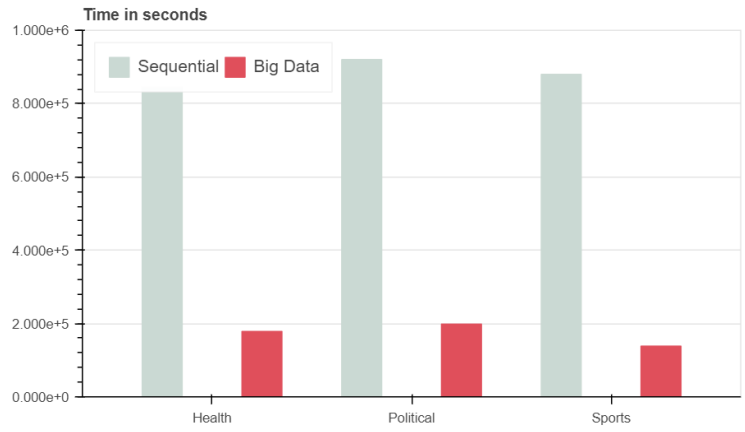

Figure 16: Execution time comparison for 5M tweets and each use case.

The contribution related to this algorithm extension was published in (Chapter II, Section 4):

Diaz-Garcia, J. A., Gutiérrez-Batista, K., Fernandez-Basso, C., Ruiz, M. D., & Martin-Bautista, M. J. (2023, September). A flexible big data system for credibility-based filtering of social media information according to expertise. *International Journal of Computational Intelligence Systems*.

6 Concluding remarks

Throughout the course of this thesis, we have presented distinct proposals, from which we can confidently assert the fulfillment of our hypothesis. Our contributions represent a significant stride in the realm of opinion mining, effectively achieving the primary objectives (**O1, O2** and **O3**) and their corresponding sub-objectives (**O2.1** and **O2.2**). These contributions have propelled the field of text and opinion mining forward and introduced groundbreaking techniques to augment the understanding and analysis of unstructured content within social media environments. Our contributions, closely aligned with our objectives, encompass:

- The review of current applications of association rules to the field of textual content in social networks that can set the groundwork for future work and applications (**O1**).
- The identification of current challenges and future problems that need to be addressed in the coming years with association rules in user-generated content (**O1** and faced in **O2.1** and **O2.2**).
- The theoretical definition of query and non-query based systems, as well as the value of the latter for big data problems (**O2**,**O3**).
- A new approach for sentiment analysis using generalized association rules, capable of summarizing a very huge set of tweets in a set of rules based on the 8 emotions (**O2.1**).
- The design of a non-query based system that is capable of working without topic filtered tweets. This differs from literature, where all the works are query based and tweets are filtered according to a specific topic depending on the problem under study (**O2.1**).
- The development of a system harnessing the power of GAR has allow us to extract sentiments, identify aspects, and entities discussed and assessed in social media, all without the necessity of prior training. This innovative approach has unlocked the potential for unsupervised sentiment analysis and opinion mining, thereby delivering valuable insights from extensive social media datasets without the constraints of predefined sentiment labels or the reliance on training data, squarely addressing our objective (**O2.1**).
- A detalied study and comparison of the performance of different word embedding algorithms and their internal representations for the task of retrieving similar search terms in social networks (**O2.2**).
- A new algorithm for the selection of credible users in the social network Twitter based on the popularity and expertise of the user. To do this we focus on the user's biographies and process it with word embedding. As far as we know, this is the first work that applies word embeddings techniques to the biographies of the user on Twitter, using this to discern the user's expertise on a certain topic (**O2.2**).
- We introduced the NOFACE system, which incorporates word embeddings and social features to evaluate opinion holders credibility and detect potential spreaders of misinformation and non-credible content. Utilizing word embeddings, our system effectively distinguishes between trustworthy and untrustworthy users, thereby elevating quality of information derived from social media environments. This enhancement in credibility assessment significantly bolsters the subsequent data mining procedure, guaranteeing that only credible and reliable opinion holders are included in further analysis (**O2.2**).

- The extension of the NOFACE system under the big data paradigm using Spark, highlighting its great usefulness in problems involving large data sets from social networks (**O2**).
- A comprehensive analysis of patterns related to fake and real news during the 2016 US presidential election campaign (**O3**).
- The validation of the NOFACE framework involves testing it with three different techniques (LDA, clustering, and association rules) and two distinct use cases (**O3**).
- A comprehensive analysis of the power of NOFACE as a data filtering technique to improve the textual clustering results on user-generated content in social networks (**O3**).

Although the thesis has derived valuable insights and made relevant contributions to the state of the art, it is important to acknowledge that all the proposed techniques and solutions have their own set of limitations and potential areas for improvement, which should be thoroughly discussed and evaluated to explore possible research directions. Additionally, the thesis has arrived at important conclusions and findings that warrant in-depth analysis and discussion. In the following section, we will delve into these final conclusions, focusing on the three main objectives of our thesis.

6.1 Discussion and conclusions

To achieve our objetive (**O1**), in our survey [DGRMB23], we showcased the significant potential of unsupervised techniques, particularly association rules, in addressing user-generated text mining problems within social networks.

As with any technique, association rules have their strengths and weaknesses in the context of analyzing social media texts.One of the major strengths of association rules in the social network text domain is their ability to provide interpretable and meaningful information. The output of the algorithm is easily understandable and can be extrapolated across various application domains, making it accessible even to non-technical users. Additionally, association rules are valued for their capacity to summarize information and reveal implicit relationships between different textual components that might otherwise remain hidden. By identifying these relationships, it becomes easier to comprehend complex patterns and gain a deeper understanding of the problem domain, such as identifying associations between diseases and medicines or cities and tourist attractions. This, in turn, allows researchers to focus on relevant information and avoid irrelevant terms, streamlining the analysis process. Furthermore, the unsupervised nature of association rules makes them ideal candidates for the initial stages of data mining, where they can be used as building blocks for hybrid models that incorporate both supervised and unsupervised components.

However, association rules also exhibit some weaknesses when applied to social network texts. The textual transaction matrix can become very sparse, making it challenging to handle and process in memory. Additionally, association rules are susceptible to the presence of lies and colloquial expressions often found in social media, which can impact the accuracy of the results.

Regarding the credibility of opinions and opinion holders, the problem has been approached from different perspectives within Artificial Intelligence. Two main aspects in this area are content-level credibility and user-level credibility. Both of these aspects heavily rely on social and content features and are particularly sensitive to the issue of deception and dishonesty prevalent in social networks. For instance, individuals might falsely claim to possess certain credentials or expertise in their biographies, leading to credibility issues in the system. Designing an automatic system capable of discerning between genuine users and deceptive ones remains a significant challenge in this domain.

Despite these challenges, our findings in the review of the state of the art reflects potential and relevance of association rules in uncovering valuable insights from diverse and voluminous social media data. Similarly, the pursuit of solutions for credibility assessment reflects the importance of ensuring the quality and reliability of information obtained from social networks.

The second objective (**O2**) holds significant importance in our thesis, as it has led to the development of two innovative techniques for mining opinions in an unsupervised and interpretable manner, while also being capable of handling vast amounts of data. To the best of our knowledge, our work represents the first attempt to perform pattern analysis in social media without the use of topic filtering. This uniqueness makes direct comparisons with similar existing systems in the literature challenging, as many of them rely on filtering by hashtags, clusters, or topics to achieve better performance in terms of execution time, memory usage or accuracy. Our system, however, takes a step further by offering a processing flow capable of handling raw and unfiltered data from social media in an unsupervised manner.

Regardless of the specific algorithm used, our system has demonstrated excellent results with unsupervised data mining techniques. The generated patterns are highly descriptive and can be effectively utilized, for instance, by the press to gather information about the tweets published within a specific time frame on particular topics of interest. Thanks to the power of the non-query based system, multiple topics can be analyzed in conjunction without the need of obtaining or loading new data. Essentially, our proposal paves the way for the development of a large-scale social media listening system, particularly for Twitter.

The effectiveness of association rules in obtaining sentiment patterns is noteworthy, as they provide intuitive and close-to-natural-language interpretations in a straightforward manner, even without prior information about the specific problem domain. However, it is essential to acknowledge some limitations of our solution. While the accuracy of the proposed approach is acceptable, it is important to recognize that deep learning techniques can outperform association rule-based proposals when sufficient data is available for training models. Therefore, in situations where extensive training data is accessible, deep learning approaches may outperforms our solution.

To address objective **O2.2**, we developed NOFACE, a framework designed to tackle the challenges posed by misinformation and non-credible opinion holders in social media. Our approach leverages the potential of user biographies in platforms like Twitter, where it is common for individuals to mention their professions. As far as we know, our work is the first to explore and utilize this information in conjunction with word embeddings.

One of the primary challenges faced by the NOFACE framework, as well as other similar systems, is dealing with the prevalence of lies and noise in social media. To mitigate this issue, NOFACE, along with other proposals in the literature [CMP11], employs multiple layers of analysis (such as engagement and credibility assessment). This approach aims to distinguish real Twitter followers from dubious accounts based on their biographical information and affiliations. The premise is that authentic followers are more likely to interact with content from credible and trustworthy users, helping to identify and prioritize genuine influencers.

However, one of the significant challenges in systems like NOFACE is the potential to detect fake influencers as relevant. These fake influencers often manage accounts with impressive interaction statistics and content generation. Brands may be misled into believing that collaborating with these accounts will yield substantial returns, but in reality, it would be a futile investment as these accounts typically interact with bots and non-genuine users, resulting in a lack of real engagement. Detecting such accounts necessitates network analysis, and according to [TAN19], these accounts tend to exhibit egocentric behaviors.

Objective **O3** aims to apply and validate the proposed techniques through real-world use cases. To validate the proposed system to address objective **O2** [DMC16], we applied it to a real-life contrasting event. We chose two well-known US politicians, Donald Trump and Hillary Clinton, to analyze the patterns and rules obtained by the system in relation to the events that occurred in 2016. While there were many relevant people in the social network Twitter discovered by the system, we focused on these two characters to facilitate the contrast with real-life events. However, the system can be applied to other topics or characters present in the dataset. The results obtained from this analysis were found to be satisfactory. Our system successfully identified patterns, association rules, and sentiments that were related to events during the election campaign. These patterns provided a descriptive representation of policies adopted or to be adopted, disputed voting places, and confrontations between candidates.

NOFACE filtering was tested on three different real use-cases: clustering, association rules, and LDA. It was observed that the framework improved clustering results in terms of silhouette and cohesion of the clusters. Additionally, the topics obtained through LDA were of higher quality in terms of coherence. The literature [JK21] emphasizes the significance of pre-processing techniques in enhancing the performance of algorithms like topical detection algorithms (LDA). Efficient pre-processing was shown to improve the execution times of the entire data mining pipeline. The filtering offered by NOFACE proved to be particularly beneficial for algorithms dealing with large datasets or those with inefficient execution times, as demonstrated with the Apriori algorithm [AMA14].

Another important concluding observation from our research is related to the dynamic nature of social networks. The landscape of social media platforms is constantly evolving, with new players entering the scene and established platforms undergoing significant changes. Recently, META, formerly known as Facebook, introduced Threads [Fac23], a new microblogging social network, while Twitter has also embarked on a transformational journey by rebranding as X [CBS23]. Additionally, platforms like Reddit [MLD19] and Mastodon [ZLG20] are experiencing a surge in user growth. While our research primarily focused on Twitter, given its leadership position in the current competition, it's essential to acknowledge that our developed techniques can be readily extended to other microblogging systems with minimal modifications or even without any changes at all.

This adaptability and scalability of our techniques allow us to explore and extract valuable insights from diverse social media platforms, enabling us to keep pace with the ever-changing landscape of social networks and continue making meaningful contributions to the field of opinion mining and sentiment analysis. As social media ecosystems continue to evolve, our research remains well-positioned to address new challenges and opportunities that emerge in this dynamic space.

7 Future work

Based on the extensive work and publications conducted during the course of this thesis, several noteworthy areas for future research and projects have emerged. These potential avenues for future work include:

- **Enhancing traditional techniques for social media analysis**: The thesis has successfully demonstrated the power of traditional techniques in extracting valuable insights from social media data. As part of future work, there is a potential to further explore and enhance these traditional techniques to democratize the use of artificial intelligence. The aim is to develop more lightweight and resource-efficient approaches that can be readily applied without the need for big computational clusters or extensive resources. This would enable wider accessibility and adoption of AI methods for analyzing social media data, empowering researchers and practitioners with valuable insights while minimizing computational requirements.
- **Real-time environment:** Extending the proposed techniques to real-time and cloud environments to enable the analysis and processing of social media data in real-time. This would allow for timely insights and responses to emerging trends and opinions.
- **Ethical considerations:** Addressing the ethical implications of social media opinion mining, including privacy concerns, bias mitigation, and responsible data usage. Developing systems that prioritize ethical practices ensures the development of accountable and trustworthy systems
- **New multidimensional version:** Creating a multidimensional system that considers additional contextual information, such as demographics, time, and location data. This expanded approach would help mitigate issues such as fake professional claims on social media platforms.
- **Extension to fake news and bot detection:** Utilizing the developed techniques, such as credibility and engagement analysis, to enrich other models and aid in the detection of fake news and bots in social networks. This extension would contribute to the identification and mitigation of misinformation and malicious activities on social media platforms.
- **Moving to more explicable models**: Future work should focus on exploring novel methodologies, algorithms, and visualization techniques that enhance model interpretability without sacrificing performance. It is necessary to investigate techniques such as rule extraction, feature importance analysis, and model-agnostic explanations. Additionally, research efforts should be directed towards developing evaluation metrics, and provide new datasets and frameworks to assess the interpretability of models in social media analysis and misinformation detection.
- **Facing misinformation as part of IR problem**: It is crucial to consider the emerging trend of discarding misinformation as part of the information retrieval process. This area, in which our research and other cutting-edge papers [PMNL21, ZVTAS22] can be located, will improve typical problems facing misinformation, such as the capability of generalization of the systems. Due to that, it is necessary to keep improving and developing systems in this area.
- **Explore and extend the systems to supervised thechniques**: To further enhance the system, it is essential to explore and extend it to leverage state-of-the-art supervised techniques based on large language models, such as BERT.

Chapter II

Publications

1 A survey on the use of association rules mining techniques in textual social media

- Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2023). A survey on the use of association rules mining techniques in textual social media. Artificial Intelligence Review, 56(2), 1175-1200.
  - Journal: Artificial Intelligence Review
  - Status: Published.
  - Impact Factor (JCR 2021): 9.588.
  - Category: Computer Science, Artificial Intelligence. Order 17/145 Q1.

Noname manuscript No.
(will be inserted by the editor)

A survey on the use of association rules mining techniques in textual social media

Jose A. Diaz-Garcia · M. Dolores Ruiz · Maria J. Martin-Bautista

Accepted: 12/05/2022

Abstract The incursion of social media in our lives has been much accentuated in the last decade. This has led to a multiplication of data mining tools aimed at obtaining knowledge from these data sources. One of the greatest challenges in this area is to be able to obtain this knowledge without the need for training processes, which requires structured information and pre-labelled datasets. This is where unsupervised data mining techniques come in. These techniques can obtain value from these unstructured and unlabelled data, providing very interesting solutions to enhance the decision-making process. In this paper, we first address the problem of social media mining, as well as the need for unsupervised techniques, in particular association rules, for its treatment. We follow with a broad overview of the applications of association rules in the domain of social media mining, specifically, their application to the problems of mining textual entities, such as tweets. We also focus on the strengths and weaknesses of using association rules for solving different tasks in textual social media. Finally, the paper provides a perspective overview of the challenges that association rules must face in the next decade within the field of social media mining.

Keywords social media mining · association rules · text mining · social networks

Jose A. Diaz-Garcia
Department of Computer Science and Artificial Intelligence, University of Granada
E-mail: jagarcia@decsai.ugr.es
ORCID: 0000-0002-9263-1402

M. Dolores Ruiz
Department of Computer Science and Artificial Intelligence, University of Granada
E-mail: mdruiz@decsai.ugr.es
ORCID: 0000-0003-1077-3173

Maria J. Martin-Bautista
Department of Computer Science and Artificial Intelligence, University of Granada
E-mail: mbautis@decsai.ugr.es
ORCID: 0000-0002-6973-477X

1 Introduction

The recent incursion of online social networks in our world has changed the economic and social paradigm of society. Thanks to social networks and social media, we can communicate with our relatives thousands of miles away, buy products, give our opinion on whether the product is good or not, be updated in real time or simply be entertained.

The volume of information circulating on social networks daily has increased considerably and will continue to do so in a massive way. This has awakened the interest of countless small and large companies, research institutes and governments. These institutions have seen, in the massive processing of social media data, the opportunity to obtain competitive advantages or simply to improve the lives of citizens. Given its importance, many works in the field of Data Mining and Artificial Intelligence have emerged that are focused on analysing these sources of social network data. This has led to the emergence of a new area, known as social media mining.

Social media mining, which will be seen in detail in the next section, seeks to extract value from data from online social networks and social media. This is done by means of text mining techniques, machine learning, natural language processing, clustering, pattern mining and deep learning. In this paper, we focus on the application of association rules to the problem of social media mining. Association rules mining is a non-supervised data mining technique that has a growing protagonism nowadays. In the literature we find many surveys [4], [63] that try to address the problem with supervised approaches, usually classification [53]. It is evident from the nature of the social media mining, that unsupervised techniques, such as association rules, have to be taken into account. This is because they provide a very interpretable solution, are able to deal with large amounts of unstructured data and do not need labelled datasets. This latter feature is very useful in online social networks where the volatility of the topics makes it difficult to have labelled data.

Despite this, even recent surveys [52] that address the use of data mining techniques in the field of user-generated content and sentiment analysis, neglect unsupervised techniques based on association rules. Therefore, we find it necessary to elaborate a review that addresses these techniques, which are so relevant nowadays due to their easily interpretable results and robustness in the absence of labelled data. Both problems are very relevant in the field of Big Data in social networks. As far as we know, this is the first survey that addresses social media mining with association rules. The paper also focuses on current challenges that are being addressed by the association rules and that open up promising avenues for future research. Thus, the major contributions of the survey to the state-of-the-art are:

- The review of current applications of association rules to the field of textual content in social networks that can set the groundwork for future work and applications.

– The identification of current challenges and future problems that need to be addressed in the coming years with association rules in user-generated content. These current challenges are a starting point for future studies.

The paper is organized as follows. The Section 2 explains the methodology followed by the literature review. In Section 3, social media mining problem is discussed, as well as the need for unsupervised techniques, such as association rules. In Section 4, we describe in detail different tasks of association rules in social media analysis. In Section 5, we present the field of application of the studied tasks. In Section 6, we look at the future challenges to be addressed by association rules, aiming to expose the main lines for future research. Section 7 offers a retrospective analysis of the survey and shows some statistics about the papers retrieved. Finally, in Section 8 we present the concluding remarks of our work.

2 Methodology

The survey is based on journals and conferences in various spheres, providing a globalised vision of what the academic, business and social world is doing with the help of association rules. The methodology followed for the creation of the survey is very similar to the one proposed in the paper [42], which is based on the methodology Systematic Literature Review [11].

2.1 Research questions

As far as we know, association rules are one of the main techniques used in Data Mining and are very popular in several sectors [76]. In this review, our aim is to determine whether the field of social networks and textual content (microblogs, posts, tags...) is one of these sectors. Therefore, our research questions will be aimed at determining in which social network tasks and application domains we find solutions based on association rules. So, the research questions (RQs) that we aim to cover with the survey are:

– RQ1: What tasks are currently being solved with association rules in user-generated text?
– RQ2: What areas or fields of application have been addressed with association rules in user-generated text?
– RQ3: What are the current trends and future problems to be faced by association rules in social media mining?

2.2 Search strategy

The search criteria employed has been based on the research questions and the main association rule mining algorithms. Concretely, using combinations

of OR logical operators, we searched for articles that included the following terms in the abstract or the title of the paper: *association rules, pattern mining, Apriori, Eclat, FP growth* and *association rule mining*. In order to refine the search to the review domain, the OR combinations of the above words have been joined by an AND operator with the following terms: *Text mining, social media mining, social media, user-generated text* and *social networks*. The databases and search engines queried are those that encompass most JCR[1] ranked journals and publications, which means that the value of the articles is proven with a rigorous peer review. We used the following databases:

- IEEE Xplore
- Google Scholar
- ScienceDirect
- Web of Science

2.3 Study selection

The papers were selected based on the following inclusion criteria:

- Use association rules in social media.
- Scope of application is user-generated text on social networks.
- Studies indexed in at least one of the above databases.
- If the paper has different versions, we consider the most recent one.
- Papers published between the year 2000 and 2020. They correspond to the year in which social networks and blogs started to gain popularity (2000) and the last expired year (2020).

On the other hand, some articles were left out of the review based on the following exclusion criteria:

- Objectives of the paper were not well defined or the application filed was not clearly related to the user-generated text.
- Papers that apply association rules on non-user-generated textual content, e.g. on graphs.
- Theoretical studies on new association rule techniques that only name possible applications.

The papers were analysed manually, initially reading the abstract and checking whether they passed the inclusion criteria. In total, 43 papers were considered to pass the filter. Once they passed this cut-off, an exhaustive reading was carried out focusing on the association rules technique used, the task in which it is applied and the scope of application.

[1] Journal Citation Reports (impact factor).

3 The social media mining problem

Social media mining, according to P. Gundecha [35], involves the process of representing, analysing and extracting meaningful and valuable social media patterns from data. In a latter stage, these patterns can be used in the decision-making processes of small or large companies. Social media mining is therefore a multidisciplinary field and its scope can be divided, according to P. Gundecha [35], into the following areas of application:

- Community analysis: By means of graph theory [86] and clustering [8], communities within a target population are obtained. These can be users with similar interests, likes or preferences.
- Collaborative recommendation systems: The recommendation system is based on the assumption that similar users will have similar likes, so that recommendation systems can be refined to take these factors into account. In the field of online social networks, collaborative recommendation systems [85] are present on multiple platforms. For example, Facebook or Instagram systems recommend whom to follow, based on users' friends or likes.
- Influence studies: These are based on obtaining the influence of brands, topics or people in certain sectors [60].
- Dissemination of information: In today's information-saturated world, knowing the best way to disseminate information to reach more people is a critical factor. Recently, this sector has come under scrutiny due to the rapid dissemination of fake news or misinformation on topics of social interest such as COVID-19 [39].
- Privacy, security [20] and truthfulness: This area focuses on the automatic verification of false accounts, identification of sources of spam as well as identifying the truthfulness of information [15] or identifying privacy violation issues.
- Opinion mining: A process by which we try to obtain relevant information from texts published on the web [87]. This information can be, for example, polarity, orientation or relationships between text entities.

If we skip those problems that focus on graph theory and distance analysis, almost all of the problems addressed by social media mining, are related to text mining and Natural Language Processing. Text mining and Natural Language Processing aim to extract non-implicit knowledge from unstructured textual entities. Therefore, we could argue that social media mining is a specification of text mining, when text mining is applied to texts from websites or social networks. At this point, the technique diverges because it has to deal with problems inherent to user-generated content, such as colloquial expressions, emoticons, jokes or sarcasm.

Many of the solutions provided in the literature about social network mining involve supervised learning [51], [64],[14]. But, what happens when the problem does not have any databases on which to train? This is the point where unsupervised methods, such as association rules, become relevant. Association

rules are able to obtain in an unsupervised way co-occurrence relationships in large databases, such as those from online social networks. This means that association rules must be taken into account in social media mining problems because they provide fast, efficient and highly interpretable solutions, which are very valuable features in the decision-making process. Association rules, were created for transactional databases, therefore, to apply them in text problems, such as user-generated content, the first step would be to create the text transactions. We will study this in detail in the next section.

3.1 Association rules for mining text from social networks

Association rules belong to the Data Mining field and have been used and studied for a long time. One of the first references to them dates back to 1993 [6]. They are used to obtain relevant information from large transactional databases. A transactional database could be, for example, a shopping basket database, where the items would be the products, or a text database, as in our case, where the items are the words, or more specifically, the entities represented by the words. In a more formal way, let t=$\{A,B,C\}$ be a transaction of three items *(A, B and C)*, and any combination of these items forms an itemset. In this case, all the possible itemsets are: $\{A,B,C\}$, $\{A,B\}$, $\{B,C\}$, $\{A,C\}$, $\{A\}$,$\{A\}$, $\{B\}$ *and* $\{C\}$. According to this, an association rule would be represented in the form $X \rightarrow Y$, where X is an itemset that represents the antecedent, and Y an itemset called consequent where ($X \cap Y = \phi$). As a result, we can conclude that consequent itemsets have a co-occurrence relation with antecedent itemsets. Therefore, association rules can be used as a method for extracting hidden relationships among items or elements within transactional databases, data warehouses or other types of data storage. The classical way of measuring the goodness of association rules regarding a given problem is using three measures: support, confidence and lift, which are defined as follows:

- Support of an itemset. It is represented as *supp (X)*, and is the proportion of transactions containing itemset X out of the total number of transactions of the dataset (D). Support is defined by equation (1).

$$supp(X) = \frac{|t \in D : X \subseteq t|}{|D|} \tag{1}$$

- Support of an association rule. It is represented as *supp*($X \rightarrow Y$) and is the total amount of transactions containing both itemsets X and Y, as defined in the following equation:

$$supp(X \rightarrow Y) = supp(X \cup Y) \tag{2}$$

- Confidence of an association rule. It is represented as *conf* ($X \rightarrow Y$) and represents the proportion of transactions containing itemset X which also contains Y. The equation is:

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{3}$$

– Lift. It is a useful measure to assess independence between itemsets of an association rule. The measure *lift (X→ Y)* represents the degree to which X is frequent when Y is present or vice versa. Lift is a very interesting measure as it relates the antecedent and the consequent through the concept of independence. A value of 1 indicates that the appearance of the consequent and the antecedent in the same rule is independent, therefore, the rule has no effect. On the other hand, lift values greater than 1 indicate a dependence between antecedent and consequent that will make the rule perfect for predicting the consequent in future datasets. Negative values indicate that the presence of one item has a negative effect on the presence of another. Lift is defined mathematically in the following way:

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)} \tag{4}$$

Since association rules demonstrated their great potential to obtain hidden co-occurrence relationships within transactional databases, they have been increasingly applied in different fields. Among other fields, one in which association rules have attracted a lot of interest is Text Mining. One of the first papers which addresses the problem of text mining with association rules, is the paper presented by Martin-Bautista et al. [59]. In this work, textual transactions are defined, on which fuzzy association rules are applied. Textual transactions are necessary in order to be able to apply association rules on text, opening the possibility of applying association rules to text mining problems. In this field, text entities (opinions, tweets,...) are handled as transactions in which each of the words is an item. In this way, we can obtain relationships and metrics about co-occurrences in large text databases. Technically, we could define a text transaction as:

Definition 1 Text transaction: Let W be a set of items (words in our context). A text transaction is defined as a subset of words, i.e. a word will be present or not in a transaction.

For example, in a Twitter database, in which each tweet is a transaction, it will be composed of each of the terms that appear in that tweet. So the items will be the words. The structure will be stored in a term matrix in which the terms that appear will be labelled with 1 and those that are not present as 0. For example for the transactional database $D = \{t1, t2\}$ being $t1 = (just, like, emails, requested, congress)$ and $t2 = (just, anyone, knows, use, delete, keys)$ the representation of text transactions is shown in Table 1.

We can see how in a real problem this matrix will be very sparse. This is a problem that is often mitigated with different data cleaning processes. For example, in some cases, terms that are similar or represent a closely related entity are exchanged for this entity so that the term matrix is less sparse.

Table 1 Example of a term matrix in a database with two textual transactions.

Transaction\Item	***anyone***	***congress***	***delete***	***emails***	***just***	***keys***	***knows***	***like***	***requested***	***use***
t1	0	1	0	1	1	0	0	1	1	0
t2	1	0	1	0	1	1	1	0	0	1

Transactions are an essential part of the association rule mining extraction process and without them we would not be able to mine frequent patterns. Association rules cannot be applied on raw text, so with this internal representation of text as textual transactions, association rules can be applied to almost any textual data problem. The internal representations of the matrix can also be calculated instead of using the absolute frequency, by employing its TF-IDF[2], TF[3] [71] or even its fuzzy membership value, which gives to this representation a great versatility still to be exploited. Having any of these internal representations for transactions, it is possible to apply any association rules mining approach.

The most widespread approach for mining association rules is based on the downward-closure property of support and consists of two stages. To be considered frequent, the itemset has to exceed the minimum support threshold. In the second stage, association rules are obtained using the confidence or other assessment measure. To obtain the frequent itemsets, the algorithms, based on a minimum support value, will generate all the possible combinations of itemsets and will check if they are frequent or not. In each iteration, all the possible different itemsets that can be formed by combining those of the previous iteration are generated, so the itemsets will grow in size. Within this category we find most of the algorithms for obtaining association rules, such as Apriori, proposed by Agrawal and Srikant [7]. Apriori is based on the premise that if an itemset is frequent, then all its subsets are also frequent. It means when we find one of unfrequent itemsets, all other itemsets containing this one do not need to be analyzed. Thus, we can prune the search tree avoiding checks and increasing efficiency. Although the most widely used algorithm is the Apriori algorithm, it is not the only one. We can find others such as FP-Growth proposed by Han et al. [37] and Eclat [65].

Eclat [65] is an improved version of the Apriori algorithm, which improves the execution times of the algorithm. The main difference with the Apriori algorithm is that the Eclat algorithm performs a vertical search, similar to the depth-first search of a graph, as opposed to the breadth-first search performed by the Apriori algorithm. The basic idea is to compute intersections between items. To do this, a list of items is created, in which the items are related to the transactions in which they appear. With this list, the algorithm can compute the support value of a candidate itemset and avoid generating subsets that will not reach the support threshold. In this way, it can reduce the computation time in obtaining rules, but the management of this list implies a higher memory consumption.

[2] Term Frequency - Inverse Document Frequency

[3] Term Frequency

The FP-Growth algorithm [37] was proposed in 2000, as a solution to the memory problems generated by typical methods such as Apriori, seen above. It is a very efficient algorithm and is widely used in problems and solutions that could be framed under the name of Big Data. FP-Growth creates a compressed model of the original database using a data structure called FP-tree, which is made up of two essential elements:

- Transaction network: Thanks to this network, the entire database can be abbreviated. At each node, an itemset and its support are described and calculated by following the path from the root to the node in question.
- Header table: This is a table of lists of items. That is, for each item, a list is created that links nodes of the network where it appears. Once the tree is constructed, using a recursive approach based on divide and conquer, frequent itemsets are extracted. To do this, first the support of each of the items that appear in the header table is obtained. Second, for each of the items that exceed the minimum support, the following steps are carried out:
  1. The section of the tree where the item appears is extracted by readjusting the support values of the items that appear in that section.
  2. Considering the extracted section, a new FP-tree is created.
  3. The itemsets that exceed the minimum support of this last FP-tree are extracted.

Thus, FP-Growth requires less memory than Apriori. Additionally, the divide and conquer principle makes FP-Growth attractive in Big Data environments. It is worth noting that one of the reasons why FP-Growth is more efficient is that it is not exhaustive, i.e. it does not get all possible rules. Apriori, on the other hand, does [33].

Before finishing this section it is necessary to mention that the internal binary or fuzzy matrix representation is not the only possible way to represent the textual transactions. In some studies a standard representation is used in which the terms and texts present in social networks are used to form a new transactional database. For example, let us imagine a database of tweets. By means of the text mining procedure we can obtain feelings about each of the tweets and build a transactional element called sentiment, which could take the values sentiment=positive, sentiment=negative, sentiment=neutral. Another transactional element could be names, which would take the output produced by a Named Entity Recognition over the tweets. The Named Entity Recognition (NER) process is an automatic process that identifies entities and assings them a category within their grammatical or lexical category [70]. One of the most famous systems, proposed by Stanford University [57], can be used to obtain the names of people, places, etc. mentioned in a given document. With these textual categories located and tagged, we can create transactions. For example, names = (Katie) or names = (Biden,Trump). In this case a concrete example would be $D = \{t1, t2\}$ being $t1 = (sentiment = positive, names = Trump)$ and $t2 = (sentiment = neutral, names = Biden)$. On these two

transactions we could perfectly apply association rules to find which feelings have more co-occurrence with certain names.

4 Tasks

In this section we have grouped the studies that apply association rules according to their data mining task. In Section 5, we will analyse the fields or domains of application of these tasks. A summary of all of them has been added in Table 2.

4.1 Summarization

The large amount of data present on the web has influenced the appearance of systems capable of presenting this information in a summarized form. Association rules have the potential to bring together words and terms by means of co-occurrences and frequent itemsets, therefore, this is one of the fields of social media mining in which association rules proliferate the most. r One of the first papers to apply association rules to summarization is [47]. Within this category, there are articles focused on summarizing information from Twitter. These articles are based on summarizing threads of conversation on a particular topic, so that the reader can get a reliable and complete idea quickly. In this area we also found the proposal in [67] that uses maximal association rules to summarize Obama's most important tweets.

4.2 Topic detection

Topic detection is based on the ability of an intelligent system to detect what is being talked about, or what a specific discussion in a social network is about. That is, they try to tag a set of textual entities according to a topic. In this sense, probabilistic and unsupervised techniques such as LDA[4] [9], [43], stand out, but given the unsupervised nature of the problem we also find some approximations by association rules.

One of the first studies in this line dates from the year 2000 with the work described in [68]. This paper focuses on obtaining new research topics on the text in the WWW[5]. To do so, it uses association rules obtained from the keywords of the publications. A more recent study in this line is the paper by Cagliero et al. [12], where a solution through generalised association rules is offered, which provides a compendium of topics used in the social network Twitter. This paper uses dynamic association rules that are generalized through the context of the posts and the content of the tweets generated by

[4] Latent Dirichlet Alocation
[5] World Wide Web

the user. In [55] Mai et al. use association rules to build a Twitter data analytics application that allows to a hypothetical user to find out what is being discussed about their profession on Twitter. The system is based on the Apriori algorithm and data visualisation, so a non-technical user could use it. To test how well it works, the system was tested on a real case with physician assistants.

In the field of topic detection, association rules stand out when being used to obtain knowledge about a specific topic, for example, about cyber bullying [58], [88]. In this way, the authors use the Apriori algorithm to obtain concrete patterns which helped to identify topics in social networks used for cyber bullying. That is, on a specific topic, association rules are used to obtain subtopics or more detailed information and knowledge. In the same direction as the previous ones, but oriented to tweets from insurance, we find the study proposed by Mosley et al. [61].

As for topic detection, but in reviews on digital banking of a Google Play Store app, we found the paper [17]. In it, Cheng and Sharmayne propose to use LDA and association rules in conjunction. They first obtain topics about the reviews, and create two clusters depending on whether they are positive or negative. Once this is done, they obtain association rules that relate terms from each topic to one of the clusters. With the result, they try to obtain which words in the customer reviews are related to positive or negative concepts of the services of the bank's app.

4.3 Event detection

Event detection is closely linked to the detection of topics, but characterized by having a temporal character, for example, the identification of possible terrorist attacks, floods or earthquakes. In this area of application association rules are playing a major role again due to the possibility of using them without prior training. In this sense, they are very appropriate and robust to be used in data from unfiltered social networks and processing them in real time.

Two papers that are related according to the data set used are the one proposed by Adedoyin et al. [5] and Ferdandez-Basso et al. [32]. Both try to detect events in politics and sports events. The first addresses the problem by matching rules on hashtags, whilst the second is a spark-based distributed solution that improves detection trends in the form of frequent itemsets without obtaining rules.

For events related to possible problems or catastrophes, we find the papers [34], [41], [3], [72]. All of them make use of a crawler on websites and social networks to obtain event detection patterns on transport and traffic in the first case, insurance and natural disaster in the second and third and on flood areas in the Manila underground in the last case. These studies try to exploit the potential of association rules to obtain relevant patterns quickly on collected data by the crawler in real time. However, these systems are unreliable compared to other systems based, for example, on classification. Supervised

systems are most suitable for event detection problems, especially when human lives may be at stake. In the field of health, and particularly in monitoring possible health-related behaviour events on social networks, we find the paper [46]. In it, association rules are used to monitor health in Korea to face the problem of yellow dust.

A paper that is very interesting due to the nature of the used data is the one proposed by Zhang et al. [90]. This paper tries to detect events on video and uses adaptive association rules on video metadata and video tags. The obtained patterns are used to feed an event classification system that also uses a Near-Duplicate Keyframe identifier.

4.4 Sentiment analysis

For the application of association rules for sentiment analysis, there are also some recent hybrid approaches that use association rules in conjunction with other techniques. Hybrid approaches can be defined as proposals that use association rules and other techniques (usually classification algorithms) to improve the results. In these cases, due to the nature of the problem, association rules cannot be used alone, but they can be used to improve the results of later stages of classification. Association rules play a role in finding relationships between terms that are grouped by patterns in the classification stage, improving the results to a significant degree.

In this case we find two papers [89], [23] where a hybrid approach is proposed using association rules to generate co-occurrences of terms related to feelings and thus create a much more powerful system of sentiment analysis. The first paper uses a dataset of crimes and motives for crime, which is compared with methods such as PCA[6], showing that the method using the rules for co-occurrences improves what already exists. The latter paper [23] applies association rules in an earlier stage of sentiment classification, summarizing the debates about Kurds on Twitter.

Closely linked to the above hybrid models but in the health field we find the paper [66]. In this, the authors use association rules mining and natural language processing techniques to mine the social network Twitter for the use of Fentanyl. After a process of sentiment analysis they use association rules to obtain the correlations of its use with other drugs and products of a dangerous nature for health.

To conclude this section, it is necessary to mention a paper that offers a version of sentiment analysis based only on association rules. This is the paper [26], in which the authors generalize the association rules obtained by the majority sentiment, cataloguing the emotions around a politician on Twitter.

[6] Principal Component Analysis

4.5 Forecasting

Due to their social nature, online social networks have been the subject of forecasting studies since their creation. This is because in many social networks and websites the data is public, and people have opinions about certain issues. These opinions and posts can be used for prediction systems in various aspects such as politics or studies of influence on a certain sector. In this latter case these studies try to report which users are more influential in a certain network with the aim, for example, of being used as seeds for spreading marketing or publicity strategies.

In this area of application we find the papers by Erlandsson et al. [30], [29]. In [30] they propose a system based on association rules to identify which users are more influential, pursuing the idea of comparing the system with the state of the art in the field, with these being the Page Rank Centrality and the Degree Centrality. The results are promising, although they highlight a problem of execution time.

On the other hand, in the paper [29] they predict, using association rules, the participation of users in Facebook, something useful that could be used for example to contact certain high participation users for possible promotions. In line with Facebook, the paper by Nancy et al. [62] uses the Apriori algorithm on a data set of universities to study the influence of gender studying a course.

In the Twitter domain the papers [2], [21] use hashtags that are trending topics, that is, those hashtags used by a several number of tweets in a concrete moment, as textual entities that feed the association rule mining algorithm. With this, they obtain patterns from the most popular users of the social network at that certain moment.

Also in Twitter, the paper [25] is oriented to discern patterns of fake news in a Twitter data set about the 2016 American presidential elections. The paper [27] proposes a solution based on big data capable of cataloguing and obtaining knowledge in an unsupervised way. The system is tested with a set of more than 1M tweets obtained from the United States, specifically, tweets of a political nature. In all the studies, solutions are provided that conclude in the potential of the use of association rules to undermine social networks for voting predictions. n the field of fake news and hoaxes forecasting, we also find the paper [80] where Utami et al. create a classifier based on Random Forest to discover hoaxes on Twitter. At this point, the association rules take part through the Apriori algorithm, used to discover associations between words that simplify the learning process. Again, this is a hybrid approach, where the frequent pattern discovery potential of association rules greatly enhances the results of classification-oriented algorithms.

Tourism is one of the areas where forecasting is more relevant. In this sector, we find the paper [19] that uses the TripAdvisor social network to train rating models from 1 to 5 stars. Within each of the models the association rules are used to see which terms are more related to others, to know which words are the ones that cause an opinion to be better or worse.

In the field of smart cities, we find the paper [73], where the authors use Twitter and spatio-temporal pattern mining with an adaptation of the apriori algorithm to determine which roads in a city are likely to be most congested at a certain time.

We conclude this section by discussing crime surveillance. Crime and criminal organisations are very present in social networks. Having systems capable of dealing with this scourge and giving early warnings is very much needed by law enforcement agencies. In the fight against these problems we find the paper [79]. This paper uses a compendium of data mining techniques to trace the similarity between users that may be related to crimes. To this end, the paper uses association rules and other techniques such as clustering.

4.6 Collaborative social systems

Data from social networks as well as user actions can be used to generate collaborative systems that improve the experience of other users. These systems exploit the premise that if something has been useful for a user, it will also be useful for a user with a similar profile. Under this premise, we find different approaches such as collaborative recommendation systems, collaborative expert systems or social tagging systems. Collaborative recommendation systems aim to be able to recommend items (films, songs...) to certain profiles based on their similarity to other profiles. They are based on the recommendations or ratings given by certain users, so that these preferences enrich the recommendation system. As for expert systems, the system simulates human reasoning in order to make decisions. In this case, certain actions or considerations offered by the system may be motivated by similarity to the situation, or requirements. Finally, social tagging systems try to recommend tags for certain items, based on the tags that another profile has left on similar items. In all three cases, there is an underlying functionality, which is the need to obtain common patterns between different items or profiles. That is why these elements are studied together at this point.

4.6.1 Recommendation and expert systems

In the field of collaborative recommendation and expert systems we find quite a few studies that make use of association rules. This is because the nature of association rules for finding co-occurrence relationships is one of the ways in which expert and recommendation systems are developed internally.

Two closely related papers that use association rules to obtain Twitter patterns that are related and can be used in a later expert system are [56] and [22]. The first deals with the problem of choosing a school, for which it obtains patterns of good schools and terms related to them from the social network Twitter. Along the same lines, but to promote cycling, the paper [22] extracts patterns that relate good habits to promote the population to use bicycles.

Also in the field of health, we find the study [36] that seeks to get patterns on Twitter that could be used by other people to give up smoking.

As far as recommendation systems are concerned, we found some important papers that make use of the rules for recommendation systems ranging from schools to movies as well as mobile services or courses [48], [84], [40]. In the paper [48] the authors use the Apriori algorithm to relate actors and metadata from films to other similar films. It is based on the premise that related films will be a good recommendation for the users. The underlying background in the other papers is very similar but applied to suggest mobile services and courses respectively. The paper [75] uses Twitter and Linkedin to create a system capable of relating users preferences to the most acceptable job applications for them. All this is possible by using just text mining techniques and association rules.

In [92] authors apply an extended version of the Apriori algorithm to a health shopping website. The purpose is to recommend products to users based on the relationship of some products with others, as well as the user's behaviour with respect to certain products. Although the results are promising, the authors conclude that the large amount of data present in these databases limits the use of Apriori, so new and more efficient versions should be explored.

Finally, in [69] Rao et al. propose an unsupervised recommendation system based on the Apriori algorithm and Named Entity Recognition. The system is designed to analyse trends on Twitter, and recommend tweets related to the context of each one in order to consolidate an opinion on them. The system first obtains named entities and then extracts frequent patterns among them. These patterns are represented in trees, having in the leaf nodes the tweets related to the context of a given trend.

4.6.2 Social tagging systems

Finally, it is necessary to talk about social tagging at this point. Social tagging is based on the possibility of tagging through the community of users, any online resource, being it text, movies or music. These tags become part of the communities domain, facilitating thus the search for a certain resource. At this point, association rules have a quite obvious role, since they offer the user words that are largely related to the ones they use as labels, favouring and enriching the system [49], [38]. Another study in this line proposed by Feng et al. [31] is very interesting because it exploits the relationship of emotions and colour with social tagging through association rules. In this study, association rules are used over an encoded image, associating pixels of a colour to certain emotions, something that shows that the use of association rules in social media mining, has innumerable and creative applications.

Table 2 Papers using association rules according to their task and field

Paper	*Task*	*Social Network*	*Academics*	*Crime*	*Health*	*Insurance*	*Influence*	*Leisure*	*Natural disasters*	*Politics*	*Sports*	*Transport*	*Trends*
[67]	Summarization	Twitter								X			
[68]	Topic Detection	Web posts	X										
[12]	Topic Detection	Twitter											X
[58]	Topic Detection	Twitter		X									
[88]	Topic Detection	Twitter		X									
[61]	Topic Detection	Twitter				X							
[55]	Topic Detection	Twitter	X		X								
[17]	Topic Detection	Google reviews						X					
[5]	Event Detection	Twitter						X		X	X		X
[32]	Event Detection	Twitter						X		X	X		X
[34]	Event Detection	Twitter										X	
[41]	Event Detection	Twitter				X			X				
[3]	Event Detection	Twitter							X				
[72]	Event Detection	Twitter							X				
[46]	Event Detection	Twitter			X								
[90]	Event Detection	Youtube		X					X	X	X		X
[89]	Sentiment Analysis	Twitter		X									
[23]	Sentiment Analysis	Twitter								X			
[66]	Sentiment Analysis	Twitter			X								
[26]	Sentiment Analysis	Twitter					X			X			
[30]	Forecasting	Twitter					X						
[29]	Forecasting	Facebook					X						
[62]	Forecasting	Facebook	X				X						X
[2]	Forecasting	Twitter					X						
[21]	Forecasting	Twitter					X						X
[25]	Forecasting	Twitter								X			
[27]	Forecasting	Twitter								X			
[80]	Forecasting	Twitter								X			X
[79]	Forecasting	Twitter		X									
[19]	Forecasting	TripAdvisor						X					
[73]	Forecasting	Twitter										X	
[56]	Expert System	Twitter	X										
[22]	Expert System	Twitter			X							X	
[36]	Expert System	Twitter			X								
[48]	Recommendation	Web Posts						X					
[84]	Recommendation	Web Posts						X					
[40]	Recommendation	Web Posts	X										
[75]	Recommendation	LinkedIn	X					X					
[69]	Recommendation	Twitter											X
[92]	Recommendation	Web Page			X			X					
[49]	Social Tagging	Web Tags						X					
[38]	Social Tagging	Web Tags						X					
[31]	Social Tagging	Web Tags						X					
Total		6	5	6	2	5	11	4	9	3	3	8	

5 Fields of application

In this section, we will go into detail on the field of application where association rules have been applied in the social media mining domain. We must take into consideration that we have gathered the fields in large groups, but they are not exclusive or the only ones, simply those with enough scientific weight for being able to categorize them. The fields of application have been distilled from the reading of the papers and the dataset used:

- Academic: In this area, we have included those papers that deal with the academic or professional world. For example, those in which a university recommendation is proposed or association rules are used to obtain information on new research fields.
- Crime: In the area of crime we have included all those actions catalogued as against the law, from theft to cyber bullying, as well as more serious crimes such as homicides.
- Health: Everything related to health, from epidemics prediction to patterns detection, useful for recommendation systems such as for giving up smoking, for example.
- Insurance: Everything related to insurance, which is closely linked to other branches such as transport and natural disasters, so we will have overlapping papers in these areas.
- Influence: In this area we have included influence studies either of users or topics, linked with trends. This can be catalogued as a kind of subsection in forecasting where the central issue is to find who or what is really influential on the web.
- Leisure: This area encompasses leisure time, news, e-commerce, digital banking, and tourism actions that can be taken to spend time on social networks and the internet.
- Natural Disasters: Floods and disasters generally influenced by the weather and about which people post their warnings and alarms on social networks.
- Politics: This area includes news about politicians, datasets on opinions and hashtags related to electoral processes or tweets and posts released by a politician.
- Sports: Related to football matches or any other type of sports competition.
- Transport: This relates to transport, either personal, citizen mobility or professional related to logistics.
- Trends: Trend analysis is one of the most studied areas in data mining applied to social networks. A trend would be that which at a particular time is heavily posted on social networks, it can be related to many topics, even a cluster of them, but must be linked to a moment or several moments in time.

For a better understanding, Table 2 comprises a compilation of the references cited above. Also Table 2 shows the application, the social network used and the field of application where they were applied. If we analyse the table, we can see how many papers are transversal, that is, they exploit more than one

field of application. This shows that association rules are versatile and there are many fields in which they can be used. One of the reasons for this is linked to their potential for interpretability. This factor is very relevant today, as all areas are asking for explainable and interpretable Artificial Intelligence systems, that are easier to understand contrary to other black box systems. This is undoubtedly a great potential in descriptive problems or previous data analysis. Furthermore, association rules cannot be considered as an alternative to other supervised techniques. This is one of the reasons why the revised works are focused on descriptive analyses and when they are predictive, association rules are used in conjunction with other supervised techniques. Finally, we can see how the field of politics and leisure is currently arousing a lot of interest. In this point, association rules stand out considerably because of their potential to correlate terms or users with certain policies or parties, for example.

6 Current challenges and future trends

As we have seen in the previous section, association rules are a very interesting tool to address the problem of social media mining when, due to restrictions or needs of the problem, it is necessary to deal with it in an unsupervised way. On the other hand, there are still some challenges that are already being addressed and that undoubtedly open up a line of future trends to be faced in the coming years in the field of association rules and social media. The most relevant of these challenges are:

- Streaming association rules: Although association rules were created to address the problem of data mining in large transactional databases, the paradigm has mutated and they are now required in streaming environments. Currently, one problem that has been solved is that of discovering frequent itemsets in streaming [13], [45], [83]. Like for association rules, there are some incipient works about this [1] [82]. Many of these articles name the process of obtaining association rules but they are really based on approximations or data structures for obtaining frequent itemsets. The main problem in obtaining association rules is that in traditional databases the algorithms can take several passes adjusting supports according to thresholds. This is unfeasible in streaming because old data has to be discarded and the supports have to be readjusted according to the time windows. The main concern in addressing this solution is to provide an efficient data structure capable of storing and calculating support and fitting measurements correctly over time and the arrival of new data. There is some work already in this line of research [54] but it still needs to be tested and used in problems related to the domain of social networks user-generated text, where to our knowledge, there are still no systems applied to solve this particular problem. Therefore, we consider it to be a current line of research and a current challenge.
- Temporal association rules: This is a similar idea to streaming but can be applied over association rules in a standard transactional database. In it,

we would have time stamps in each transaction. The obtained rules could be compared in time and would be very relevant in problems of influence detection, because something that today is influential can stop being so in a posterior period of time. There are already some proposals that use temporal fuzzy association rules aiming to address this challenge [16].

- Socialized association rules: Generalized association rules are a great tool to improve the performance of association rules in environments with a large dispersion of items, such as social media environments. This technique was introduced by Srikant and Agrawal [77] in 1995. They propose that the rule $\{Strawberries,\ Oranges\} \rightarrow \{Milk\}$ could be replaced by $\{Fruit\} \rightarrow \{Milk\}$. This hierarchical point of view allows a higher level of abstraction that offers the possibility of obtaining even more information from our data. Creating a topology of generalized association rules by social flags could be a great solution to problems such as event or topic detection. In this area there are already some studies. On the one hand, the research [26] tries to extend and generalize association rules by feelings. On the other hand, the proposal in [81] extends the rules using graph theory, something that could have innumerable applications in online marketing. This point also offers the possibility of enrichment by other elements. For example, for a crime detection application in social networks it would be very interesting to enrich and extend the rules or frequent terms by means of police databases.
- Data dispersion: One of the main problems with user-generated text association rules in online social networks is the granularity and dispersion of terms. That is, when the domain under study occurs in Twitter or TripAdvisor, it is very complicated for a term to appear several times in a post. Therefore, the internal binary representation for transactions should be very smooth for dispersed data. Finding an internal representation with weights and capable of capturing the nature of the analysis in the matrix will offer great results. In this context, there is a large research line in the problem of sparse matrices [44], [24], [91]. But an appropriate solution for matrices obtained from social networks is still missing.

7 Discussion

In this paper, we demonstrated the great potential of unsupervised techniques and, more specifically, of association rules to deal with user-generated text mining problems in social networks. Regarding the strengths and weaknesses of association rules as they apply to user-generated content, it should be noted that they are inherent to the technique and are common across the different fields of application. The greatest strength of association rules in the field of social network text comes from their capacity to provide information. The output of the algorithms is easily interpretable and extrapolated across all the application domains seen. This facilitates their use by non-technical users. We have also detected that association rules are used because of their potential to summarise information and relate certain textual components with others

that would be difficult to see explicitly. These relationships allow a better understanding of the problem domain, for example which disease is related to which medicine or which city is related to which tourist attraction. In this way, terms with little relationship can be avoided, favouring the analysis of what is really necessary. Finally, it should be noted that the unsupervised potential of association rules makes them perfect candidates for a first stage of data mining on which to build hybrid models with both supervised and unsupervised components. As for their weaknesses in the field of their application to social network texts, it should be noted that the textual transaction matrix is very sparse, which makes it difficult to load in memory. Also, it is a technique that is very sensitive to lies and colloquial expressions in social networks. It is clear that its potential is greater than its weaknesses and that is why interest in this subject has grown considerably in the last years.

If we look at the publication years (Figure 1) we can see how the publication trend is increasing. This is due to the fact that social networks have a very recent period of assimilation that starts in 2010 and onwards, so in the first years the number of papers is lower. On the other hand, in recent years the number of papers has increased reaching almost 75% of the total number of papers in the last 3 years. This is due to two reasons. Firstly, it is now when the social networks have a very relevant role in our lives. Secondly, association rules are arousing interest in applications related to social networks given their descriptive potential and interpretability. This is undoubtedly a warning that in the coming years the number of publications such as those contained in this survey will increase considerably, especially those in the health field marked by the COVID-19 pandemic. The pandemic is arousing a lot of interest in the field of artificial intelligence and more specifically in the field of social networks. In this field, we can find numerous papers that deal with sentiment analysis on prevention or social distancing measures [18], [74]. Regarding association rules and COVID-19, we can see how there is a current application and research trend that uses the great interpretability of association rules to obtain patterns and relationships between symptoms [78] or variants of COVID-19 and places [10]. Based on user-generated content, it is only a matter of time before new articles can be added to the review. This can be seen from the fact that there are already some pre-prints, for example [28], that attempt to address the scope of Twitter conversations related to COVID-19. On the other hand, the incursion of COVID-19 also increases other problems present in social networks, such as the dissemination of misinformation or fake news. This problem is also being addressed by techniques based on pattern mining. In [50] association rules are used in conjunction with a named entity detection process to discern if there is any pattern in the use of named entities in real news versus fake news. This article shows the growing interest and usefulness of association rules in conjunction with other less interpretable techniques. This is undoubtedly a symptom of the good health of association rules in the digital world.

Another important factor of the review can be outlined if we look at the kind of publication. The number of publications is around 90% in journals

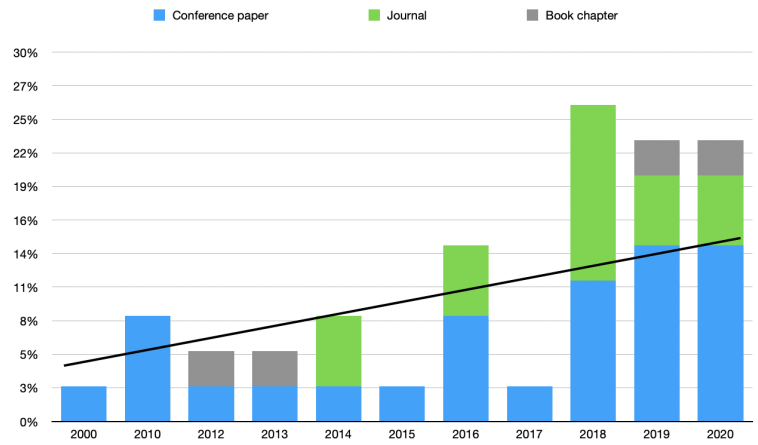

Fig. 1 Papers according to their type and year of publication in social media mining

and conference papers. The remaining 10% are book chapters. The 90% is divided in a 56% of conference papers and a 44% of journal articles. Here we can see how journal publications, which normally have more impact, are almost entirely situated in recent years, which again leads us to think about the health and strength that association rules hold in the current academic community.

Regarding the type of data, as we can see in the Figure 2, Twitter continues to be the most widely used social network for data mining applications and work. This is because it offers APIS to obtain data freely and therefore it is very easy to obtain information about accounts or user-generated content that can feed these systems. Although Twitter stands out the most, it is interesting to mention how association rules are being used in very interesting applications that take into account networks as disparate as Facebook, YouTube, Linkedin or TripAdvisor.

8 Conclusions

User-generated text in social media provides a great opportunity of knowledge as long as we are able to extract it correctly. We have demonstrated how association rules are necessary in social media mining problems. Following the analysis carried out, we can conclude that, at this moment, the power of the association rules lies in the interpretability of the model and the results.

Recently, explainable Artificial Intelligence is becoming increasingly relevant to business processes, as other black box solutions, such as neural networks, offer solutions that are difficult to interpret and maintain. Association rules offer a great interpretability of the results and the model, which makes them very easily extendable to business processes that have to deal with content from social networks. This is very important because these results will

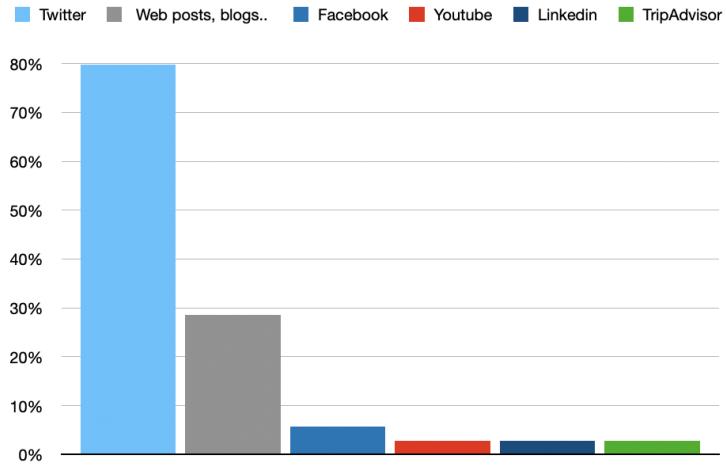

Fig. 2 Papers according to their type of social networks data

probably not have to be interpreted by a data engineer but could be interpreted by someone with less technical knowledge. In addition to interpretability, the potential of association rules in social media mining solutions framed in Big Data lies in the possibility of making use of pattern mining algorithms without having pre-labelled databases, i.e. without prior training. This offers the possibility of having systems capable of responding with sufficient speed to the volatility of social networks. Other models, such as those based on deep learning, need pre-trained networks for being able to respond quickly. Having these pre-trained networks is not trivial, and a change of context or trends in the networks would invalidate it completely. This offers a competitive advantage to association rule-based systems as they can adapt to these changes without the need for training. However, association rules cannot be positioned as an alternative to neural networks or other classification techniques in certain problems such as predictive or classificatory problems, as association rules in these cases may be not suitable.

In this paper, we have seen a lot of problems that are currently addressed using association rules, as well as their future challenges. With this in mind, we have tried to bring together the current knowledge of these techniques and their applications in different fields of textual entities present in social media. n the current research work, we have read, studied and classified several papers related to association rules in the field of social media mining. Our results indicate that the use of this technique is robust. Finally, we could summarize the conclusions obtained in this paper, through the answers to the research questions that we introduced in Section 1.

– RQ1: According to Section 4 the most relevant tasks implemented by association rules are summarization, event and topic detection, sentiment analysis, forecasting and collaborative social systems. It is necessary to

point out that these may not be the only applications, but they are the most relevant ones, or at least the ones that bring together a considerable number of articles.
- RQ2: As we have seen in Section 5, there are several fields of application where association rules have been applied. These areas are: academic, crime, health, insurance, influence, leisure, natural disasters, politics, sports, transport and trends.
- RQ3: The most relevant future problems to be achieved by association rules are the streaming association rules, the temporal association rules, and the extended association rules by social flags or graphs [81], as pointed out in Section 7.

Acknowledgements The research reported in this paper was partially supported by the COPKIT project under the European Union's Horizon 2020 research and innovation program (grant agreement No 786687), the Andalusian government and the FEDER operative program under the project BigDataMed (P18-RT-2947 and B-TIC-145-UGR18). Finally the project is also partially supported by the Spanish Ministry of Education, Culture and Sport (FPU18/00150).

References

1. Abd Elaty, A.A., Salem, R., Elkader, H.A.: Efficient association rules mining from streaming data with a fault tolerance. In: 2018 13th International Conference on Computer Engineering and Systems (ICCES), pp. 627–632 (2018). DOI 10.1109/ICCES.2018.8639433
2. Abu Daher, L., Elkabani, I., Zantout, R.: Identifying influential users on twitter: A case study from paris attacks. Applied Mathematics and Information Sciences **12**, 1021–1032 (2018). DOI 10.18576/amis/120515
3. Acosta, M.E., Palaoag, T.D.: Characterization of disaster related tweets according to its urgency: A pattern recognition. In: Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, pp. 30–37 (2019)
4. Adedoyin-Olowe, M., Gaber, M., Stahl, F.: A survey of data mining techniques for social network analysis. Journal of Data Mining & Digital Humanities (2014)
5. Adedoyin-Olowe, M., Gaber, M.M., Dancausa, C.M., Stahl, F., Gomes, J.B.: A rule dynamics approach to event detection in twitter with its application to sports and politics. Expert Systems with Applications **55**, 351–360 (2016)
6. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Acm sigmod record, vol. 22, pp. 207–216. ACM (1993)
7. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. very large data bases, VLDB, vol. 1215, pp. 487–499 (1994)
8. Alanezi, M., et al.: Community detection in facebook using visual approach and clustering. In: Journal of Physics: Conference Series, vol. 1804, p. 012047. IOP Publishing (2021)
9. AlSumait, L., Barbará, D., Domeniconi, C.: On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: 2008 eighth IEEE international conference on data mining, pp. 3–12. IEEE (2008)
10. Atsa'am, D.D., Wario, R.: Association rules on the covid-19 variants of concern to guide choices of tourism destinations. Current Issues in Tourism pp. 1–5 (2021)
11. Budgen, D., Brereton, P.: Performing systematic literature reviews in software engineering. In: Proceedings of the 28th international conference on Software engineering, pp. 1051–1052 (2006)
12. Cagliero, L., Fiori, A.: Analyzing twitter user behaviors and topic trends by exploiting dynamic rules. In: Behavior Computing, pp. 267–287. Springer (2012)

13. Calders, T., Dexters, N., Goethals, B.: Mining frequent itemsets in a stream. In: Seventh IEEE International Conference on Data Mining (ICDM 2007), pp. 83–92. IEEE (2007)
14. Cambria, E., Speer, R., Havasi, C., Hussain, A.: Senticnet: A publicly available semantic resource for opinion mining. In: AAAI fall symposium: commonsense knowledge, vol. 10 (2010)
15. Cardinale, Y., Dongo, I., Robayo, G., Cabeza, D., Aguilera, A., Medina, S.: T-creo: A twitter credibility analysis framework. IEEE Access **9**, 32498–32516 (2021)
16. Chen, C., Chou, H., Hong, T., Nojima, Y.: Cluster-based membership function acquisition approaches for mining fuzzy temporal association rules. IEEE Access **8**, 123996–124006 (2020). DOI 10.1109/ACCESS.2020.3004095. URL https://doi.org/10.1109/ACCESS.2020.3004095
17. Cheng, L.C., Sharmayne, L.R.: Analysing digital banking reviews using text mining. In: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 914–918. IEEE (2020)
18. Chintalapudi, N., Battineni, G., Amenta, F.: Sentimental analysis of covid-19 tweets using deep learning models. Infectious Disease Reports **13**(2), 329–339 (2021)
19. Chugh, N., Phumchusri, N.: Bangkok tours and activities data analysis via user-generated content. In: 2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 98–102. IEEE (2020)
20. Dadkhah, S., Shoeleh, F., Yadollahi, M.M., Zhang, X., Ghorbani, A.A.: A real-time hostile activities analyses and detection system. Applied Soft Computing **104**, 107175 (2021)
21. Daher, L.A., Elkabani, I., Zantout, R.: Identifying influential users on twitter's trendy hashtags using association rule learning. In: 2018 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1293–1296. IEEE (2018)
22. Das, S., Dutta, A., Medina, G., Minjares-Kyle, L., Elgart, Z.: Extracting patterns from twitter to promote biking. IATSS Research **43**(1), 51–59 (2019)
23. Dehkharghani, R., Mercan, H., Javeed, A., Saygin, Y.: Sentimental causal rule discovery from twitter. Expert Systems with Applications **41**(10), 4950–4958 (2014)
24. Demirci, G.V., Aykanat, C.: Scaling sparse matrix-matrix multiplication in the accumulo database. Distributed and Parallel Databases **38**(1), 31–62 (2020)
25. Diaz-Garcia, J.A., Fernandez-Basso, C., Ruiz, M.D., Martin-Bautista, M.J.: Mining text patterns over fake and real tweets. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 648–660. Springer (2020)
26. Diaz-Garcia, J.A., Ruiz, M.D., Martin-Bautista, M.J.: Generalized association rules for sentiment analysis in twitter. In: International Conference on Flexible Query Answering Systems, pp. 166–175. Springer (2019)
27. Diaz-Garcia, J.A., Ruiz, M.D., Martin-Bautista, M.J.: Non-query-based pattern mining and sentiment analysis for massive microblogging online texts. IEEE Access **8**, 78166–78182 (2020)
28. Drias, H.H., Drias, Y.: Mining twitter data on covid-19 for sentiment analysis and frequent patterns discovery. medRxiv (2020)
29. Erlandsson, F., Borg, A., Johnson, H., Bródka, P.: Predicting user participation in social media. In: International Conference and School on Network Science, pp. 126–135. Springer (2016)
30. Erlandsson, F., Bródka, P., Borg, A., Johnson, H.: Finding influential users in social media using association rule learning. Entropy **18**(5), 164 (2016)
31. Feng, H., Lesot, M.J., Detyniecki, M.: Using association rules to discover color-emotion relationships based on social tagging. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 544–553. Springer (2010)
32. Fernandez-Basso, C., Francisco-Agra, A.J., Martin-Bautista, M.J., Ruiz, M.D.: Finding tendencies in streaming data using big data frequent itemset mining. Knowledge-Based Systems **163**, 666–674 (2019)
33. Fernandez-Basso, C., Ruiz, M.D., Martin-Bautista, M.J.: Extraction of association rules using big data technologies. International Journal of Design & Nature and Ecodynamics **11**(3), 178–185 (2016)

34. Fu, K., Lu, C.T., Nune, R., Tao, J.X.: Steds: Social media based transportation event detection with text summarization. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pp. 1952–1957. IEEE (2015)
35. Gundecha, P., Liu, H.: Mining social media: a brief introduction. In: New Directions in Informatics, Optimization, Logistics, and Production, pp. 1–17. Informs (2012)
36. Hamed, A.A., Wu, X., Rubin, A.: A twitter recruitment intelligent system: association rule mining for smoking cessation. Social Network Analysis and Mining **4**(1), 212 (2014)
37. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: ACM sigmod record, vol. 29, pp. 1–12. ACM (2000)
38. He, K., He, L., Lin, X., Lu, W.: Social view based user modeling for recommendation in tagging systems by association rules. In: 2010 2nd International Workshop on Intelligent Systems and Applications, pp. 1–5. IEEE (2010)
39. Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H.A., Epstein, D.H., Leggio, L., Curtis, B., et al.: Bots and misinformation spread on social media: Implications for covid-19. Journal of Medical Internet Research **23**(5), e26933 (2021)
40. Huang, X., Tang, Y., Qu, R., Li, C., Yuan, C., Sun, S., Xu, B.: Course recommendation model in academic social networks based on association rules and multi-similarity. In: 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)), pp. 277–282. IEEE (2018)
41. Huizinga, T., Ayanso, A., Smoor, M., Wronski, T.: Exploring insurance and natural disaster tweets using text analytics. International Journal of Business Analytics (IJBAN) **4**(1), 1–17 (2017)
42. Injadat, M., Salo, F., Nassif, A.B.: Data mining techniques in social media: A survey. Neurocomputing **214**, 654–670 (2016)
43. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimedia Tools and Applications **78**(11), 15169–15211 (2019)
44. Jiang, P., Hong, C., Agrawal, G.: A novel data transformation and execution strategy for accelerating sparse matrix multiplication on gpus. In: Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, pp. 376–388 (2020)
45. Jin, R., Agrawal, G.: An algorithm for in-core frequent itemset mining on streaming data. In: Fifth IEEE International Conference on Data Mining (ICDM'05), pp. 8–pp. IEEE (2005)
46. Jung, Y.H., Seo, M.S., Yoo, H.H.: An analysis on the citizen's health by using the twitter data of yellow dust. Journal of Korean Society for Geospatial Information Science **24**(2), 55–62 (2016)
47. Kacprzyk, J., Zadrozny, S.: Linguistic summarization of data sets using association rules. In: The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ'03., vol. 1, pp. 702–707. IEEE (2003)
48. Kakulapati, V., Reddy, S.M.: Mining social networks: Tollywood reviews for analyzing upc by using big data framework. In: Smart Innovations in Communication and Computational Sciences, pp. 323–334. Springer (2019)
49. Kammergruber, W.C., Viermetz, M., Ehms, K., Langen, M.: Using association rules for discovering tag bundles in social tagging data. In: 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pp. 414–419. IEEE (2010)
50. Kasseropoulos, D.P., Tjortjis, C.: An approach utilizing linguistic features for fake news detection. In: I. Maglogiannis, L. Macintyre Johnnand Iliadis (eds.) Artificial Intelligence Applications and Innovations, pp. 646–658. Springer International Publishing, Cham (2021)
51. Krawczyk, B., McInnes, B.T., Cano, A.: Sentiment classification from multi-class imbalanced twitter data using binarization. In: International Conference on Hybrid Artificial Intelligence Systems, pp. 26–37. Springer (2017)
52. Kumar, A., Jaiswal, A.: Systematic literature review of sentiment analysis on twitter using soft computing techniques. Concurrency and Computation: Practice and Experience **32**(1), e5107 (2020)

53. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining text data, pp. 415–463. Springer (2012)
54. Liu, L., Wen, J., Zheng, Z., Su, H.: An improved approach for mining association rules in parallel using spark streaming. International Journal of Circuit Theory and Applications **49**(4), 1028–1039 (2021)
55. Mai, M., Leung, C.K., Choi, J.M., Kwan, L.K.R.: Big data analytics of twitter data and its application for physician assistants: who is talking about your profession in twitter? In: Data Management and Analysis, pp. 17–32. Springer (2020)
56. Mamgain, N., Pant, B., Mittal, A.: Categorical data analysis and pattern mining of top colleges in india by using twitter data. In: 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 341–345. IEEE (2016)
57. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55–60 (2014)
58. Margono, H., Yi, X., Raikundalia, G.K.: Mining indonesian cyber bullying patterns in social networks. In: Proceedings of the Thirty-Seventh Australasian Computer Science Conference-Volume 147, pp. 115–124. Australian Computer Society, Inc. (2014)
59. Martin-Bautista, M., Sánchez, D., Serrano, J., Vila, M.: Text mining using fuzzy association rules. In: Fuzzy logic and the internet, pp. 173–189. Springer (2004)
60. Mora-Cantallops, M., Sánchez-Alonso, S., Visvizi, A.: The influence of external political events on social networks: The case of the brexit twitter network. Journal of Ambient Intelligence and Humanized Computing **12**(4), 4363–4375 (2021)
61. Mosley Jr, R.C.: Social media analytics: Data mining applied to insurance twitter posts. In: Casualty Actuarial Society E-Forum, vol. 2, p. 1. Citeseer (2012)
62. Nancy, P., Ramani, R.G., Jacob, S.G.: Mining of association patterns in social network data (face book 100 universities) through data mining techniques and methods. In: Advances in Computing and Information Technology, pp. 107–117. Springer (2013)
63. Nenkova, A., McKeown, K.: A survey of text summarization techniques. In: Mining text data, pp. 43–76. Springer (2012)
64. Noferesti, S., Shamsfard, M.: Resource construction and evaluation for indirect opinion mining of drug reviews. PLOS ONE **10**(5), 1–25 (2015)
65. Ogihara, Z.P., Zaki, M., Parthasarathy, S., Ogihara, M., Li, W.: New algorithms for fast discovery of association rules. In: In 3rd Intl. Conf. on Knowledge Discovery and Data Mining. Citeseer (1997)
66. Paulose, R., Samy, B.G., Jegatheesan, K.: Text mining and natural language processing on social media data giving insights for pharmacovigilance: A case study with fentanyl. Indian Journal of Pharmaceutical Sciences **80**(4), 762–766 (2018)
67. Phan, H.T., Nguyen, N.T., Hwang, D.: A tweet summarization method based on maximal association rules. In: International Conference on Computational Collective Intelligence, pp. 373–382. Springer (2018)
68. Ramamonjisoa, D., Suzuki, E., Hamid, I.: Research topics discovery from www by keywords association rules. In: International Conference on Rough Sets and Current Trends in Computing, pp. 412–419. Springer (2000)
69. Rao, P.G., Khan, M.Z., Rafeeq, M., Hitesh, N., Shenoy, P.D., Venugopal, K.: An automated learning system for twitter trends. In: 2019 Fifteenth International Conference on Information Processing (ICINPRO), pp. 1–6. IEEE (2019)
70. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the 2011 conference on empirical methods in natural language processing, pp. 1524–1534 (2011)
71. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation (2004)
72. Rodavia, M.R.D., Cuison, L.T., Barcelo, A.: Detecting flood vulnerable areas in social media stream using association rule mining. In: 2018 International Conference on Platform Technology and Service (PlatCon), pp. 1–6. IEEE (2018)
73. Shen, D., Zhang, L., Cao, J., Wang, S.: Forecasting citywide traffic congestion based on social media. Wireless Personal Communications **103**(1), 1037–1057 (2018)

74. Shofiya, C., Abidi, S.: Sentiment analysis on covid-19-related social distancing in canada using twitter data. International Journal of Environmental Research and Public Health **18**(11), 5993 (2021)
75. Si, H., Zhou, J., Chen, Z., Wan, J., Xiong, N.N., Zhang, W., Vasilakos, A.V.: Association rules mining among interests and applications for users on social networks. IEEE Access **7**, 116014–116026 (2019)
76. Solanki, S.K., Patel, J.T.: A survey on association rule mining. In: 2015 fifth international conference on advanced computing & communication technologies, pp. 212–216. IEEE (2015)
77. Srikant, R., Agrawal, R.: Mining generalized association rules. Future generation computer systems **13**(2-3), 161–180 (1997)
78. Tandan, M., Acharya, Y., Pokharel, S., Timilsina, M.: Discovering symptom patterns of covid-19 patients using association rule mining. Computers in biology and medicine **131**, 104249 (2021)
79. Tundis, A., Jain, A., Bhatia, G., Muhlhauser, M.: Similarity analysis of criminals on social networks: An example on twitter. In: 2019 28th International Conference on Computer Communication and Networks (ICCCN), pp. 1–9. IEEE (2019)
80. Utami, M.P., Nurhayati, O.D., Warsito, B.: Hoax information detection system using apriori algorithm and random forest algorithm in twitter. In: 2020 6th International Conference on Interactive Digital Media (ICIDM), pp. 1–5. IEEE (2020)
81. Wang, X., Xu, Y., Zhan, H.: Extending association rules with graph patterns. Expert Syst. Appl. **141** (2020). DOI 10.1016/j.eswa.2019.112897. URL https://doi.org/10.1016/j.eswa.2019.112897
82. Xiao, Y., Zhang, R., Kaku, I.: A new framework of mining association rules with time-windows on real-time transaction database. Int. J. Innov. Comput. Inf. Control **7**(6), 3239–3253 (2011)
83. Xie, M., Tan, L.: An efficient algorithm for frequent pattern mining over uncertain data stream. In: 2019 12th International Symposium on Computational Intelligence and Design (ISCID), vol. 1, pp. 84–88. IEEE (2019)
84. Xin, M., Wu, L., Liang, W., Shu, J.: An approach to the mobile social services recommendation algorithm based on association rules mining. International Journal of Services Technology and Management **26**(2-3), 115–130 (2020)
85. Yadav, N., Mundotiya, R.K., Singh, A.K.: Tag-based personalized collaborative movie recommender system. Journal of Information Assurance & Security **16**(1) (2021)
86. Yang, Z., Tang, J., Li, J., Yang, W.: Social community analysis via a factor graph model. IEEE Intelligent Systems **26**(03), 58–65 (2011)
87. Yousaf, A., Umer, M., Sadiq, S., Ullah, S., Mirjalili, S., Rupapara, V., Nappi, M.: Emotion recognition by textual tweets classification using voting classifier (lr-sgd). IEEE Access **9**, 6286–6295 (2020)
88. Zainol, Z., Wani, S., Nohuddin, P.N., Noormanshah, W.M., Marzukhi, S.: Association analysis of cyberbullying on social media using apriori algorithm. International Journal of Engineering & Technology **7**(4.29), 72–75 (2018)
89. Zainuddin, N., Selamat, A., Ibrahim, R.: Hybrid sentiment classification on twitter aspect-based sentiment analysis. Applied Intelligence **48**(5), 1218–1232 (2018)
90. Zhang, C., Wu, X., Shyu, M.L., Peng, Q.: Adaptive association rule mining for web video event classification. In: 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), pp. 618–625. IEEE (2013)
91. Zhang, Z., Wang, H., Han, S., Dally, W.J.: Sparch: Efficient architecture for sparse matrix multiplication. In: 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 261–274. IEEE (2020)
92. Zheng, L.: Research on e-commerce potential client mining applied to apriori association rule algorithm. In: 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), pp. 667–670. IEEE (2020)

2 Non-Query-Based Pattern Mining and Sentiment Analysis for Massive Microblogging Online Texts

2.1 Non-Query-Based Pattern Mining and Sentiment Analysis for Massive Microblogging Online Texts

- Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2020). Non-query-based pattern mining and sentiment analysis for massive microblogging online texts. IEEE Access, 8, 78166-78182.
  - Journal: IEEE Access
  - Status: Published.
  - Impact Factor (JCR 2021): 3.476.
  - Category: Computer Science, Information Systems. Order 79/164 Q2.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

Non-query based pattern mining and sentiment analysis for massive microblogging online texts

JOSE A. DIAZ-GARCIA[1], M. DOLORES RUIZ [2], AND MARIA J. MARTIN-BAUTISTA[3]
[1]Department of Computer Science and A.I., University Of Granada, Daniel Saucedo Aranda, s/n, 18071 Granada (e-mail: jagarcia@decsai.ugr.es)
[2]Department of Statistics and O.R., University Of Granada, Avenida de Fuente Nueva, s/n, 18071 Granada (e-mail: mariloruiz@ugr.es)
[3]Department of Computer Science and A.I., University Of Granada, Daniel Saucedo Aranda, s/n, 18071 Granada (e-mail: mbautis@decsai.ugr.es)

Corresponding author: J. Angel Diaz-Garcia (e-mail:jagarcia@decsai.ugr.es).

This research paper is part of the COPKIT project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687.

ABSTRACT Pattern mining has been widely studied in the last decade given its great interest for research and its numerous applications in the real world. In this paper the definition of query and non-query based systems is proposed, highlighting the needs of non-query based systems in the era of Big Data. For this, we propose a new approach of a non-query based system that combines association rules, generalized rules and sentiment analysis in order to catalogue and discover opinion patterns in the social network Twitter. Association rules have been previously applied for sentiment analysis, but in most cases, they are used once the process of sentiment analysis is finished to see which tokens appear commonly related to a certain sentiment. On the other hand, they have also been used to discover patterns between sentiments. Our work differs from these in that it proposes a non-query based system which combines both techniques, in a mixed proposal of sentiment analysis and association rules to discover patterns and sentiment patterns in microblogging texts. The obtained rules generalize and summarize the sentiments obtained from a group of tweets about any character, brand or product mentioned in them. To study the performance of the proposed system, an initial set of 1.7 million tweets have been employed to analyse the most salient sentiments during the American pre-election campaign. The analysis of the obtained results supports the capability of the system of obtaining association rules and patterns with great descriptive value in this use case. Parallelisms can be established in these patterns that match perfectly with real life events.

INDEX TERMS Query systems, non-query systems, pattern mining, association rules, sentiment analysis, social media mining

I. INTRODUCTION

Data Mining techniques, despite their recent novelty, are present in almost all research and development areas that human beings are currently working on. There are certain areas where these techniques stand out, remarkably influenced by the new economic and social tendencies where social networks have gained importance. These areas are, for instance, the detection of communities [1], studies and tools focused on marketing [2], the development of predictive models in financial or insurance fields [3] and of course, mining of social networks or sentiment analysis [4], [5].

This last one has currently become one of the most studied aspects due to the growing interest in understanding users habits using more reliable analysis tools. In this field, known as Sentiment Analysis, Data Mining techniques are used to obtain relevant information from textual data coming from online social networks. Sentiment analysis includes the techniques of text mining, natural language processing and automatic learning that focus on obtaining sentiment aspects from texts. The final objective is to obtain sentiments or polarities from unstructured data coming, for example, from consumer opinions of certain products. This information is very valuable for brands and can provide competitive advantages. For this reason, every day there are more companies using these techniques in their technological surveillance processes to obtain consumer feedback.

In this area, the approaches about sentiment classification predominate [6], [7], [8], but given the textual character of the input data, other techniques, such as association rules have also been applied over data from social networks with

remarkable results. The purpose of association rules is the discovery of patterns in transactional databases. These patterns represent hidden co-occurring relations between various items (products, words) within a database. The discovery of patterns is therefore called pattern mining and is one of the most used techniques due to its easy interpretation and fields of application. We propose a mixed approach that can be used to obtain patterns and make, in a latter stage, a sentiment analysis based on these patterns. The purpose of our proposal is to develop a system capable of obtaining descriptive patterns, both textual and opinion, in an unsupervised manner. In other words, a system that listens to the social network Twitter and is able to discover the most talked about topics at the time, and the sentiments behind the comments (tweets). Therefore, the proposed system is designed to listen, find and highlight any type of relation between topics or terms on Twitter during the creation of the data lake. To contrast the performance, it has faced a political case use in which we can see, for example, the connection between Donald Trump and Iowa or Hillary Clinton's e-mails.

To achieve this, our methodology obtains opinion patterns and their relation within a small textual transaction (tweet) using an approach based on association rules. Moreover, once this has been obtained, the opinion concepts (words) will be automatically tagged using sentiment analysis into the 8 sentiments or basic emotions (trust, anger, anticipation, disgust, joy, fear, sadness and surprise) characterized by Plutchik [9] in order to generalise and offer a second source of information to complement the opinion patterns obtained in the first stage. As we have previously introduced, we will rely on the use of association rules and generalized association rules, concepts that are explained in next section.

Following the above discussion, our proposal presents a new mixed approach for the fields of pattern mining and sentiment analysis from the point of view of a non-query based system, which as we will define in Section II are those systems whose collected data is not influenced by the topic under study, i.e. the core of this kind of systems lies in the absence of prior filtering. Additionally, our approach combines two well-differentiated techniques: Association Rules and Sentiment Analysis. These techniques have been employed in numerous studies where the value of association rules to summarize and discover knowledge from large data sets has been verified [10], as well as the great importance of sentiment analysis to complement the analysis in domains where these techniques can be applied. The present work uses both techniques generalized association rules and sentiment analysis to improve the core of the process. This differs practically from all previous studies, where association rules are applied to improve the step after sentiment analysis, classifying textual entities, such as tweets, into good, neutral or bad opinion without obtaining patterns on the factors that imply those results.

The contribution of this study to the fields of pattern mining and sentiment analysis are:

- The theoretical definition of query and non-query based systems, as well as the value of the latter for Big Data problems.
- The design of a non-query based system that is capable of working without topic filtered tweets. This differs from literature, where all the works are query based and tweets are filtered according to a specific topic depending on the problem under study.
- The proposal of a methodology capable of processing a large set of tweets in an efficient way which transforms the corpus of tweets into textual transactions. This point differs from other studies because the volume of data studied in most of them is very limited and far from real problems. To validate the performance of the proposed system and offer the best solution, it has been experimented and compared with three of the most widely used pattern mining algorithms.
- A detailed review of published studies applying association rules in the field of social media mining (including Twitter analysis) has been carried out.
- Finally, we propose a new approach for sentiment analysis using generalized association rules, capable of summarizing a very huge set of tweets in a set of rules based on the 8 emotions characterized by Plutchik [9]. These rules will represent the sentiments aroused by the items under study.

The methodology followed by the system to achieve this goal is shown in Figure 1. The first step is to get the Twitter data using a crawler. After this, the data is stored in NoSQL databases, creating a large data lake of social media data. In later stages, the data is loaded from these data lake and the preprocessing procedure begins, cleaning the data and identifying the interesting items. The core of the methodology is based on two stages, on the one hand, the identification of sentiments, using for that sentiments lexicons. On the other hand, the extraction of patterns using the words that form each tweet. The final stage connects these two previous steps into one, using the identified sentiments to generalize the association rules. The results are then visualized by a cloud of terms about a topic, for a character in our case, and a set of rules.

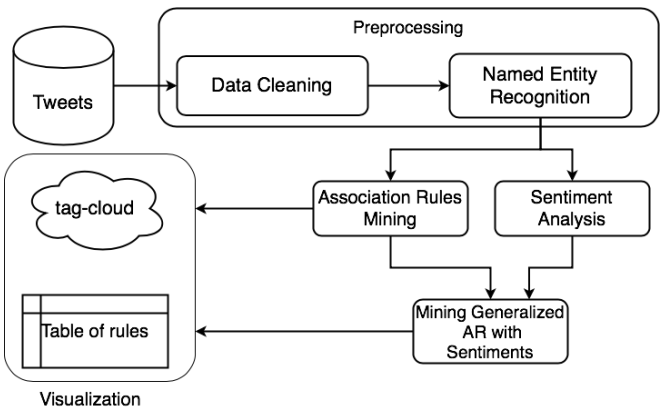

FIGURE 1—Methodology flow

After reviewing the literature we have not found any article or application that can be used as a benchmark for the proposed system for obtaining opinion patterns, so for the validation of the system, we apply it to an use case of a contrasting event in real life. Two well-known US politicians, Donald Trump, and Hillary Clinton have been chosen. The reason for choosing these characters, among all of the people that the system discovered as relevant in the social network Twitter, is that we can contrast obtained results according to the events that occurred in 2016. With this purpose, we perform an analysis of the patterns, rules and generalized rules by sentiments that the system is capable of obtaining. It should be noted that the system can be applied to other topics or other characters present in the data set. Therefore, it could be extended to other people in the political panorama or in the world of entertainment as well as in other topics such as products in marketing.

Regarding the results obtained on the use case described above and which will be analysed in detail in Section VI we can conclude that they are satisfactory. In this sense, our system is capable of obtaining patterns, association rules and sentiments which can be related to events that took place in the election campaign. This relation between real-life events and the obtained patterns can be traced in a descriptive way, as the patterns correspond to policies adopted or to be adopted, to disputed voting places or to confrontations between candidates.

The paper is structured as follows: Section II reviews some of the related theoretical concepts that allow to understand the following sections. Section III describes the related work. Section IV explains the followed methodology. Section V-A includes the experimentation carried out. Finally, Section VI puts in value the system, by means of a real use case in which obtained patterns and information are compared with real life events.The paper ends with a discussion and analysis of the proposed approach and the future lines that this work opens.

II. BASIC CONCEPTS

In this section, we introduce some theoretical concepts that are required to understand the techniques employed in our proposal. Firstly, we start with the definition of query and non-query based systems. After this, we review association rules and then we continue with their generalization.

QUERY AND NON-QUERY BASED SYSTEMS

Nowadays, in the Big Data era, organizations and companies can generate a great volume of data, from which they will be able to obtain great advantages in the future, but which are unknown at the moment of gathering and storing the data. This has led companies to invest more and more resources in the generation of data lakes, large volume of non-relational data stored without prior knowledge. Many of these companies operate in social networks, and collect data and conversations that users generate about them. Therefore the need to have systems that can handle with these data and obtain information in the future is accentuated.

It is at this point where we distinguish between query and non-query based systems. A system query based, will obtain a reduced dataset which will be limited to its domain according to a concrete need of information. Afterwards, the typical tasks of pre-processing and data mining will be carried out to obtain results. On the opposite, a non-query based system does not impose a filter before collecting data, so a huge data lake is created. In this case, the need of information, that will guide the mining process, will come later and will be linked to the pre-processing. In this case data pre-processing, will be more difficult because the data volume is higher, but it opens a world of possibilities for the extraction of inter-topic and cross-subject knowledge. It is necessary to mention that non-query based systems are also the most appropriate for Big Data problems, and more specifically those coming from social networks, due to the large amount of data produced, as well as, the speed of generation of these, which makes it almost impossible to have the pertinent queries before knowing the kind of data that will be generated in the social channel. Therefore, non-query based systems are the most appropriate solution in social media applications or when the topic under study are not fixed beforehand. For a better understanding of these definitions, their comparison can been seen in Figure 2.

According to Figure 2, in non-query based systems the user has the possibility of creating a large data lake on which to perform unsupervised analysis without the interference of data that could come from a previous filtering. These data coming from a query based system would be more cohesive but far from a real social network problem.

ASSOCIATION RULES

Association rules belong to Data Mining field and have been used and studied for a long time. One of the first references to them dates back to 1993 [11]. They are used to obtain relevant knowledge from large transactional databases. A transactional database could be for example, a shopping basket database, where the items would be the products, or a text database, as is our case, where the items are the words. In a more formal way, let $t=\{A,B,C\}$ be a transaction of three items *(A, B and C)*, and any combination of comprised them forms an e.g. itemset we would have *{A,B,C}, {A,B}, {B,C}, {A,C}, {A},{A}, {B} and {C}*. According to this, an association rule would be represented in the form $X \rightarrow Y$ where X is an itemset that represents the antecedent and Y an itemset called consequent. As a result, we can conclude that consequent items have a co-occurrence relation with antecedent items. Therefore, association rules can be used as a method of extracting hidden relation between items or elements within transactional databases, data warehouses or other types of data storage from which it is interesting to extract information to help in decision-making processes. The classical way of measuring the goodness of association rules regarding a given problem is with three measures: support, confidence and lift, which are defined as follows:

- Support of an itemset. It is represented as *supp (X)*, and

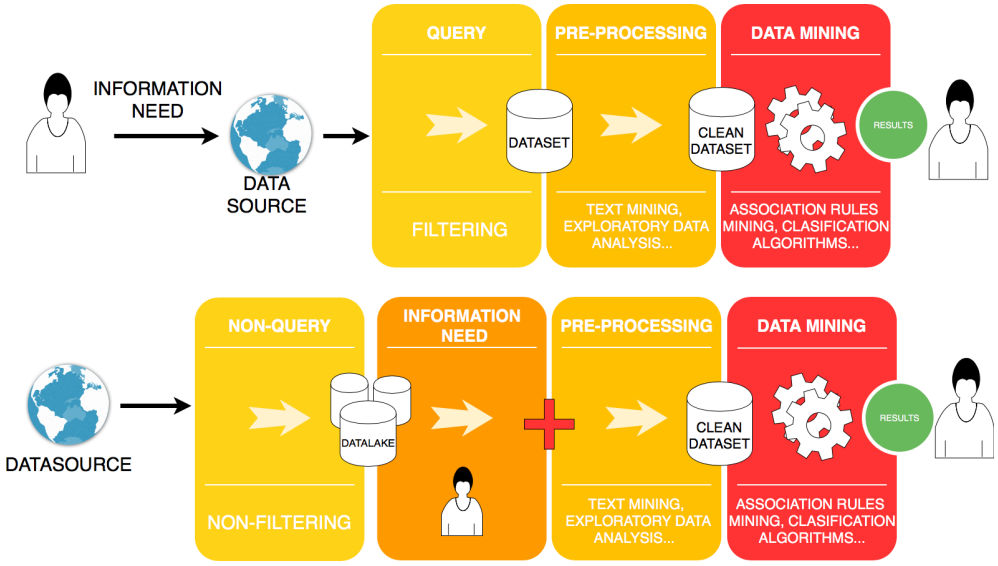

FIGURE 2—Comparison between query and non-query based systems

is the proportion of transactions containing item X out of the total amount of transactions of the dataset (D). The equation to define the support of an itemset is:

$$supp(X) = \frac{|t \in D : X \subseteq t|}{|D|} \quad (1)$$

- Support of an association rule. It is represented as *supp(X → Y)*, is the total amount of transactions containing both items X and Y, as defined in the following equation:

$$supp(X \rightarrow Y) = supp(X \cup Y) \quad (2)$$

- Confidence of an association rule. It is represented as *conf (X→Y)* and represents the proportion of transactions containing item X which also contains Y. The equation is:

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \quad (3)$$

- Lift. It is a useful measure to assess independence between items of a certain association rule. The measure *lift (X→Y)* represents the degree to which X is frequent when Y is present or vice versa. Lift is defined mathematically in the following way:

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)} \quad (4)$$

Association rules can be extracted using several approaches. One option is a brute-force based approach which is not very efficient. The most widespread approach is based on two stages using the downward-closure property. The first of these stages is the generation of frequent itemsets. To be considered frequent the itemset have to exceed the minimum support threshold. In the second stage the association rules are obtained using the minimum confidence threshold. Within this category we find most of the algorithms for obtaining association rules, such as Apriori, proposed by Agrawal and Srikant [12], FP-Growth proposed by Han et al. [13] and Eclat [14]. Although these are the most widespread approaches, there are other frequent itemset extraction techniques such as vertical mining or pattern growth.

GENERALIZED ASSOCIATION RULES

Association rules can be studied and interpreted from a hierarchical point of view [15], for example, in a shopping basket, the rule *{Apples, Bananas} → {Yogurt}* could be replaced by *{Fruit} → {Yogurt}*. This allows us to achieve a greater degree of data abstraction, which is interesting in order to obtain new relevant information. This abstraction or processing it also useful to summarise the set of rules enormously. This will result in a simpler analysis of the problems without losing relevant information which, in any case, could be easily recovered. Generalized association rules are also interesting for Big Data environments because the size of the obtained results can be highly summarized, improving consequently processing time and resources. This will have an impact on goodness indicators, providing stronger rules.

III. RELATED WORK

The field of opinion mining in social networks is relatively new, due, undoubtedly, to the novelty of social networks. Twitter was founded in 2006 and Facebook in 2005, which gives us an average of about 14 years of life for the most famous social networks. On the other hand, we must bear in mind that their establishment within society did not take place since their foundation, so that their 'age' is even lower. If we focus only on the IT aspects of the study, they are also remarkably new, because despite being widely studied we are still at the dawn of what data mining could offer

in the future. The novelty of these techniques justifies the few approaches more related to our proposal that we have found in the literature so far. However, it is also one of the research fields generating more interest among the scientific community.

In this section we will start by studying and discussing traditional and deep learning based classification techniques in the field of sentiment analysis and text mining. Finally, we will analyse those proposals that use association rules, in which we will go into detail because they are the main thread of this paper's research.

MACHINE LEARNING AND DEEP LEARNING IN SOCIAL MEDIA

In the field of classification of textual entities by their sentiments, supervised methods stand out. We can find several very recent papers such as [16], [17] in which different directed methods are compared, such as decision trees for the classification of tweets according to sentiments. We also can find probabilistic methods such as those described in [18], [19] or [20] where the authors applied Latent Dirichlet Allocation in [18], [19] and Bayesian networks in [20]. All the papers are focused on classifying the sentiments in different scenarios of great activity in Twitter, as well as in streaming. Finally, it is necessary to mention that deep learning [21], [22], [23], [24] has been employed in the field of text mining. Deep learning solutions offer a priori good results, but they present two main drawbacks. On one hand, the fact that they are black boxes with a lack of interpretability of their outputs. On the other hand, they need a strong effort in collecting already classified cases in order to adapt them to each use case. It is at this point where our work stands out trying to provide an easily interpretable unsupervised solution.

SENTIMENT ANALYSIS IN SOCIAL MEDIA BASED ON NETWORK METRICS

In the scope of Twitter and because this is a social network there are also approaches that attempt to address the problem of sentiment analysis in social networks through the usual developed methods in graph theory. This is due to the fact that the relations in Twitter are established in a directed way between user and follower. In this area we can find the paper [25], in which the authors by means of network measures such as betweenness centrality and applying machine learning techniques, obtain correlations between the sentiments in retweets and regular tweets. According to the factor of the network metrics in this field, papers such as [26] and [25] predominate.These papers try to find relevant users in the social network by analysing their publications and connections. This is a very interesting research path, which deals with the relation of textual entities and network topology.

In our proposal, we want to analyse the content generated in the social network as if it were modelled in a large dataset or data lake, from which we can obtain information, regardless of where it was generated.

ASSOCIATION RULES AND SOCIAL MEDIA

One of the main studies in the field of association rules is the one proposed in 2000 by Silverstein et al. [27]. In this study association rules are used for the well-known problem of shopping baskets, which relates the purchase of a certain product with the possibility of buying a different one. After this, the topic has been extensively studied and applied in many interesting research articles, although the use of association rules in social media does not appear in an article until 2010. The article is proposed by Oktay et al. [28] and studies the relation between the appearance of terms in questions of the Stack Overflow website with the appearance of terms in the answers to these questions. In a way it is related to our study, in which we try to obtain and interpret the relation between terms, although the techniques and domain differ completely. The use of association rules in social networks has been shown in papers such as the one proposed by Erlandsson et al. [29], in which, an analysis based on association rules to find influencers on Facebook is put forward. If we focus on our domain, Twitter, the work of Pak and Paroubek [30], has been a starting point to stand out Twitter as an important source for opinion and sentiment analysis. Attending to the use of association rules in Twitter the domain of studies and applications is diverse.

One of the most studied research areas using association rules on Twitter is the information summarization, either of users by their influence on the microblogging network [31] or of the most relevant items (tweets, post) [32]. In these areas, graph theories and other types of association rules, such as maximal rules, also come into the equation. Again, in both proposals the problem resides in the number of tweets whose volume is very low. This makes difficult to transfer the results to a real problem, where the volume of data and variables would be higher. On the other hand, both proposals highlight the power of association rules for summarizing information and obtaining patterns in the social network, something that our work also does but on a larger scale (i. e., with a higher number of tweets). A paper that can be included within the scope of summarization is the paper [33]. In this paper, the authors use association rules for datasets coming from Twitter trying to obtain hashtags and the terms related to them. This paper, although interesting, does not make a descriptive study of the obtained results but compares different algorithms to generate a new hashtag-oriented proposal without offering a real use case.

The approaches in [34] and [35] propose the detection of word patterns associated with cyberbullying to detect these behaviours through the social network Twitter, although in these experiments the domain is very small and the number of tweets is very limited (see Table 1). This specificity of the domain is also found in the work [36] where Hamed et al. proposed a system based on association rules to determine the co-occurrence of hashtags in the field of smoking, with the aim of creating an expert system to stop smoking. Also in the field of pattern mining, but in the domain of insurance and with a higher volume of input tweets for the mining process,

we find the work proposed by Mosley and Roosevelt [37]. In this work the authors use association rules and clustering to obtain interesting patterns related to people and insurance. Our proposal is linked to these articles, with the difference that we use a non-query based system and the amount of the data used is much higher (see Table 1), so that the patterns that are obtained can be considered stronger, because they appear with more representation.

All the previous approaches show that association rules are used with enough regularity in the domain of the microblogging networks like Twitter, although, as it has been verified, with a serious limitation with the size of the input. This limitation is saved by some works that could be framed within the scope of Big Data. Here we find the work proposed by Adedoyin-Olowe et al. [38] and the work proposed by Fernandez-Basso et.al [39] where in streaming, association rules were employed in the first case, and frequent itemset extraction in the latter case, for the detection of events in the sports field, or politics. In the first work around 3.8 millions of tweets are used although, they are partitioned to later simulate the streaming. Also framed within the Big Data paradigm, but only for the methodology, because the volume of tweets, is very small we find the work [40] which makes a system of film recommendations that feeds on Imdb [1] and Twitter. Our proposal differs from these two approaches, in the volume of data used and plus our system can obtain association rules while in the presented by Fernandez-Basso only frequent itemsets are obtained. Moreover, the other two systems proposed in [38] and [40], use association rules for the detection of streaming topics ignoring the analysis of sentiments about, for example, the detected topics, something that our proposal does. According to the patterns revealed about people, a later stage of identification of sentiments is carried out by our proposal offering a great amount of information to the final user.

ASSOCIATION RULES AND SENTIMENT ANALYSIS IN SOCIAL MEDIA

Regarding the fields of sentiment analysis and association rules, there are few related studies due to the predominance of classification methods [41], [42] in these areas. We find studies such as the one of Hai et al. [43] where an approach based on association rules, co-occurrence of words and clustering is applied to obtain the most common characteristics regarding certain groups of words that can represent an opinion. The purpose of the study is to go a step forward in sentiment analysis, which usually only classifies an opinion. The proposed method not only classifies, but also the user can see what words or opinion characteristics have been employed in the classification. The work of Yuan et al. [44] proposes a new measure for the discrimination of frequent terms without apparent orientation of the opinions, which favours the subsequent process of sentiment analysis. Linked to this point is the study made by Dehkharghani et al. [45] where it is proposed the use of association rules to link the co-occurrence of terms in tweets, which are subsequently classified according to the sentiments of these terms included in the obtained rules. In broad terms, the link between these studies is the use of association rules and frequent itemsets to improve the process of sentiment analysis. This differs from our study in that once the rules are mined, we use a hierarchical approach using generalized rules to improve the interpretation of the association rules.

In this field of study, we have found two methods proposing a mixed approach of association rules and sentiment analysis to obtain patterns on Twitter. The one proposed by Mamgain et al. [46] and the one proposed by Bing et al. [47]. Both propose a previous stage of sentiment analysis, by associating sentiments to each item and, afterwards, they obtain patterns using the Apriori algorithm. The former work creates a model that can help students to choose the best college in India and the latter applies it for stock market prediction. The strength of using both tools from a mixed approach is therefore contrasted in the literature, although in both proposals the number of tweets they employed is very limited (see Table 1) and the domain of use and application very specific. Our proposal differs from these, in that the domain of the problem is not limited or filtered previously, as well as the volume of tweets used is much higher. Our proposal also differs from these previous ones since we use generalized association rules for sentiment analysis.

GENERALIZED ASSOCIATION RULES IN SOCIAL MEDIA

Hierarchical approaches in the process of mining association rules have being studied lately, due in large part to the need of condensing the information they represent, for example, to improve visualization processes. A recent example of this use is put forward by Hahsler and Karpienko [48] where a matrix-based visualization, which makes use of a hierarchical simplification of the items that form the association rules, is proposed. In the present study, the hierarchical approach is also used to simplify or generalized the rules, but instead of doing this by categories of items, we do it by sentiments. Other approaches that use generalized association rules on Twitter are [49] and [50] both proposed by Cagliero and Fiori. In the former, the authors use dynamic association rules, that is, rules where confidence and support measures change over time, in order to obtain data on user habits and behaviours on Twitter, and those rules are latter generalized to get stronger rules. In the latter, the authors proposed to generalize the rules obtained from tweets according to taxonomies such as places, time or context, so that they can be used to analyse content propagation or evolution in time. Our work employs generalized association rules by using the sentiments obtained in the previous process of sentiment analysis, instead of using places or contexts like the other proposals described in this section. With this use, the system is able to start from a set of unfiltered data and then obtain patterns showing the distribution of sentiments in data, based on a specific topic on which the user wants to

[1] Internet Movie Database

obtain information. This functionality was already developed in our paper [51], where there was a first preliminary test of using generalized association rules but on a dataset filtered on two people (Hillary Clinton and Donald Trump). That is, this was a query-based system and on which non traditional pattern mining analysis was carried out.

To conclude this section we have compiled in Table 1 all the related work reviewed that use Twitter as a corpus for the subsequent process of sentiment analysis. It is necessary to mention, that all the systems seen in this point, are query based, because they all filter the data to generate a condensed dataset over which to apply the techniques of data mining, something that our proposal does not make being, as far as we know, the first non-query based proposal in the field of social media mining.

IV. PROPOSED METHODOLOGY

In this section, we present our methodology. A summary of it is depicted in Figure 1 where we can see the four stages of the methodology: pre-processing, named entity recognition, data mining (composed of association rule mining and sentiment analysis) and the combination of both using generalized association rules by sentiments and their corresponding associated visualization.

A. PRE-PROCESSING

The data coming from Twitter is very varied and noisy. This is because it is user-generated content and is susceptible to typing errors, colloquial expressions and other possible variations of a textual entity. Because of this, a first pre-processing stage is necessary to normalize and clean the data. With this, we will get better results in future stages.

In the Figure 3, we have added an example of processing two tweets, to better understand Sections IV-A and IV-B. In the figure, we can see the flow how the of two tweets are transformed until the moment we start using association rules.

1) Data cleaning

The cleaning process uses very standardized methods in the field of text mining. These techniques are:

1) *Elimination of empty words in English*. We have eliminated empty English words, such as articles, pronouns and prepositions. Empty words from the problem domain have been also added, such as, the word via, which can be considered empty since in Twitter it is common to use this word to reference some account from which information is extracted.
2) *Links removal*. Given the scope of the problem, this task aims to identify the main social networks that are used to share links on Twittter, such as Facebook, Youtube, SmartURL, Vine, OwLy or BitLy among others. This identification made by means of regular expressions eliminating their occurrence.
3) *Elimination of punctuation marks and non-alphanumeric characters*.

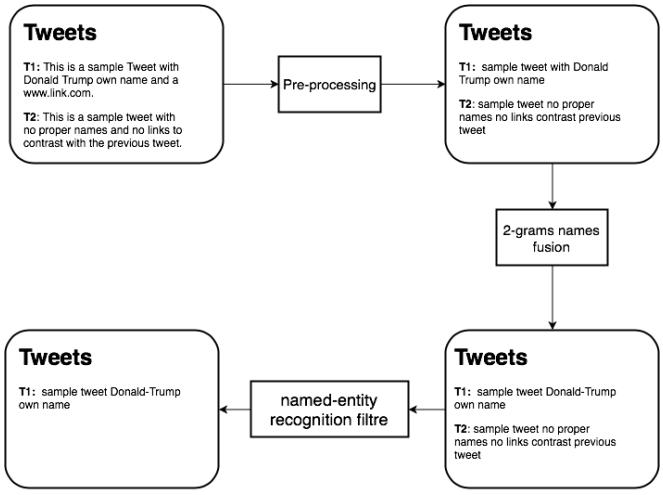

FIGURE 3—Preprocessing and named entity recognition flow.

4) *Empty tweets removal*. After the cleaning process, we may find tweets formed only by empty words, links or any other type of previously deleted string, these tweets will have an empty string in the text column. To reduce the problem, the tweets without text are located and eliminated from the data lake for not taking them into account in later stages.
5) *Unusual terms*. If we try to identify trends or opinion patterns, a word that appears in the dataset very infrequently could hardly be considered part of a trend or opinion. These words only introduce noise in the dataset so they are eliminated to avoid future problems or incoherent rules. Therefore, the words with an occurrence frequency of less than 30 are eliminated, in addition to those words that, despite having more than 30 occurrences, have length longer than 13 letters[2], which would indicate they come from hash-tags or unions of words that are not meaningful for our purpose. Although the process of obtaining frequent items would obviate these items because they are not frequent, we have made this previous elimination to enhance the size of the intermediate data structures since we are facing a Big Data approach problem.

It should be noted that we have avoided the steaming process (i.e saving only the lexical roots of each word) because we believe that interpretability could be lost in subsequent processes to mine association rules.

2) N-grams

Since our interest will focus on tweets that refer to people, we can expect the possibility of obtaining compound names for which a joint analysis is much more interesting and to a certain extent, this will avoid the appearance of redundant

[2]The average English word length is 5 letters [52], and the largest meaningful words found in our dataset are at most 13 letters long, such as international or relationship.

Reference	N tweets	Purpose	SA	PM	GAR	Query or Non-Query
[34]	8275	Detection of cyberbullying patterns.	No	Yes	No	Query
[35]	14000	Detection of cyberbullying patterns.	No	Yes	No	Query
[36]	35000	Co-occurrence of hashtags for expert system elaboration to stop smoking.	No	Yes	No	Query
[37]	68370	Obtaining patterns on the field of insurance.	No	Yes	No	Query
[31]	24026	Identifies the most active users on Twitter during attacks in Paris.	No	Yes	No	Query
[32]	500	Automatically summarizes the most relevant tweets about Barack Obama.	No	Yes	No	Query
[40]	20000	Movie recommendations.	No	Yes	No	Query
[38]	3837291	Detects streaming events in politics and sports.	No	Yes	No	Query
[41]	57000	Analyse political tweets about the Australian elections.	Yes	No	No	Query
[42]	80563	Obtain patterns for the promotion of cycling.	Yes	Yes	No	Query
[46]	8772	Create a system of obtaining patterns to choose the best university in India.	Yes	Yes	No	Query
[47]	150000	Stock prediction.	Yes	Yes	No	Query
[45]	3000	Resume Twitter debates about the Kurds in Turkey.	Yes	Yes	No	Query
[49]	450000	Topic detection.	No	Yes	Yes	Query
[50]	450000	Studies of the propagation and the temporal evolution.	No	Yes	Yes	Query
[51]	140000	Sentiment Analysis about two politicians using Association Rules.	Yes	No	Yes	Query
Our proposal	**1700000**	**Get patterns about sentiments and sentiments in Twitter.**	**Yes**	**Yes**	**Yes**	Non-query

TABLE 1: Comparison of proposals according to the number of tweets, use of sentiment analysis (SA), pattern mining (PM) and generalized association rules (GAR).

association rules. The idea is to merge terms like donald trump into a single term, donald-trump, so we get stronger rules. N-grams are a widely used technique in text mining and information retrieval, which is based on the probability of co-occurrence [53], [54]. That is, for a given term we study the following n terms to discover proper names formed by two words. We will carry out a study of our tweets based on bigrams to get better results in a later data mining stage. To obtain the bigrams, we use a tokenizer from the RWeka package [55], after which we can see which words appear most frequently together with a simple bar graph as shown in Figure 4.

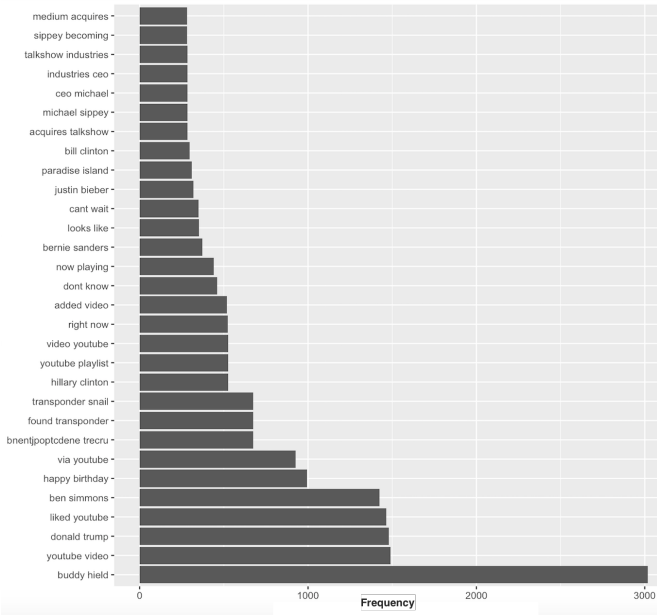

FIGURE 4—Most frequent bigrams.

According to the figure, we can confirm that the most common bigrams correspond to proper names, at least to a large extent. Due to this and to improve the association rule mining process, we will merge the most common names identified in this step, for managing them as one word instead of two. In addition, bigrams analysis provides information about the data domain and the conversations in the social network that will help to guide the information discovery process in a later stage. For instance, in the case of the US presidential election bigrams such as Bill-Clinton, Bernie-Sanders, Hillary-Clinton and Donald-Trump will guide the discovery process.

B. NAMED ENTITY RECOGNITION

Since we are focusing on obtaining sentiments or trends about influential people in the US presidential election, we have carried out an instance selection process keeping only the tweets that refer to people. This approach has been used as an example to illustrate how the model works in later steps, but the same model could be used to obtain opinions, for example, about products, brands or places. We need to recognize, therefore, the entities that are present in a tweet and this can be done using the Named Entity Recognition technique [56], from now on NER. Proposed by the University of Stanford, the method is implemented in Java, although it is integrated with several packages for R, and it obtains named-entities in a text according to the type of entity we are looking for.

This process is slow, this is why it has been parallelized and executed in a distributed processing cluster, so that the NER process which could be a bottleneck, is carried out efficiently. After executing the NER process, we obtain quite acceptable results with 140,718 tweets referring to people.

To avoid conflicts between a word written in capital letters and another occurrence of the same word in lower case letters, all the content is transformed into lower case letters. Although this transformation is one of the main steps in text

mining, in our case it is applied after the NER process, since the use of capital letters in the proper name aids and improves the results of the NER process.

At this stage, we focus on the person entity to carry out an oriented experimental process for our case of the US presidential election.

C. OBTAINING ASSOCIATION RULES

To obtain the association rules, the typical text mining corpus of tweets used so far has to be transformed into a transactional database. This structure requires a lot of memory since it is a very scattered matrix, taking into account that each item will be a word and each transaction will be a tweet. To create the transactions, we have used a binary version in which if an item appears in a transaction it is internally denoted with a 1, and if it does not appear in that transaction the matrix will have a 0.

Since our system cannot be tested against any of the proposals in the literature, because as far as we know it is the first one that deals with sentiment and pattern analysis in social media in the same way as our proposal, we have tested the system using three different pattern mining algorithms widely applied in the ambit or association rule mining. These algorithms are Apriori [12], Eclat [57] and FP-Growth [13].

D. SENTIMENT IDENTIFICATION

The sentiment analysis of our approach is based on a generalization of association rules taking into account the words associated to emotions appearing in the antecedent or the consequent of discovered association rules. To this aim, we employed generalized association rules using a hierarchy of words and sentiments. In this regard, R offers a large number of packages for sentiment analysis. In our proposal, the *syuzhet package* has been employed, which uses in its lower layer some famous sentiment dictionaries like the coreNLP proposed by Manning et al. [58] at Stanford University. This package contains very powerful dictionaries for sentiment identification. We use the dictionary NRC Word-Emotion Association Lexicon, created by Saif M. Mohammad which takes into account the 8 basic emotions (trust, anger, anticipation, disgust, fear, joy, sadness and surprise) proposed by the psychologist Plutchik [9]. Given the nature of our problem, our interest lies in obtaining the general emotion for each word.

To achieve this, we propose an iterative process in which the system obtains the feeling associated to each word in that tweet. This generates a data structure, in which for each word it can be obtained how many occurrences are for each sentiment. Finally, a majority sentiment is assigned to each word. The result of these associations of words to sentiments can be represented by a cloud of words (see example in Figure 5) where the words are classified by sentiments and represented in a certain colour.

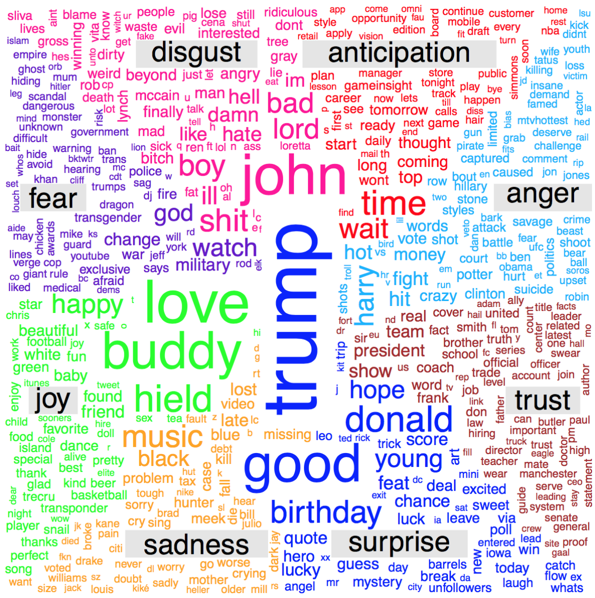

FIGURE 5—Words associated to emotions.

E. GENERALIZED ASSOCIATION RULES MINING BASED ON SENTIMENTS

The last step of the methodology is to combine the previously stages seen in IV-C and IV-D. One of the biggest differences of this study compared to previous ones is the use of sentiments to improve the process of association rules discovery. For this, we will use the sentiments associated with the terms to substitute these terms in the antecedents of the generated association rules, as long as these are not a proper name. In this way, we will obtain association rules involving people who are talked about on Twitter and their associated sentiments. This will make easier the results interpretation of a certain character. For instance, the rules of the type *{ignored,rape}*$\rightarrow$ *{donald-trump}* are transformed to *{anger}*$\rightarrow$ *{donald-trump}* because both the terms *ignored* and *rape* are associated with the anger sentiment, according to the sentiment identification stage. Some examples of how words are generalized and then rules are obtained based on these sentiments can be seen in Figure 6.

The main contribution of this study to the state-of-the-art in sentiment analysis is highlighted in this point, where we manage to summarize thousands of tweets about a character. We achieve this through the use of generalized rules by sentiments. The obtained generalized rules will be very confident and provide a new way to analyse the rules obtained in other stages based on the sentiments that the items (terms) provoke, which could even be considered closer to the human interpretation of an opinion or an emotion.

V. EXPERIMENTAL STUDY

Several experiments have been carried out with the three pattern mining algorithms used (Apriori, Eclat and FP-Growth). The implementations of the algorithms correspond to the data analysis software R. The choice of these algorithms

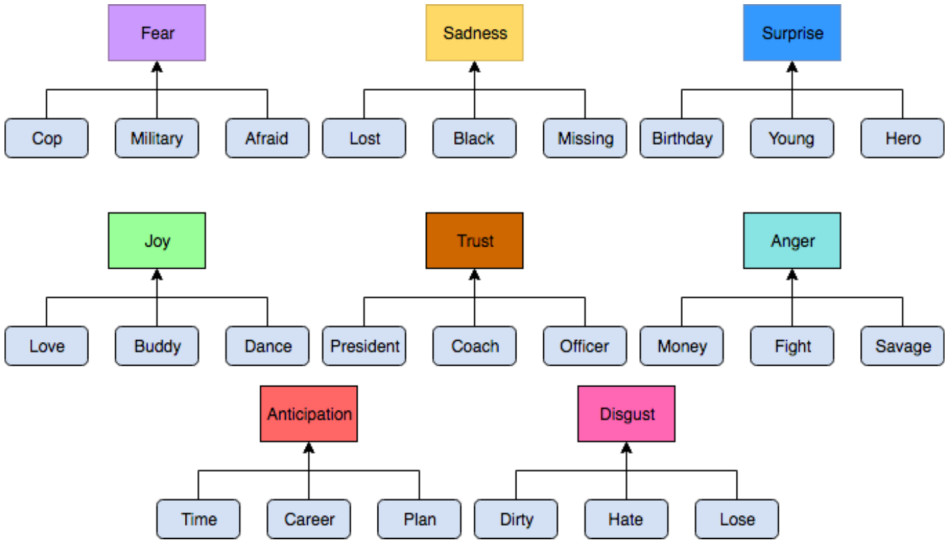

FIGURE 6—Generalization of words based on emotions.

Component	Features
CPU	2,6 GHz Intel Core i5
RAM	8 GB 1600 MHz DDR3
Hard Disk	SATA SSD de 120 GB

TABLE 2: Machine specifications.

Component	Features
CPU	Intel Xeon E5-2665
RAM	32 GB
Cores	8

TABLE 3: Cluster specifications.

corresponds to use one of each of the most extended aspects within the pattern mining, because each of them uses different data structures for the generation of rules. With them, we have carried out a comparison in terms of memory, time and number of rules obtained. Using this comparison, we try to affirm that the system is independent of the technique and obtains great results. The experiments have been carried out in terms of variation of the support value of the rules, maintaining a constant value of confidence.

For the thresholds of the algorithms we have used a confidence of 0.7, and support values of 0.01, 0.001 and 0.0001. The choice of these values is not random, they are justified because for higher threshold values very few rules were extracted. Additionally, the low support value against high confidence will offer us an acceptable and cohesive number of rules. According to the support, although the value may seem rather low, considering that the number of transactions is very high, these are values that can obtain interesting trends. For example, taking into account that we have 1.7M of transactions, a support of 0.001 will give us that there are about 1700 transactions in which related items appear, so this can be considered a trend.

According to the computer equipment used, for preprocessing, visualization and generation of transactions, it has been used the machine whose specifications can be seen in the Table 2. For the process of mining association rules, a processing cluster is used whose specifications can be studied in the Table 3.

A. DATA COLLECTION

The power of data from Twitter for mining or analysing sentiments is closely linked to the definition given by Liu et al. [59] for the concept of opinion. An opinion is then identified as a quintuple formed by entity, opinion holder, aspect, sentiment and time. If we compare this with the definition and anatomy of a tweet, in essence, we find the same elements, which can be summarized as:

- Entity: About what is the published tweet, for example, a brand.
- Opinion holder: User publishing the tweet.
- Aspect: What is valued in the tweet.
- Sentiment: We can publish a tweet of support or anger, among others.
- Time: Date and time the tweet was published.

The power of this type of data (tweets) is, therefore, verified for the process of sentiment analysis in Twitter, where extrapolating the previous definition we will try to obtain the sentiment on certain aspects of entities in an automatic way.

For the use case under study, we have collected a random sample of tweets containing the topics addressed in the social network that allow to obtain the trends in a certain period of time. In order to obtain a sample of data of such a size that could be considered as random, the native applications of Twitter were rejected for data collection, since these make use of Application Programming Interfaces (APIs) that limit the amount of data to be collected and only allow temporary connection windows. Due to these restrictions,

instead of making requests to the Twitter API, a crawler was implemented in Python that obtains, processes and stores the tweets directly from the Twitter search site. The only filter applied was to limit to tweets obtained in the US and English-speaking, between the months of January and June 2016. Since the collection of tweets has been made without applying any filter by keyword we could rely on the randomness of the sample. The final volume of the obtained sample was 1.7 million tweets.

The first difficulty we have encountered in processing the data was its volume and the need to convert each of the tweets to intermediate data structures that can be easily handled. At the end of the data collection process, we collected 1.7M of tweets divided into MongoDB[3] data collections according to the month of origin. The native connections between R and MongoDB do not allow the loading of this large volume of data, so a distributed approach based on Spark [60] was used to achieve its load. Once the data were loaded in the form of a data frame, they were integrated into a well-known structure in text mining, the corpus that facilitates the handling of texts and maintains metadata for subsequent consultation and cleaning processes.

After the pre-processing and data cleaning process (see Section IV-A), the original 1.7M tweet corpus is reduced to 140,000 tweets. The vocabulary was comprised of 7222 different terms, so due to its size we can frame it again within a Big Data problem. Finally, it should be noted that the size of the corpus is much higher than the size of the related works seen in the Table 1.

B. RESULTS

In this section we have studied the obtained results through the use of different association rule mining algorithms: Apriori, Eclat and FP-Growth. This comparison enables to choose the best algorithm to apply our procedure. In Table 4 we can find the results for the Apriori algorithm, in Table 5 the results for Eclat and in Table 6 the results for FP-Growth.

Support	***0.01***	***0.001***	***0.0001***
time	1s 378ms	1s 649ms	10s 811ms
memory (MB)	16,7	18,8	207,1
amount of rules	4	33735	2873227

TABLE 4: Results for the Apriori algorithm

Support	***0.01***	***0.001***	***0.0001***
time	50s 851ms	54s 4ms	1min 21s 849ms
memory (MB)	16,7	18,8	207,1
amount of rules	4	33735	2873227

TABLE 5: Results for the Eclat algorithm

The first obvious result, it can be observed in the comparison between Apriori and Eclat, is that both obtain the same results, in terms of association rules, since they are exhaustive

[3]NoSQL database used in Big Data problems.

Support	***0.01***	***0.001***	***0.0001***
time	45s	1min 39s	5min 12s
memory (MB)	0,0023	131,29	7.902,41
amount of rules	5	229961	13365093

TABLE 6: Results for the FP-Growth algorithm

algorithms. Regarding the time consumption, since Eclat uses lower performing data structures, it takes much more time compared to Apriori, so this leads us to discard this algorithm for our system. A more visual comparison of the time consumed by the algorithms can be seen in Figure 7.

From this graph we can also see how the FP-Growth algorithm consumes more time, as well as getting more rules (Figure 8) and therefore consumes much more memory (Figure 9). This is due to the fact that this algorithm is very efficient for very low support values so that it can obtain rules that Eclat or Apriori cannot, because these would saturate the capacities of their data structures.

An interesting comparison, is the one deduced from the Apriori algorithm and the FP-Growth. The FP-Growth (see Table 6) can obtain more rules and therefore takes longer and consumes more memory. The Apriori algorithm does not obtain the same set of rules than the FP-Growth because this latter obtain many redundant rules. That is, the same rule with the items in different positions are obtained. So the results offered by the Apriori implementation in R are more suitable for later interpretation purposes, facilitating the inspection of results.It is also important to note that Apriori can, in just few seconds, obtain almost 3 million rules. So in terms of time and number of rules, this algorithm is more interesting for social media studies.

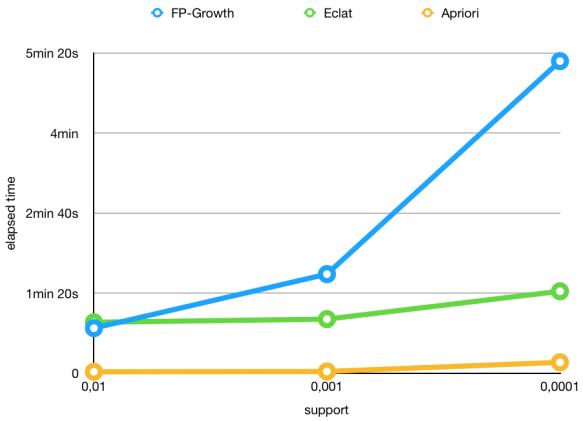

FIGURE 7—Algorithm comparison regarding the execution time

VI. USE CASE: US PRESIDENTIAL ELECTION

The final results of our study can be seen in this section, where we describe the problems and solutions we have found during the application of our system we also discuss the visualization methods to interpret the trends obtained from the rules, concluding that the proposed system helps to analyse

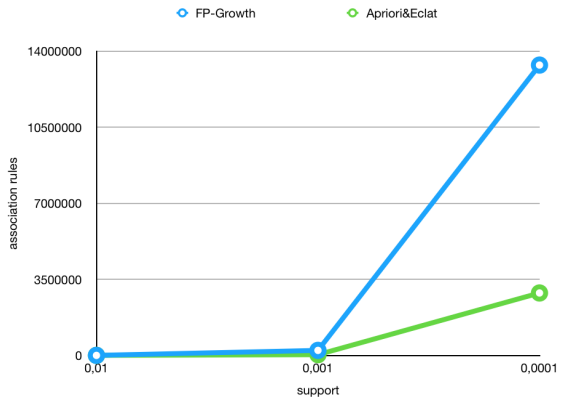

FIGURE 8—Comparison of the amount of rules discovered by the algorithms

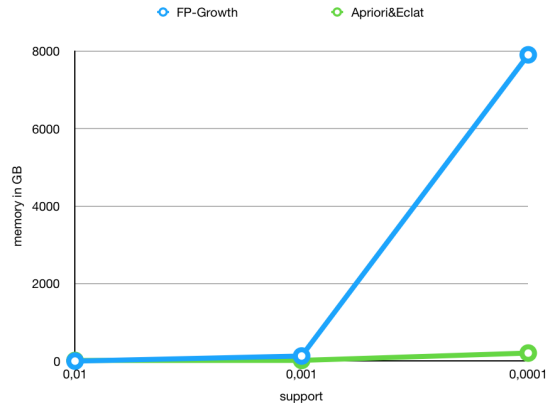

FIGURE 9—Comparison of the memory used

sentiments in microblogging texts such as Twitter. To test obtained results by the model and corroborate the utility of association rules as a descriptive method in mining trends and patterns, we will focus on two of the characters that our exploratory analysis process based on 2-grams (Figure 4) revealed: Donald Trump and Hillary Clinton. The reason for choosing a use case comparable with real-life events is due to the impossibility of measuring the system against another work of the same type. This is because the volume of tweets used in other proposals is very low and, as far as we know, there are no more studies that perform this type of pattern mining method and generalized association rules for sentiment analysis.

Given that the period of time coincides with the US election campaign, we have opted to obtain the rules generated with consequent equal to Donald-Trump or Hillary-Clinton. It would be hard to exhaustively analysed all the rules generated for proper names on Twitter during this period, so these two have been taken as an example, but it is necessary to point out that the same study could be applied to other names. However, since the chosen names belong to the political world and we know the electoral results, this will give the opportunity of corroborating the obtained results as we will explain in the forthcoming paragraphs.

Antecedent	Consequent	Supp	Conf	Lift
{military,people,transgender}	*{donald-trump}*	3.5e-04	0.71	68.79
{military,serve,transgender}	*{donald-trump}*	1.9e-04	0.79	76.48
{bans,serving, transgender}	*{donald-trump}*	8.5e-05	0.92	88.90
{ignored,rape}	*{donald-trump}*	9.9e-05	1	96.31
{child,rape}	*{donald-trump}*	9.9e-05	0.93	89.89
{caucus,lead}	*{donald-trump}*	8.5e-05	0.85	82.55

TABLE 7: Interesting rules about Donald Trump.

Afterwords, we filter the rules where the consequent is Donald-Trump or Hillary-Clinton and we focus the analysis in these two sets of rules. At the end of this process, we obtained a set of 156 rules for Donald Trump and a set of 93 rules for Hillary Clinton. Given the number of these, in the following sub-sections, we will study, visualize and interpret some of the most interesting results that our proposal obtained. For this use case we have used the Apriori agorithm with 0.0001 minimum support and 0.7 for minimum confidence thresholds.

1) Donald Trump

For Donald Trump, a set of 156 rules was found, whose distribution as a function of support, confidence, and the number of items in the rule can be seen in Figure 10. Considering this graph, we can see how the practical totality of the rules are placed on the left side, which indicates that the support values are rather low, although acceptable according to the amount of stored data. On the other hand, confidence is distributed normally and the majority of the rules are comprised of three or four items. This type of graphics is useful to see what kind of rules have been generated, but it will be necessary to study these rules manually and discern about their importance or not in the sought objective to obtain trends.

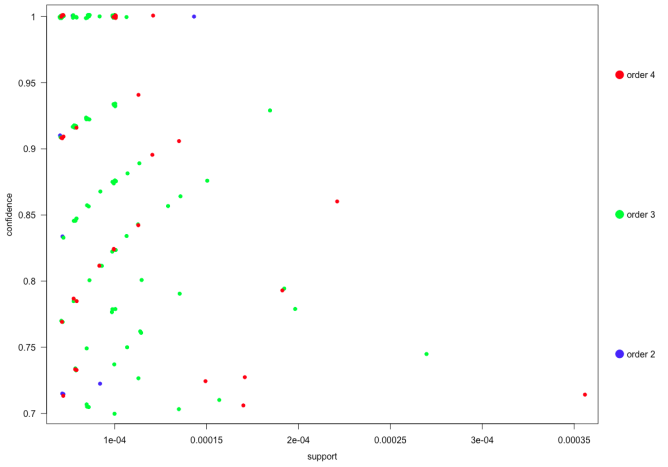

FIGURE 10—Rules distribution for Donald Trump. Support in x-axis and confidence in y-axis.

After their inspection, some interesting rules obtained about Donald Trump are shown in Table 7. Focusing on the table, the first three rules have been selected for their joint analysis, where we can see a clear pattern in terms of

Trump's policies with transgender people and their ability to serve in the United States. Specifically, the rule *{bans, serving, transgender}* → *{donald-trump}* shows that the current president was aligned with the prohibition of the service of these people in the sector, something that was already being considered in 2016 and that it was confirmed in 2017.

Another interesting pattern can be marked by the following two rules, *{ignored, rape}*→ *{donald-trump}* and *{child, rape}* → *{donald-trump}*, which indicate the scandals related to violations that Donald Trump was involved, and also the non-condemnation of these. Finally, we also find an interesting rule in *{caucus, lead}* → *{donald-trump}* that confirms the proven fact that all the polls considered this leader in voting intention in the caucus, a kind of primary election that takes place in the United States.

If we focus on the lift of the rules, we can see how their high values tell us that there is a great relation of dependence between the itemsets of the obtained rules, which leads us to affirm that the relations of these within the dataset are very strong and can be considered a trend.

Although a manual study is necessary, it can be tedious. For this reason, it is interesting to have different ways of visualization helping the user to get an idea of the generated rules, even more, if the set of them is large. Since we try to represent and obtain patterns in Twitter, in Figure 11 we have represented in the form of a word cloud, which represents the most used words in the antecedent of the rules that are consistent with our goal, in this case, Donald Trump.

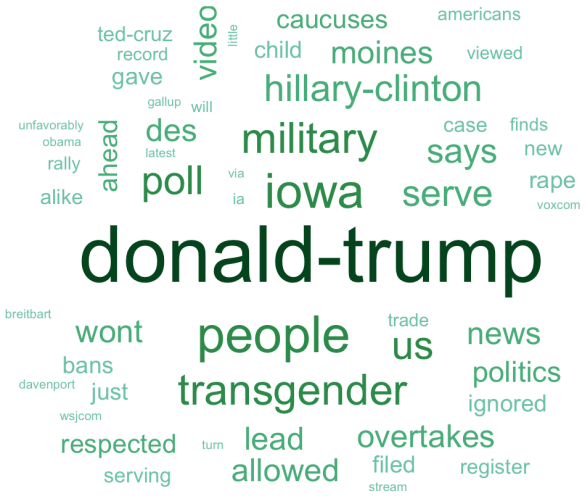

FIGURE 11—Word Cloud for Donald Trump.

If we, therefore, attend to the representation of the rules with a cloud of terms, even a person without knowledge about the topic could deduce what is being said on Twitter and what are the main tendencies in relation with the candidate. For example, we find the words *transgender, rape, child* over which we have been able to obtain trends by an inspection of rules. *Iowa* also appears as relevant, a word that we previously ignored in the manual process and, now thanks to this graphic, we see that it is interesting. If we search the rules having Iowa in the antecedent, we will see that this was a decisive and very rivalled state during the presidential elections, because the polls and the public opinion continuously generated information about that.

Antecedent	Consequent	Supp	Conf	Lift
{better,vote}	*{hillary-clinton}*	3.90e-04	0.90	246.84
{musician,squad}	*{hillary-clinton}*	3.83e-04	1	273.77
{musician,support}	*{hillary-clinton}*	3.83e-04	1	88.90
{bernie-sanders,vs}	*{hillary-clinton}*	3.83e-04	1	273.77
{bernie-sanders,better}	*{hillary-clinton}*	3.83e-04	0.94	259.36
{bernie-sanders,race}	*{hillary-clinton}*	2.20e-04	0.96	265.21
{emails,republicans}	*{hillary-clinton}*	1.56e-04	1	273.77
{attack,emails}	*{hillary-clinton}*	1.56e-04	1	273.77
{attack,republicans}	*{hillary-clinton}*	1.56e-04	0.88	240.91

TABLE 8: Interesting rules about Hillary Clinton.

2) Hillary Clinton

For Hillary Clinton, a total of 93 association rules were obtained. The distribution of them, according to their measures of goodness, can be seen in Figure 12. Looking at it, we can see how in this case the rules are placed along with the x-axis as uniform as they did in Figure 10, with some rules placed on the right which indicates good support values in them. Unlike what happened with Trump, here almost all the rules involve 3 items, having very few rules with orders different than 3.

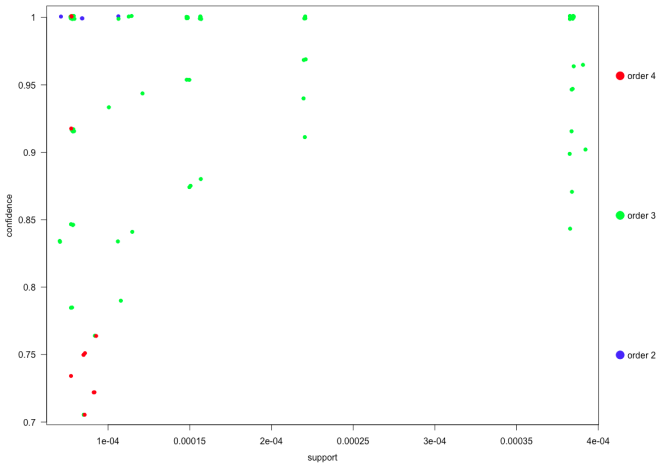

FIGURE 12—Distribution of rules for Hillary Clinton. Support in x-axis and confidence in y-axis.

Once the generated set of rules has been obtained, we can inspect them to obtain relevant information about Hillary Clinton, in the same way as what we did with Trump. Once we have analysed them, we have chosen the rules of Table 8 as the most interesting. If we perform an interpretation by groups, we could clearly define three trends and groups of patterns in the tweets related to Hillary Clinton:

1) The commitment of the show-business with her candidacy: the first three rules of the table, *{better, vote}* → *{hillary-clinton}*, *{musician, squad}* → *{hillary-clinton}* and *{musician, support}* → *{hillary-clinton}*,

refer to support received by the candidate by famous show stars that soon came out to defend her candidacy for the presidency in major public events.

2) The race against her democrat competitor Bernie Sanders: this pattern was clear before the analysis since the exploratory process unveiled Bernie Sanders as one of the most used names on Twitter in that period. The rules *{bernie-sanders, vs}* $\rightarrow$ *{hillary-clinton}*, *{bernie-sanders, better}* $\rightarrow$ *{hillary-clinton}* and *{bernie-sanders, race}* $\rightarrow$ *{hillary-clinton}*, therefore confirm the trend on Twitter to argue about which of the two candidates deserved the most the candidate position and its associated policies.
3) The scandal of the mails: the last three rules *{emails, republicans}* $\rightarrow$ *{hillary-clinton}*, *{attack, republicans}* $\rightarrow$ *{hillary-clinton}* and *{attack, emails}* $\rightarrow$ *{hillary-clinton}*, refer to the filtered mails of Hillary Clinton and their use as an attack that the republicans made of them.

Regarding the lift, again we have very strong rules that tell us that the terms appearing in the rules have a high dependence, and its descriptive and predictive power is high. Finally, we show the results using a word cloud, where we can delve easily into other trends or patterns that we might not have taken into account a priori. The graphic can be seen in Figure 13 and, in this case, we corroborate the totality of the conclusions obtained previously, like for instance the importance of the relation between the opinions of Bernie Sanders and Hillary Clinton herself. On the other hand, we see Iowa again, something that we would expect from the moment we studied the cloud of terms related to rules involving Donald Trump, since both candidates competed for the votes in that state, the related rules will be bidirectional.

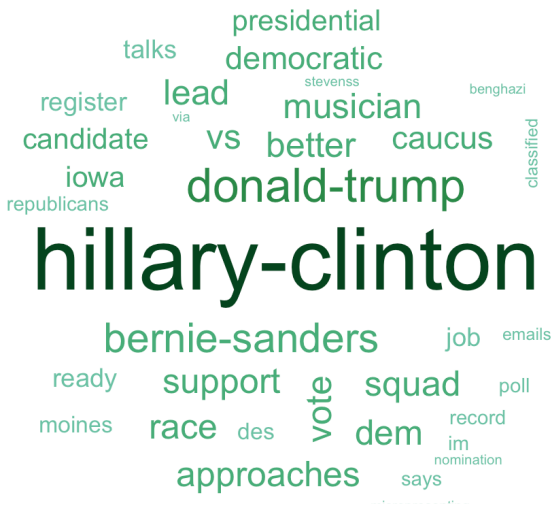

FIGURE 13—Word Cloud for Hillary Clinton.

3) Generalized approach based on sentiments

Thanks to our analysis of sentiments, we have categorized each of the words present in the domain of our problem, so based on what has been previously seen for the generalized association rules, we can organize them hierarchically according to the sentiments that each word represents.

Antedecent	Consequent	Supp	Conf	Lift
{trust}	*{donald-trump}*	0.945927	1	1
{anticipation}	*{donald-trump}*	0.594113	1	1
{surprise}	*{donald-trump}*	0.425051	1	1
{anger}	*{donald-trump}*	0.345656	1	1
{fear}	*{donald-trump}*	0.295003	1	1
{joy}	*{donald-trump}*	0.226557	1	1
{disgust}	*{donald-trump}*	0.112936	1	1
{sadness}	*{donald-trump}*	0.074606	1	1

TABLE 9: Rules based on sentiments about Donald Trump

A previous step before carrying out this phase, goes through the interpretation of the results obtained during the categorization process of the words. For this, we use the analysis shown in Figure 5, where the data have been categorized by sentiments attending to colours. There are also interesting things like relating anger with politicians, murders or piracy among other cases. A very interesting feeling which appears frequently is fear, related to the terms related to police or army. The transgender word also appears associated with this and the country's politics, news that appeared during the electoral campaign.

Restricting to our use case, the results for Donald Trump can be seen in Table 9 while the results obtained for Hillary Clinton can be seen in Table 10. The first thing that is interesting to emphasize about the obtained rules is that we have developed a ranking of the sentiments that identify each of the analysed person by employing the association rules assessment measures (i.e. support, confidence and lift). One thing that comes to light and that we could conclude is that Twitter has been issued more support tweets against both candidates than other types of sentiments because the feeling of trust is present in almost 95% of the tweets that talked about them. A very interesting discovery is the association of anger with Hilary Clinton supported by 50% of the tweets. On the contrary, Trump has 20% of tweets related to this sentiment, so it seems that American society, despite the perception in Spain, was more aligned against Hillary Clinton than Trump, something that was later confirmed with the victory of the Republican candidate. Also noteworthy is the feeling of *surprise* in Donald Trump, as he is known worldwide for his outbursts in social and political networks, so it was expected that this sentiment would have great relevance in the tweets about the current president of the United States of America.

VII. DISCUSSION

In this section, the main contributions as well as the difficulties encountered during the experimental phase are addressed. As far as we know, this is the first work that deals with pattern analysis in social media without topic filtering. For this reason, it is not possible to compare our system with

Antedecent	Consequent	Supp	Conf	Lift
{trust}	*{hillary-clinton}*	0.939688	1	1
{anger}	*{hillary-clinton}*	0.492217	1	1
{anticipation}	*{hillary-clinton}*	0.486381	1	1
{fear}	*{hillary-clinton}*	0.299610	1	1
{surprise}	*{hillary-clinton}*	0.200389	1	1
{joy}	*{hillary-clinton}*	0.145914	1	1
{sadness}	*{hillary-clinton}*	0.079766	1	1
{disgust}	*{hillary-clinton}*	0.077821	1	1

TABLE 10: Rules based on sentiments about Hillary Clinton

similar ones reviewed in the literature, because systems that filter by hashtags, clusters or topics will obviously have better performance in terms of execution, memory or accuracy. Our system has therefore gone a step further by offering a processing flow capable of working with data as it is found in social media, in a unsupervised way.

Independently of the algorithm, our system offers great results in terms of unsupervised data mining algorithms. The patterns are very descriptive and could be used, for example, by the press to obtain information about the tweets published in a specific period of time about the topic that concerns them at that time. Due to the power of the non-query based system the topic under analysis can be combined with other topics without the need to obtain or load new data. That is, a first version of a massive listening system of the social network Twitter has been proposed.

It is worth noting how the system offers descriptive results with support values that are not excessively low and in very acceptable times. In this sense, the Apriori algorithm with 0.001 obtains very relevant association rules in very short time, so that results can be latter catalogued, obtaining, for example, the patterns about all the people who have spoken in Twitter in a few months in just a few seconds about a topic. If we pay attention to the FP-Growth algorithm, it obtains more rules because, in terms of efficiency, it can explore more solutions than the Apriori without saturating the system. Finally, it is necessary to mention that the increase of rules of this algorithm, is largely due to the number of redundant rules.

VIII. CONCLUSIONS AND FUTURE WORK

In the course of elaboration of this work, the increasing interest of the application of traditional data mining techniques to new domains such as social networks is pointed out. Based on these techniques, a study of the state of the art about the application of association rules to the field of pattern mining and social media mining has been carried out. Theoretical notions about query and non-query based systems have been established, differentiating them and placing the value of non-query systems and data lakes in the field of Big Data. Also, it has been shown how the system can obtain interesting patterns from one of these data lakes without the need to filter the input, that is, using an unsupervised approach being able to obtain cross-sectional information from the content generated in social networks.

We also developed a system capable of obtaining sentiment patterns in microblogging platforms such as Twitter. These patterns could be catalogued as trends since we have seen that they are very relevant in the use case demonstrated. If we look at the obtained results we have compared obtained patterns with the events that have taken place in real life. In this way we have highlighted the great potential of the system in terms of its descriptive power.

We have been able to show that the techniques like association rules are also relevant and should be taken into account in similar studies since they provide very close to natural language interpretations in a straightforward way, even without having a priori information about the problem. Finally, it is necessary to highlight the loud and interesting information extracted by our system from the myriad of different topics addressed on Twitter.

According to the above, we have verified the power of association rules for obtaining sentiment patterns. It would, therefore, be very interesting to extend the work to a focus on the cloud so that it could be kept running in virtual machines of some cloud service provider. This would eliminate the restrictions of personal machines and allow a more detailed analysis that could make use of streaming data from Twitter, categorizing trends in real-time. A future work will be the development of a real-time procedure, based on data mining in stream flows to analyse opinions and sentiments of a certain country and region about a certain topic in real-time.

ACKNOWLEDGMENT

The research reported in this paper was partially supported by the Spanish Ministry for Economy and Competitiveness by the project (grant TIN2015-64776-C3-1-R), the COPKIT project under the European Union's Horizon 2020 research and innovation program (grant agreement No 786687), the Andalusian government under the project P18-RT-2947, Data Analysis in Medicine: from Medical Records to Big Data. Finally the project is also partially supported by the Spanish Ministry of Education, Culture and Sport (FPU18/00150) and the program of research initiation for master students of the University of Granada.

REFERENCES

[1] S. A. Moosavi and M. Jalali, "Community detection in online social networks using actions of users," in Intelligent Systems (ICIS), 2014 Iranian Conference on. IEEE, 2014, pp. 1–7.
[2] J. Serrano-Cobos, "Big data y analítica web. estudiar las corrientes y pescar en un océano de datos," El profesional de la información, vol. 23, no. 6, pp. 561–565, 2014.
[3] E. W. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decision support systems, vol. 50, no. 3, pp. 559–569, 2011.
[4] K. Kwon, Y. Jeon, C. Cho, J. Seo, I.-J. Chung, and H. Park, "Sentiment trend analysis in social web environments," in Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on. IEEE, 2017, pp. 261–268.
[5] M. d. P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. A. Rodriguez-Garcia, and R. Valencia-García, "Sentiment analysis

on tweets about diabetes: an aspect-level approach," Computational and mathematical methods in medicine, vol. 2017, 2017.
[6] B. Krawczyk, B. T. McInnes, and A. Cano, "Sentiment classification from multi-class imbalanced twitter data using binarization," in International Conference on Hybrid Artificial Intelligence Systems. Springer, 2017, pp. 26–37.
[7] S. Noferesti and M. Shamsfard, "Resource construction and evaluation for indirect opinion mining of drug reviews," PLOS ONE, vol. 10, no. 5, pp. 1–25, 2015.
[8] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "Senticnet: A publicly available semantic resource for opinion mining," in AAAI fall symposium: commonsense knowledge, vol. 10, no. 0, 2010.
[9] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," American scientist, vol. 89, no. 4, pp. 344–350, 2001.
[10] M. D. Ruiz, J. Gómez-Romero, M. Molina-Solana, J. R. Campaña, and M. J. Martín-Bautista, "Meta-association rules for mining interesting associations in multiple datasets," Applied Soft Computing, vol. 49, pp. 212–223, 2016.
[11] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in Acm sigmod record, vol. 22, no. 2. ACM, 1993, pp. 207–216.
[12] R. Agrawal, R. Srikant et al., "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. very large data bases, VLDB, vol. 1215, 1994, pp. 487–499.
[13] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in ACM sigmod record, vol. 29, no. 2. ACM, 2000, pp. 1–12.
[14] Z. P. Ogihara, M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in In 3rd Intl. Conf. on Knowledge Discovery and Data Mining. Citeseer, 1997.
[15] R. Srikant and R. Agrawal, "Mining generalized association rules," Future Generation Computer Systems, vol. 13, no. 2, pp. 161 – 180, 1997.
[16] A. Khan, U. Younis, A. S. Kundi, M. Z. Asghar, I. Ullah, N. Aslam, and I. Ahmed, "Sentiment classification of user reviews using supervised learning techniques with comparative opinion mining perspective," in Science and Information Conference. Springer, 2019, pp. 23–29.
[17] R. P. Mehta, M. A. Sanghvi, D. K. Shah, and A. Singh, "Sentiment analysis of tweets using supervised learning algorithms," in First International Conference on Sustainable Technologies for Computational Intelligence. Springer, 2020, pp. 323–338.
[18] F. Colace, M. De Santo, and L. Greco, "A probabilistic approach to tweets' sentiment classification," in 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. IEEE, 2013, pp. 37–42.
[19] F. Colace, M. De Santo, L. Greco, V. Moscato, and A. Picariello, "Probabilistic approaches for sentiment analysis: Latent dirichlet allocation for ontology building and sentiment extraction," in Sentiment Analysis and Ontology Engineering. Springer, 2016, pp. 75–91.
[20] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of twitter data during critical events through bayesian networks classifiers," Future Generation Computer Systems, vol. 106, pp. 92–104, 2020.
[21] J. Tao and X. Fang, "Toward multi-label sentiment analysis: a transfer learning based approach," Journal of Big Data, vol. 7, no. 1, pp. 1–26, 2020.
[22] A. Mohammed and R. Kora, "Deep learning approaches for arabic sentiment analysis," Social Network Analysis and Mining, vol. 9, no. 1, p. 52, 2019.
[23] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1253, 2018.
[24] A. Da'u, N. Salim, I. Rabiu, and A. Osman, "Recommendation system exploiting aspect-based opinion mining with deep learning method," Information Sciences, vol. 512, pp. 1279–1292, 2020.
[25] J. Chen, M. S. Hossain, and H. Zhang, "Analyzing the sentiment correlation between regular tweets and retweets," Social Network Analysis and Mining, vol. 10, no. 1, p. 13, 2020.
[26] P. Dey, A. Chaterjee, and S. Roy, "Influence maximization in online social network using different centrality measures as seed node of information propagation," Sādhanā, vol. 44, no. 9, p. 205, 2019.
[27] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable techniques for mining causal structures," Data Mining and Knowledge Discovery, vol. 4, no. 2-3, pp. 163–192, 2000.
[28] H. Oktay, B. J. Taylor, and D. D. Jensen, "Causal discovery in social media using quasi-experimental designs," in Proceedings of the First Workshop on Social Media Analytics. ACM, 2010, pp. 1–9.
[29] F. Erlandsson, P. Bródka, A. Borg, and H. Johnson, "Finding influential users in social media using association rule learning," Entropy, vol. 18, no. 5, p. 164, 2016.
[30] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in LREc, vol. 10, no. 2010, 2010, pp. 1320–1326.
[31] L. Abu Daher, I. Elkabani, and R. Zantout, "Identifying influential users on twitter: A case study from paris attacks," Applied Mathematics and Information Sciences, vol. 12, pp. 1021–1032, 09 2018.
[32] H. T. Phan, N. T. Nguyen, and D. Hwang, "A tweet summarization method based on maximal association rules," in International Conference on Computational Collective Intelligence. Springer, 2018, pp. 373–382.
[33] A. Belhadi, Y. Djenouri, J. C.-W. Lin, C. Zhang, and A. Cano, "Exploring pattern mining algorithms for hashtag retrieval problem," IEEE Access, vol. 8, pp. 10 569–10 583, 2020.
[34] Z. Zainol, S. Wani, P. N. Nohuddin, W. M. Noormanshah, and S. Marzukhi, "Association analysis of cyberbullying on social media using apriori algorithm," International Journal of Engineering & Technology, vol. 7, no. 4.29, pp. 72–75, 2018.
[35] H. Margono, X. Yi, and G. K. Raikundalia, "Mining indonesian cyber bullying patterns in social networks," in Proceedings of the Thirty-Seventh Australasian Computer Science Conference-Volume 147. Australian Computer Society, Inc., 2014, pp. 115–124.
[36] A. A. Hamed, X. Wu, and A. Rubin, "A twitter recruitment intelligent system: association rule mining for smoking cessation," Social Network Analysis and Mining, vol. 4, no. 1, p. 212, 2014.
[37] R. C. Mosley Jr, "Social media analytics: Data mining applied to insurance twitter posts," in Casualty Actuarial Society E-Forum, vol. 2. Citeseer, 2012, p. 1.
[38] M. Adedoyin-Olowe, M. M. Gaber, C. M. Dancausa, F. Stahl, and J. B. Gomes, "A rule dynamics approach to event detection in twitter with its application to sports and politics," Expert Systems with Applications, vol. 55, pp. 351–360, 2016.
[39] C. Fernandez-Basso, A. J. Francisco-Agra, M. J. Martin-Bautista, and M. D. Ruiz, "Finding tendencies in streaming data using big data frequent itemset mining," Knowledge-Based Systems, vol. 163, pp. 666–674, 2019.
[40] V. Kakulapati and S. M. Reddy, "Mining social networks: Tollywood reviews for analyzing upc by using big data framework," in Smart Innovations in Communication and Computational Sciences. Springer, 2019, pp. 323–334.
[41] X. Zhou, X. Tao, J. Yong, and Z. Yang, "Sentiment analysis on tweets for social events," in Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2013, pp. 557–562.
[42] S. Das, A. Dutta, G. Medina, L. Minjares-Kyle, and Z. Elgart, "Extracting patterns from twitter to promote biking," IATSS Research, vol. 43, no. 1, pp. 51–59, 2019.
[43] Z. Hai, K. Chang, and J.-j. Kim, "Implicit feature identification via co-occurrence association rule mining," in International Conference on Intelligent Text Processing and Computational Linguistics. Springer, 2011, pp. 393–404.
[44] M. Yuan, Y. Ouyang, Z. Xiong, and H. Sheng, "Sentiment classification of web review using association rules," in International Conference on Online Communities and Social Computing. Springer, 2013, pp. 442–450.
[45] R. Dehkharghani, H. Mercan, A. Javeed, and Y. Saygin, "Sentimental causal rule discovery from twitter," Expert Systems with Applications, vol. 41, no. 10, pp. 4950–4958, 2014.
[46] N. Mamgain, B. Pant, and A. Mittal, "Categorical data analysis and pattern mining of top colleges in india by using twitter data," in 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2016, pp. 341–345.
[47] L. Bing, K. C. Chan, and C. Ou, "Public sentiment analysis in twitter data for prediction of a company's stock price movements," in 2014 IEEE 11th International Conference on e-Business Engineering. IEEE, 2014, pp. 232–239.
[48] M. Hahsler and R. Karpienko, "Visualizing association rules in hierarchical groups," Journal of Business Economics, vol. 87, no. 3, pp. 317–335, 2017.
[49] L. Cagliero and A. Fiori, "Analyzing twitter user behaviors and topic trends by exploiting dynamic rules," in Behavior Computing. Springer, 2012, pp. 267–287.

[50] ——, "Discovering generalized association rules from twitter," Intelligent Data Analysis, vol. 17, 01 2013.
[51] J. A. Diaz-Garcia, M. D. Ruiz, and M. J. Martin-Bautista, "Generalized association rules for sentiment analysis in twitter," in International Conference on Flexible Query Answering Systems. Springer, 2019, pp. 166–175.
[52] V. V. Bochkarev, A. V. Shevlyakova, and V. D. Solovyev, "The average word length dynamics as an indicator of cultural changes in society," Social Evolution & History, vol. 14, no. 2, pp. 153–175, 2015.
[53] M. Damashek, "Gauging similarity with n-grams: Language-independent categorization of text," Science, vol. 267, no. 5199, pp. 843–848, 1995.
[54] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE, 2007, pp. 697–702.
[55] K. Hornik, C. Buchta, and A. Zeileis, "Open-source machine learning: R meets Weka," Computational Statistics, vol. 24, no. 2, pp. 225–232, 2009.
[56] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005, pp. 363–370.
[57] C. Borgelt, "Efficient implementations of apriori and eclat," in FIMI'03: Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations, 2003.
[58] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System demonstrations, 2014, pp. 55–60.
[59] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data. Springer, 2012, pp. 415–463.
[60] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin et al., "Apache spark: a unified engine for big data processing," Communications of the ACM, vol. 59, no. 11, pp. 56–65, 2016.

J. ANGEL DIAZ-GARCIA is a predoctoral fellow in the Department of Computer Science and Artificial Intelligence at the University of Granada. He obtained his degree in Computer Engineering at the University of Granada in 2016. He then obtains the degree of Master in Computer Engineering in 2017, and continues his studies in the Master of Data Science, which concludes in 2019 with the obtention of his second Master's degree. He works in the research group of Databases and Intelligent Information Systems, collaborating in projects such as COPKIT, in the topic of text mining, Big Data and social media mining. Currently, he is in his first year of studies to obtain a Phd degree, in the field of Data Mining.

M. DOLORES RUIZ received her Mathematics degree in 2005 and the European Ph.D. degree in computer science in 2010, both from the University of Granada, Spain.

She held non-permanent teaching positions with the Universities of Jaen, Granada and Cadiz. She is currently with the Department of Statistics and O.R., University of Granada (Spain). She has organized several special sessions about Data Mining in International conferences and was part of the organization committee of the FQAS'2013 and SUM'2017 conference. She has participated in more than 10 projects including the FP7 projects ePOOLICE and Energy IN TIME. Her research interests include data mining, information retrieval, energy efficiency, big data, correlation statistical measures, sentence quantification, and fuzzy sets theory.

Dr. Ruiz belongs to the Approximate Reasoning and AI research group and the security Lab at the University of Granada. She has been the principal investigator of the project "Exception and Anomaly detection by means of Fuzzy Rules using the RL-theory. Application to Fraud Detection".

MARIA J. MARTIN-BAUTISTA is a Full Professor at the Department of Computer Science and Artificial Intelligence at the University of Granada, Spain. She is a member of the IDBIS (Intelligent Data Bases and Information Systems) research group. Her current research interests include Big Data Analytics in Data, Text, Web Mining, Social Mining Intelligent Information Systems, Knowledge Representation and Uncertainty. She has supervised several Ph. D. Thesis and published more than 100 papers in high impact international journals and conferences. She has participated in more than 20 R+D projects, including several European Projects and has supervised several research technology transfers with companies. She has served as a program committee member for several international conferences.

2.2 Generalized association rules for sentiment analysis in twitter

- Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2019). Generalized association rules for sentiment analysis in twitter. In Flexible Query Answering Systems: 13th International Conference, FQAS 2019, Amantea, Italy, July 2–5, 2019, Proceedings 13 (pp. 166-175). Springer International Publishing.
  - Conference: Flexible Query Answering Systems: 13th International Conference, FQAS 2019, Amantea, Italy, July 2–5, 2019, Proceedings 13.
  - Status: Published.

Generalized Association Rules for Sentiment Analysis in Twitter

J. Angel Diaz-Garcia[b] and **M. Dolores Ruiz**[c] and **Maria J. Martin-Bautista**[d]
[a]Department of Computer Science and A.I., University Of Granada, Daniel Saucedo Aranda, s/n, 18014 Granada
[b] joseangeldiazg@ugr.es
[c] mdruiz@decsai.ugr.es
[d] mbautis@decsai.ugr.es

No Institute Given

Abstract. Association rules have been widely applied in a variety of fields over the last few years, given their potential for descriptive problems. One of the areas where the association rules have been most prominent in recent years is social media mining. In this paper, we propose the use of association rules and a novel generalization of these based on emotions to analyze data from the social network Twitter. With this, it is possible to summarize a great set of tweets in rules based on 8 basic emotions. These rules can be used to categorize the feelings of the social network according to, for example, a specific character.

Keywords: association rules, sentiment analysis, social media mining, generalized association rules

1 Introduction

Social media mining is defined as a branch of data mining that encompasses all those techniques that are used to extract valuable knowledge from a social network. It uses techniques such as text mining, natural language processing, unsupervised and supervised learning. Recently, social media mining has taken a great relevance in the current world, since it allows automatic systems to obtain information about products, brands, services or people that can be transformed into competitive advantages. Among the most used techniques in social media mining we can find, association rules and sentiment analysis. These techniques have been analysed in numerous studies where the value of association rules to summarize and discover knowledge from large data sets has been verified [14], also as the great importance of sentiment analysis [11],[15] for completing subjective analysis of problems or domains where these techniques are applied.

In this paper we propose a social media mining system based on generalized association rules capable of summarizing a large set of tweets into a reduced set of rules. This set of rules can serve, among other things, to categorize the

feelings associated with a certain character, place or product at a certain time. The novelty of the system, is in a mixed approach that uses association rules and sentiments analysis. The fusion of both techniques is carried out through the generalized association rules. To this end, the terms associated to each association rule are swapped for its associated feeling, allowing to obtain strong and cohesive rules.

To validate the correct functioning of the proposed system, two well-known US politicians, Bernie Sanders and Hillary Clinton, have been chosen. Choosing these characters, among all of the people that the system discovered as relevant in the social network Twitter, is that we can contrast the results, according to the events that occurred during the US electoral campaing.

The paper is structured as follows: Section 2 reviews some of the related theoretical concepts that allow us to understand perfectly the following sections and also describes the related works . Section 3 explains our proposal and finally in Section 4 we explain the experimentation carried out. The paper concludes with an analysis of the proposed approach and the future lines that this work opens.

2 Preliminar Concepts & Related Work

In this section, we will see the basic concepts related to association rules and generalized association rules. We will also study previous related works following the scope of generalized association rules and the use of association rules with sentiment analysis in the field of social media mining.

2.1 Association Rules

Association rules belong to Automatic Learning and Data Mining fields. One of the first references to them dates back to 1993 [1]. They are used to obtain relevant knowledge from large databases and their representation is given by the form $X \rightarrow Y$ where X is an itemset that represents the antecedent and Y an item or itemset called consequent. As a result, we can conclude that consequent items have a co-occurrence relationship with antecedent items. Therefore, association rules can be used as a method for extracting apparently hidden relationships. The classical way of measuring the goodness of association rules is with two measures: support and confidence. In the following definitions we will see how these measures can be defined.

Definition 1. *Support of a itemset is represented as supp (X), and is the proportion of transactions containing item X out of the total amount of transactions of the dataset (D). The equation to define the support of an itemset is:*

$$supp(X) = \frac{||t \in D : X \subseteq t||}{|D|} \tag{1}$$

Definition 2. *Support of an association rule is represented as supp($X \rightarrow Y$), consequently, is the total amount of transactions containing both items X and Y, as defined in the following equation:*

$$supp(X \rightarrow Y) = supp(X \cup Y) \tag{2}$$

Definition 3. *Confidence is represented as conf ($X \rightarrow Y$) and represents the proportion of transactions containing item Y out of the transactions containing item X. The equation is:*

$$conf(X \rightarrow Y) = \frac{supp(X \rightarrow Y)}{supp(X)} \tag{3}$$

If we focus on how to obtain the rules, they can be approached from two perspectives, brute force solution (prohibitive) or from a two-step approach. The first of these stages is the generation of frequent itemsets, from which, in the second stage, the association rules are obtained. It is in this last approach where we find the most famous algorithms for mining association rules. Among these, the most famous is the Apriori proposed by Agrawal and Srikant [2], an exhaustive algorithm (gets all the rules), compared to for example the FP-Growth algorithm proposed by Han et al. [10] a very fast and appropriate for Big Data problems but not exhaustive.

2.2 Generalized Association Rules

Association rules can be interpreted and studied in different ways. One of the multiple, along with the analysis of negative association rules [18] or association rules with absent items [6], is the use of generalized association rules. This technique was introduced by Srikant and Agrawal [17] in 1995. Also knows as multilevel association rules, they propose that the rule $\{Strawberries,\ Oranges\} \rightarrow \{Milk\}$ could be replaced by $\{Fruit\} \rightarrow \{Milk\}$. This hierarchical point of view allows a higher level of abstraction that offers us the possibility of obtaining even more information from our data. They also allow us to summarize the data in a very important way, which for Big Data environments can be of vital importance. Finally, it should be noted that the rules obtained with this interpretation tend to be very strong.

2.3 Related Work

Since they were defined by Agrawal, association rules have been widely used in various problems. One of the main studies in the field of association rules is the one proposed in 2000 by Silverstein et al. [16], where they are used for the well-known problem of shopping baskets. If we look at the problem of social media mining, we can find studies such as the one proposed by Cagliero and Fiori [4] or the one proposed by Erlandsson et al. [7]. In the first study, the authors use dynamic association rules where confidence and support measures change over

time, in order to obtain data on user habits and behaviours on Twitter. In the latter, an analysis based on association rules to find influencers on Twitter is put forward.

The use of association rules in conjunction with sentiment analysis, has been less studied, so it represents an incipient problem in which we can find some studies such as Hai et al. [9] where an approach based on association rules, co-occurrence of words and clustering is applied to obtain the most common characteristics regarding certain groups of words that represent an opinion and its polarization.

If we focus on generalized association rules, they have been used in diverse approaches. For example, to improve visualization of the rules, summarizing the set of rules to be visualized [8] and have also been used in other applications such as a data mining application applied to library recommendations [12]. It is necessary to mention, their most famous use, to obtain stronger rules and better interpretation in shopping baskets problems [3]. As far as we know, the only application of generalized association rules to the field of social media mining is the work of Cagliero and Fiori [5]. In this paper, the authors propose obtaining generalized association rules through a taxonomy created by twitter topics and contexts. This work is related to ours, but in ours we use the emotions related to each word as taxonomy when extracting generalized association rules. This, offers us a powerful way to apply sentiment analysis to the Twitter environment, instead of a summary of them as the previous work does.

According to the foregoing, our work differs from all the others in that it proposes a novel mixed technique that combines the use of generalized association rules, the sentiment analysis and applies it to the field of social media mining, more specifically the microblogging platform, Twitter, although it could be extended to other fields.

3 Our Proposal

In this section we will see in detail the proposal introduced in the previous sections. We can find a graphic summary of our proposal in the Figure 1.

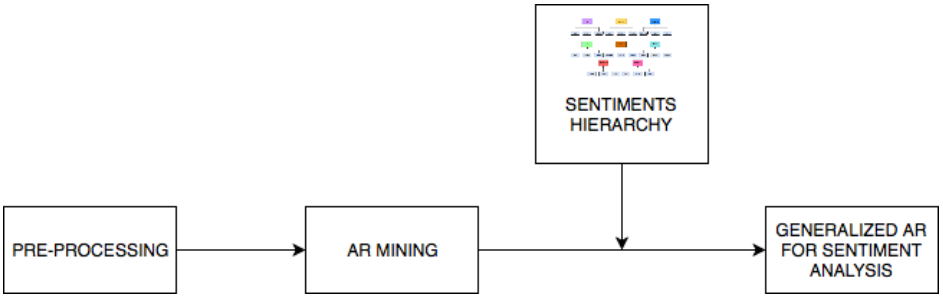

Fig. 1. Metolodogy followed.

3.1 Pre-processing

The data obtained from Twitter are very noisy so it is necessary a pre-processing step before working with them. The techniques used have been:

- Elimination of empty words in English.
- Removal of links, removal of punctuation marks, non-alphanumeric characters, and missing values (empty tweets).
- Identification and removal of unusual terms.
- Named-Entity Recognition to get those tuits that talk about people.
- Content transformation to lower case letters.
- Union of compound names.

At this point, we have a set of clean tuits on which we can apply the association rules mining techniques.

3.2 Generalized Association Rules for Sentiment Analysis

The first step of our proposal is based on obtaining the feelings associated with each word present in the data set. For this, we used the dictionary with the same name of the package, Syuzhet, created by the Nebraska Literature Laboratory. This approach takes into account the 8 basic emotions proposed by the psychologist Plutchik [13]. These emotions are *trust, anger, anticipation, disgust, fear, joy, sadness* and *surprise*. We can find an example of the emotions associated with certain words in the Figure 2.

The next step is to obtain association rules about our data set. To do this, we have use the Apriori algorithm with minimum support threshold of 0.001 and minimum confidence threshold of 0.7. The last step of the proposal is to combine the last two steps. For this, we will use the sentiments associated to the terms to substitute these terms (some examples can be see in the Figure 3) in the antecedents of the generated association rules, as long as these are not a proper name. To choose the sentiment, a majority vote is used. Finally, we will obtain association rules involving people who are talked about on Twitter and their associated sentiments.

At this point we find the major contribution of this article to the state-of-the-art of using association rules for social media mining, demonstrating that generalized association rules by feelings, generates strong rules that can categorize, for example, a character. These affirmations will be demonstrated with two real examples in the process of experimentation.

4 Experimentation

The purpose of the experimentation is to categorize by means of association rules generalized by feelings, a certain character. On the one hand, this categorization will be the reflection of the feelings that this character raises in the social network. On the other hand, this will allow us to contrast the result obtained by

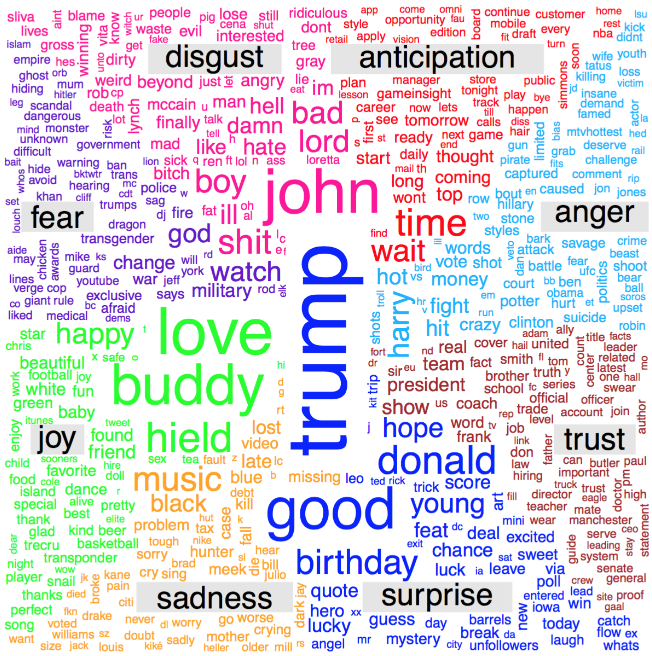

Fig. 2. Words associated to emotions.

our proposal. For this, we should have a sample large enough to be determined randomly.

The restrictions of the Twitter API do not allow to get a large number of tweets, which is why a crawler was developed that obtains English-speaking data, tweet in the United States during the first 6 months of 2016. The final data set is comprised of 1.7M tweets in total, and 140000 after the NER process. It is on these 140000 tuits that we will focus. For the general data mining processes, the computer whose specifications can be seen in the Table 1 has been used. For the NER process, a processing cluster was used whose specifications can be seen in the Table 2.

Component	Features
CPU	2,6 GHz Intel Core i5
RAM	8 GB 1600 MHz DDR3
Hard Disk	SATA SSD de 120 GB

Table 1. Computer specifications.

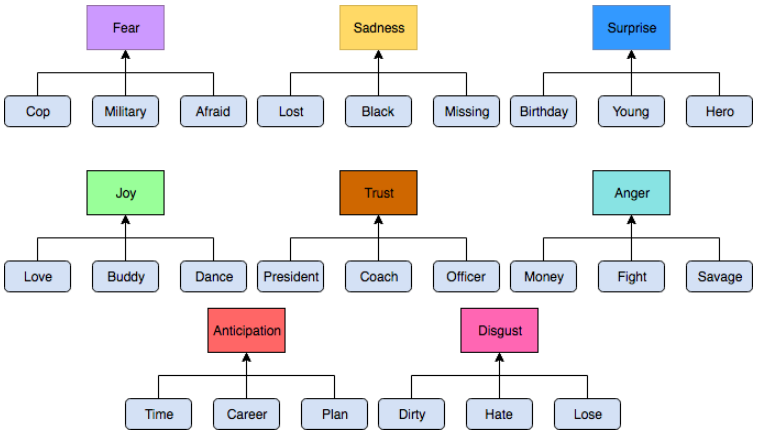

Fig. 3. Example of Generalization of words based on emotions.

Component	Features
CPU	Intel Xeon E5-2665
RAM	32 GB
Cores	8

Table 2. Cluster specifications.

4.1 Results

After applying the Apriori algorithm to our dataset, we get 34.119 rules with minimum support threshold of 0.001. Once these rules have been obtained, the terms have been swapped for their associated emotion as seen in Section 3.2. In this way, we obtain for each proper name present in the dataset a set of 8 very cohesive and strong rules that can be contrasted with reality. In this case we demostrate the system with the candidates of the primary elections of the Democratic Party of the United States in the last elections.

Use case: Hillary Clinton and Bernie Sanders After generalizing the rules, the results the system found for Hillary Clinton are those that can be seen in the Table 3. On the other hand, for Bernie Sanders we can see them in the Table4. Thus, the system has obtained an ordered list of the emotions that a certain character awakens in Twitter. According to emotions, there are interesting cases such as the emotion of fear, placed in the same position according to their values of support in both politicians. This shows that the American society, tweeted words of fear according to these two candidates equally, which it can be translated into a fear sentiment towards the Democratic Party, something that was contrasted with the victory of the Republican Party. In this way of interpretation also comes the similarity of the emotion anger. In this last emotion, Hillary

Clinton stands out. One interpretation that can be shelled is that there were many angry tweets about this character, knowing that he would be the candidate of the Democratic Party and that most Americans preferred a change of party. Again, this was corroborated by Donald Trump's victory.

Antedecent	**Consequent**	**Supp**	**Conf**
{*trust*}	=>{*hillary-clinton*}	0.93968872	1
{*anger*}	=>{*hillary-clinton*}	0.49221790	1
{*anticipation*}	=>{*hillary-clinton*}	0.48638132	1
{*fear*}	=>{*hillary-clinton*}	0.29961089	1
{*surprise*}	=>{*hillary-clinton*}	0.20038911	1
{*joy*}	=>{*hillary-clinton*}	0.14591440	1
{*sadness*}	=>{*hillary-clinton*}	0.07976654	1
{*disgust*}	=>{*hillary-clinton*}	0.07782101	1

Table 3. Rules based on feelings about Hillary Clinton.

Antedecent	**Consequent**	**Supp**	**Conf**
{*trust*}	=>{*bernie-sanders*}	0.97297297	1
{*anticipation*}	=>{*bernie-sanders*}	0.52432432	1
{*anger*}	=>{*bernie-sanders*}	0.47027027	1
{*fear*}	=>{*bernie-sandersn*}	0.22162162	1
{*joy*}	=>{*bernie-sanders*}	0.21351351	1
{*surprise*}	=>{*bernie-sandersn*}	0.19459459	1
{*disgust*}	=>{*bernie-sanders*}	0.09459459	1
{*sadness*}	=>{*bernie-sanders*}	0.08378378	1

Table 4. Rules based on feelings about Bernie Sanders.

5 Conclusions and Future Work

The proposed system based on Generalized Association Rules has shown with a concrete case, that it can be used to analyze the social network Twitter. The power of association rules for problems of the social media mining type has been demonstrated. For the Twitter data, it should be mentioned that it is complicated its treatment due to the noise they offer. If we look at future avenues of work, it would be interesting to apply Fuzzy Association Rules in the mining process and compare with the work done in this study. Also, it would be interesting to contrast the application with other characters from the social network.

Acknowledgment

This research paper is part of the COPKIT project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Acm sigmod record. vol. 22, pp. 207–216. ACM (1993)
2. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB. vol. 1215, pp. 487–499 (1994)
3. Boztuğ, Y., Reutterer, T.: A combined approach for segment-specific market basket analysis. European Journal of Operational Research **187**(1), 294–312 (2008)
4. Cagliero, L., Fiori, A.: Analyzing twitter user behaviors and topic trends by exploiting dynamic rules. In: Behavior Computing, pp. 267–287. Springer (2012)
5. Cagliero, L., Fiori, A.: Discovering generalized association rules from twitter. Intelligent Data Analysis **17**(4), 627–648 (2013)
6. Delgado, M., Ruiz, M.D., Sanchez, D., Serrano, J.M.: A fuzzy rule mining approach involving absent items. In: Proceedings of the 7th Conference of the European Society for Fuzzy Logic and Technology. pp. 275–282. Atlantis Press (2011)
7. Erlandsson, F., Bródka, P., Borg, A., Johnson, H.: Finding influential users in social media using association rule learning. Entropy **18**(5), 164 (2016)
8. Hahsler, M., Karpienko, R.: Visualizing association rules in hierarchical groups. Journal of Business Economics **87**(3), 317–335 (2017)
9. Hai, Z., Chang, K., Kim, J.j.: Implicit feature identification via co-occurrence association rule mining. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 393–404. Springer (2011)
10. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: ACM sigmod record. vol. 29, pp. 1–12. ACM (2000)
11. Kwon, K., Jeon, Y., Cho, C., Seo, J., Chung, I.J., Park, H.: Sentiment trend analysis in social web environments. In: Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on. pp. 261–268. IEEE (2017)
12. Michail, A.: Data mining library reuse patterns using generalized association rules. In: Proceedings of the 22nd international conference on Software engineering. pp. 167–176. ACM (2000)
13. Plutchik, R.: The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American scientist **89**(4), 344–350 (2001)
14. Ruiz, M.D., Gómez-Romero, J., Molina-Solana, M., Campaña, J.R., Martín-Bautista, M.J.: Meta-association rules for mining interesting associations in multiple datasets. Applied Soft Computing **49**, 212–223 (2016)
15. Salas-Zárate, M.d.P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M.A., Valencia-García, R.: Sentiment analysis on tweets about diabetes: an aspect-level approach. Computational and mathematical methods in medicine **2017** (2017)
16. Silverstein, C., Brin, S., Motwani, R., Ullman, J.: Scalable techniques for mining causal structures. Data Mining and Knowledge Discovery **4**(2-3), 163–192 (2000)

17. Srikant, R., Agrawal, R.: Mining generalized association rules. Future generation computer systems **13**(2-3), 161–180 (1997)
18. Yuan, X., Buckles, B.P., Yuan, Z., Zhang, J.: Mining negative association rules. In: Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on. pp. 623–628. IEEE (2002)

2.3 Mining text patterns over fake and real tweets

- Díaz-García, J. A., Fernandez-Basso, C., Ruiz, M. D., Martin-Bautista, M. J. (2020, June). Mining text patterns over fake and real tweets. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 648-660). Cham: Springer International Publishing.
  - Conference: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems.
  - Status: Published.

Mining text patterns over fake and real tweets

Jose A. Diaz-Garcia[a,b][0000−0002−9263−1402] **Carlos Fernandez-Basso**[a,c][0000−0002−8809−8676] and **M. Dolores Ruiz**[d][0000−0003−1077−3173] and **Maria J. Martin-Bautista**[a,e][0000−0002−6973−477X]

[a]Department of Computer Science and A.I., University of Granada
[d]Department of Statistics and O.R., University of Granada
[b] joseangeldiazg@ugr.es
[c] cjferba@decsai.ugr.es
[d] mariloruiz@ugr.es
[e] mbautis@decsai.ugr.es

Abstract. With the exponential growth of users and user-generated content present on online social networks, fake news and its detection have become a major problem. Through these, smear campaigns can be generated, aimed for example at trying to change the political orientation of some people. Twitter has become one of the main spreaders of fake news in the network. Therefore, in this paper, we present a solution based on Text Mining that tries to find which text patterns are related to tweets that refer to fake news and which patterns in the tweets are related to true news. To test and validate the results, the system faces a pre-labelled dataset of fake and real tweets during the U.S. presidential election in 2016. In terms of results interesting patterns are obtained that relate the size and subtle changes of the real news to create fake news. Finally, different ways to visualize the results are provided.

Keywords: Association Rules, social media mining, fake news, Text Mining, Twitter

1 Introduction

With the rise of social networks and the ease with which users can generate content, publish it and share it around the world, it was only a matter of time before accounts and people would appear to generate and share fake news. These fake news can be a real problem as it usually includes content that can go viral and be taken as true by a large number of people. In this way, political orientations, confidence in products and services, etc. can be conditioned. The textual nature of these news, has made it perfectly approachable by Data Mining techniques such as Text Mining, a sub area of Data Mining that tries to obtain relevant information from unstructured texts.

Because of the potential of these techniques in similar problems, in this paper we address the analysis of tweets that deal with fake content and real content by

using text mining by means of association rules. With this, we intend to prove that through these techniques relevant information can be obtained that can be used for the detection of patterns related to fake news. The contribution to the state of the art of paper is twofold:

- A reusable workflow that can get patterns on fake and real news, that can be the input of a posterior classification algorithm in order to discern between both types of news.
- A comprehensive analysis of patterns related to fake and real news during the 2016 US presidential election campaign.

In order to test and validate the system a tweet dataset has been used in which the tweets have been previously labelled as fake and real. The dataset [4] corresponds to tweets from the 2016 presidential elections in the United States. On this dataset, very interesting conclusions and patterns have been drawn, such as the tendency of fake news to slightly change real news to make it appear real. Different visualization methods are also offered to allow a better analysis of the patterns obtained.

The paper is structured as follows: Section 2 reviews some of the related theoretical concepts that allow to understand the following sections. Section 3 describes the related work. Section 4 explains the methodology followed. Finally Section 5 includes the experimentation carried out. The paper concludes with an analysis of the proposed approach and the future lines that this work opens.

2 Preliminar Concepts

In this section we will see the theoretical background of the Data Mining techniques that will be mentioned throughout the paper and that were used for the experimental development.

2.1 Association Rules

Association rules belong to the Data Mining field and have been used and studied for a long time. One of the first references to them dates back to 1993 [1]. They are used to obtain relevant knowledge from large transactional databases. A transactional database could be for example, a shopping basket database, where the items would be the products, or a text database, as in our case, where the items are the words. In a more formal way, let t=$\{A,B,C\}$ be a transaction of three items *(A, B and C)*, and any combination of them forms an itemset. Examples of direrents itemsets are $\{A,B,C\}$, $\{A,B\}$, $\{B,C\}$, $\{A,C\}$, $\{A\}$,$\{A\}$, $\{B\}$ *and* $\{C\}$. According to this, an association rule would be represented in the form $X \rightarrow Y$ where X is an itemset that represents the antecedent and Y an itemset called consequent. As a result, we can conclude that consequent items have a co-occurrence relationship with antecedent items. Therefore, association rules can be used as a method of extracting hidden relationships between items or

elements within transactional databases, data warehouses or other types of data storage from which it is interesting to extract information to help in decision-making processes. The classical way of measuring the goodness of association rules regarding a given problem is with two measures: support and confidence. To these metrics, new metrics have been added over time, among which the certainty factor [5] stands out, which we have used in our experimental process and we will define together with the support and confidence in the following lines.

- Support of an itemset. It is represented as $supp(X)$, and is the proportion of transactions containing item X out of the total amount of transactions of the dataset (D). The equation to define the support of an itemset is:

$$supp(X) = \frac{|t \in D : X \subseteq t|}{|D|} \tag{1}$$

- Support of an association rule. It is represented as $supp(X \rightarrow Y)$, is the total amount of transactions containing both items X and Y, as defined in the following equation:

$$supp(X \rightarrow Y) = supp(X \cup Y) \tag{2}$$

- Confidence of an association rule. It is represented as $conf(X \rightarrow Y)$ and represents the proportion of transactions containing item X which also contains Y. The equation is:

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{3}$$

- Certainty factor. It is used to represent uncertainty in rule-based expert systems. It has been shown to be one of the best models for measuring the fit of rules. Represented as $CF(X \rightarrow Y)$, a positive CF measures the decrease of probability that Y is not in a transaction when X appears. If we have a negative CF, the interpretation will be analogous. It can be represented mathematically as follows:

$$CF(X \rightarrow Y) = \begin{cases} \dfrac{conf(X \rightarrow Y) - supp(Y)}{1 - supp(Y)} & \text{if } conf(X \rightarrow Y) > supp(Y) \\ \dfrac{conf(X \rightarrow Y) - supp(Y)}{supp(Y)} & \text{if } conf(X \rightarrow Y) < supp(Y) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The most widespread approach to obtain association rules is based on two stages using the downward-closure property. The first of these stages is the generation of frequent itemsets. To be considered frequent the itemset have to exceed the minimum support threshold. In the second stage the association rules are

obtained using the minimum confidence threshold. In our approach, we will employ the certainty factor to extract more accurate association rules due to the goo properties of this assessment measure (see for instance [9]). Within this category we find the majority of the algorithms for obtaining association rules, such as Apriori, proposed by Agrawal and Srikant [2] and FP-Growth proposed by Han et al. [10]. Although these are the most widespread approaches, there are other frequent itemset extraction techniques such as vertical mining or pattern growth.

2.2 Association Rules and Text Mining

Since association rules demonstrated their great potential to obtain hidden co-occurrence relationships within transactional databases, they have been increasingly applied in different fields. One of the fields is Text Mining [14]. In this field, text entities (paragraphs, tweets,...) are handled as a transaction in which each of the words is an item. In this way, we can obtain relationships and metrics about co-occurrences in large text databases. Technically, we could define a text transaction as:

Definition 1. *Text transaction: Let W be a set of words (items in our context). A text transaction is defined as a subset of words, i.e. a word will be present or not in a transaction.*

In a text database, in which each tweet is a transaction, it will be composed of each of the terms that appear in that tweet once the cleaning processes have been carried out. So the items will be the words. The structure will be stored in a matrix of terms in which the terms that appear will be labelled with 1 and those that are not present as 0. For example for the transactional database $D = \{t1, t2\}$ being $t1 = (just, like, emails, requested, congress)$ and $t2 = (just, anyone, knows, use, delete, keys)$ the representation of text transactions would be as we can see in Table 1.

Transaction\Item	***anyone***	***congress***	***delete***	***emails***	***just***	***keys***	***knows***	***like***	***requested***	***use***
t1	0	1	0	1	1	0	0	1	1	0
t2	1	0	1	0	1	1	1	0	0	1

Table 1. Example of a database with two textual transactions.

3 Related Work

In this section, we will see in perspective the use of Data Mining techniques applied in the field of fake news. This is a thriving area within Data Mining and more specifically Text Mining, in which there are more and more related articles published.

Within the field of text analysis or Natural Language Processing for the detection of fake news, solutions based on Machine Learning and concretely classification problems stand out. This is corroborated in the paper [7], where the authors make a complete review of the approaches to address the problem of analysing fakes news and clearly highlight the problems of classification either by traditional techniques or by deep learning. According to the traditional techniques we find works like [17], in which Ozbay and Alatas, apply 23 different classification algorithms over a set previously labelled fake news coming from the political scene. With this same approach we find the paper [8] in which, the authors apply again a battery of different classification methods that go from the traditional decision trees to the neural networks, all of them with great results. If we look at the branch of deep learning, we also find some works [15], [13], [16] in which the authors try to train neural network models to classify texts in fake news or real. If we look at other Machine Learning methods, another interesting work that focuses on selecting which features are interesting to classify fake news is the paper [18]. On the other hand, we also find solutions based on linear regression as presented by Luca Alfaro et al. in the paper [3]. These works, despite being at the dawn of their development, work quite well but are difficult to generalize to other domains in which they have not been trained.

Because of this, within the aspect of textual entities based on fake news, another series of studies appear that try to address the problem from the descriptive and unsupervised perspective of Text Mining. A very interesting work in this sense, because it combines NLP metrics with a rule-based system is [11], in which in a very descriptive way a solution is provided that is based on the combination of a rule-based system with metrics such as the length of the title, the % of stop-words or the proper names. In the same line there is the proposal in [6] in which authors try to improve the behaviour of a random forest classifier using Text Mining metrics like bigrams, or word frequencies. Finally, in this more descriptive aspect that combines classification and NLP or Text Mining techniques, we also find the social network analysis aspect [12], where the authors classify fake or real news in twitter according to network topologies, information dissemination and especially patterns in retweets.

As far as we know, this is the first work that applies association rules in the field of fakes news. By using this technique we will try to find out which patterns are related to fake news within our domain and try to generalize to possible general patterns related to fake news in other domains of the political field. Due to the impossibility of confronting the system against a similar one, we will carry out in the next sections a descriptive study of the obtained rules.

4 Our Proposal

In this section we will depict the procedure followed in our proposal. For that we will detail the pre-processing carried out on the data. We will also look at the pattern mining process on the textual transactions. For a better understanding we can look at Figure 1. In it we can see how the first part of the process passes

through pre-processing the data, then the textual transactions are obtained, the association rules are applied and results are obtained for fake and real news.

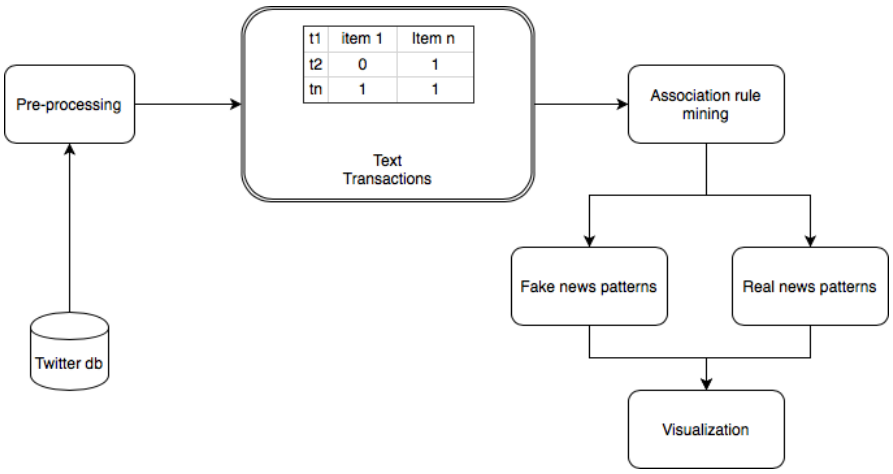

Fig. 1. Process flow for Association Rule extraction in Twitter transactions

Through this processing flow, we offer a system that discovers patterns on fake and real news that can set the basis of new interesting input values for a latter system to, for instance, obtain and classify new coming patterns into real or fake news. In this first approach the system is able to obtain, in a very friendly and interpretable way for the user, which patterns or rules can be related to fake and/or real news.

4.1 Pre-processing

The data obtained from Twitter are often very noisy so it is necessary a pre-processing step before working with them. The techniques used have been:

- Language detection. We are only interested in English tweets.
- Removal of links, removal of punctuation marks, non-alphanumeric characters, and missing values (empty tweets).
- Removal of numbers.
- Removal of additional white spaces.
- Elimination of empty words in English.We have eliminated empty English words, such as articles, pronouns and prepositions. Empty words from the problem domain have been also added, such as, the word via or rt, which can be considered empty since in Twitter it is common to use this word to reference some account from which information is extracted.

- Hashtags representing readable and interpretable terms are taken as normal words, and longer words which do not represent an analysable entity are eliminated.
- Retweets are removed.
- Content transformation to lower case letters.

At this point, we have a set of clean tweets on which we can apply the association rules mining techniques.

4.2 Mining text patterns

The first step in working with association rules and pattern mining in text is to obtain the text entities. To achieve this, the typical text mining corpus of tweets used so far has to be transformed into a transactional database. This structure requires a lot of memory since it is a very scattered matrix, taking into account that each item will be a word and each transaction will be a tweet. To create the transactions, the tweets have been transformed into text transactions as we saw in Section 2.2. We have used a binary version in which if an item appears in a transaction it is internally denoted with a 1, and if it does not appear in that transaction the matrix will have a 0.

The association rule extraction algorithm described in [1] has been used for the results. For this purpose, the parameters of minimum support threshold of 0.005 and minimum certainty factor of 0.7 have been chosen. For experimentation, we have varied the support value from 0.05 to 0.001, with fixed values of confidence and certainty factor.

5 Experimentation

In this section we will go into detail on the experimental process. We will study the dataset, the results obtained according to the input thresholds for the Apriori algorithm and finally the visualization methods used to interpret the operation of the system.

5.1 Dataset

In order to compare patterns from fake news and on the other from real news we have divided the dataset [4] into two datasets depending on whether they are labelled as fake news or not.

After this, we have two datasets, which will be analysed together but being able to know which patterns correspond to each one. The fake news dataset is composed of 1370 transactions (tweets), on the other hand, the real news dataset is composed of 5195 transactions.

5.2 Results

The experimentation has been carried out with different values of supports aiming to obtain interesting patterns within the two sets of data. It is possible to observe in Figure 2 how the execution time is greater as the support decreases, due to the large set of items that we find with these support values.

In the Figure 3 we can see the number of rules generated for the different support values. According to the comparison of both graphs we could draw a correlation between this graph and the previous runtime graph. As for the volume of rules generated and also the time in generating them (that as we have seen offers a graph of equal tendency), it is necessary to emphasize as the dataset fake offers more time and rules, in spite of having less transactions something that comes offered by the variability of the items inside this dataset.

Moreover, in the Figures we see how the AprioriTID algorithm has an exponential increase in the number of rules and execution time when it is executed with low support values or with more transactions. This would rule out in versions based on Big Data, where the volume of input data increases and support must be lowered.

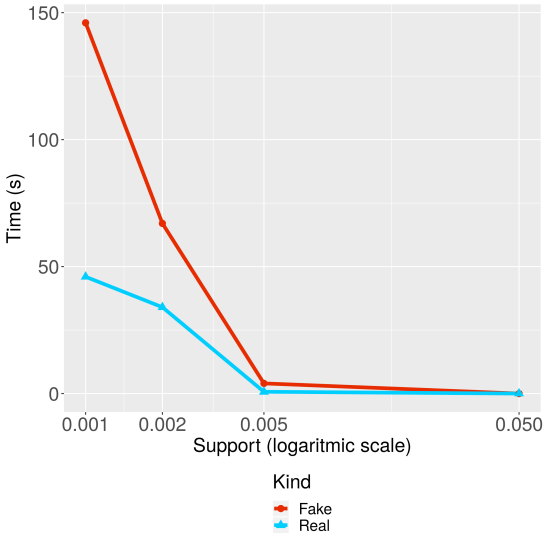

Fig. 2. Execution time of the experiments with different supports

This variability and an interpretation of the obtained patterns can be seen attending to the Table 2 where we have the strongest rules of both datasets. If we pay attention to its interpretation, it is curious how for both datasets we can find very similar rules but with some differences. This may be due to the fact that fake news are usually generated with real news to which some small

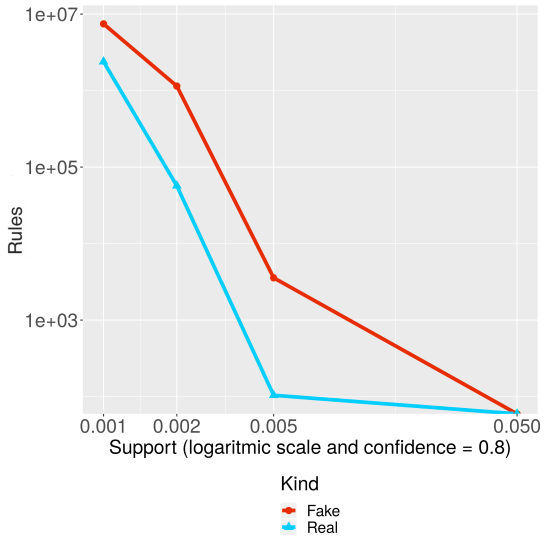

Fig. 3. Number of rules of the experiments with different supports

element is changed. This is something that the rules of association discover for example in the rules $\{sexism,\ won\} \rightarrow \{electionnight,\ hate\}$ for fake news and the rule $\{sexism,\ won\} \rightarrow \{electionnight\}$ for real news. We can also observe how the tendency is to discover more items in the rules corresponding to fake news, probably caused by these sensationalist adornments that are usually charged to fake news.

antecedent	*consequent*	*supp*	*conf*	*Dataset*
electionnight, won, hate	**sexism**	0.0075	0.909	Fake
sexism, won	**electionnight;hate**	0.0075	0.9	Fake
didnt, trump, won, electionnight, sexism, win, racism	**hate**	0.06	1	Fake
sexism, won	**electionnight**	0.0051	0.97	Real
projects	**foxnews**	0.005	1	Real

Table 2. Example of rules obtained in the experiments

5.3 Visualization

A system that is easily interpretable must have visualization methods so we have focused part of the work on obtaining and interpreting interesting and friendly graphics on the fake and real news. We can observe the results obtained through the graphics of the Figures. In Figure 4 we can see the rules obtained for the fake news, where we can appreciate that the resulting rules associate in great

quantity of occasions to trump with sexist, winning or racist. But some of them are interesting because the indicate the opposite, like the rule that relates*racist, trump, didnt and sexist*.

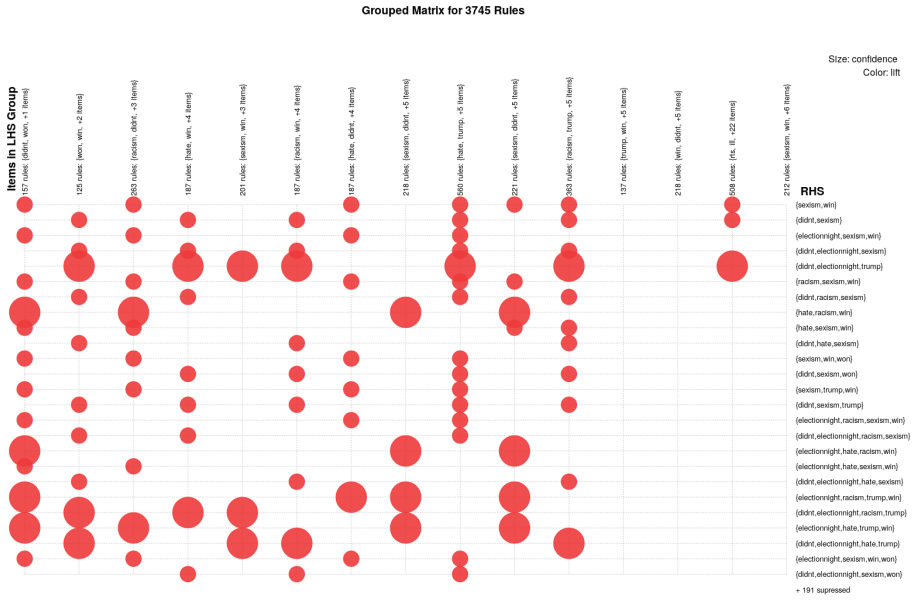

Fig. 4. Example of rules in fake news

On the other hand, in Figure 5 we can see the rules obtained for the real news. Here we can see how fewer rules are obtained for experimentation and that the terms that appear in them encompass media such as $fox, news, usa$ or *winning*. Studying the terms that appear in both examples we can see racist that in this case is associated with fox and *donald*.

Finally, a graph has been generated, which can be seen in Figure 6, with the results of the fake news filtering the 80 rules with a higher certainty factor. It can be seen that there are three groups of terms, one with very interconnected negative terms and another with very frequent terms due to the subject matter.

6 Conclusions and Future Work

In conclusion, we can see how the application of Data Mining on this kind of data allows us to extract hidden patterns. These patterns allow us to know better the terms more used in each type of news according to if it is false or real in addition to the interrelations between them.

Fig. 5. Example of rules in real news

Fig. 6. Example of rules in fake news

Data mining techniques and, in particular, association rules have also been corroborated as techniques that can provide relevant and user-friendly information in Text Mining domains such as this.

In future works we will extend this technique in order to classify new tweets using the information provided after the application of association rule mining. Another application would be the use of the extracted patterns in order to create a knowledge base that can be applied in real time data.

Acknowledgement

This research paper is part of the COPKIT project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687 and the program of research initiation for master students of the University of Granada.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Acm sigmod record. vol. 22, pp. 207–216. ACM (1993)
2. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB. vol. 1215, pp. 487–499 (1994)
3. de Alfaro, L., Di Pierro, M., Agrawal, R., Tacchini, E., Ballarin, G., Della Vedova, M.L., Moret, S.: Identifying fake news from twitter sharing data: A large-scale study (2019)
4. Amador Diaz Lopez, J., Oehmichen, A., Molina-Solana, M.: Fakenews on 2016 us elections viral tweets (november 2016 - march 2017) (Nov 2017). https://doi.org/10.5281/zenodo.1048826, https://doi.org/10.5281/zenodo.1048826
5. Berzal, F., Blanco, I., Sánchez, D., Vila, M.A.: Measuring the accuracy and interest of association rules: A new framework. Intelligent Data Analysis **6**(3), 221–235 (2002)
6. Bharadwaj, P., Shao, Z.: Fake news detection with semantic features and text mining. International Journal on Natural Language Computing (IJNLC) Vol **8** (2019)
7. Bondielli, A., Marcelloni, F.: A survey on fake news and rumour detection techniques. Information Sciences **497**, 38–55 (2019)
8. CORDEIRO, P.R.D., Pinheiro, V., Moreira, R., Carvalho, C., Freire, L.: What is real or fake?-machine learning approaches for rumor verification using stance classification. In: IEEE/WIC/ACM International Conference on Web Intelligence. pp. 429–432 (2019)
9. Delgado, M., Ruiz, M.D., Sanchez, D.: New approaches for discovering exception and anomalous rules. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **19**(02), 361–399 (2011)
10. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: ACM sigmod record. vol. 29, pp. 1–12. ACM (2000)
11. Ibrishimova, M.D., Li, K.F.: A machine learning approach to fake news detection using knowledge verification and natural language processing. In: Barolli, L., Nishino, H., Miwa, H. (eds.) Advances in Intelligent Networking and Collaborative Systems. pp. 223–234. Springer International Publishing, Cham (2020)

12. Jang, Y., Park, C.H., Seo, Y.S.: Fake news analysis modeling using quote retweet. Electronics **8**(12), 1377 (2019)
13. Kaliyar, R.K.: Fake news detection using a deep neural network. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA). pp. 1–7. IEEE (2018)
14. Martin-Bautista, M., Sánchez, D., Serrano, J., Vila, M.: Text mining using fuzzy association rules. In: Fuzzy logic and the internet, pp. 173–189. Springer (2004)
15. Molina-Solana, M., Amador Diaz Lopez, J., Gomez, J.: Deep learning for fake news classification. In: I Workshop in Deep Learning, 2018 conference spanish association of artificial intelligence. pp. 1197–1201
16. Monti, F., Frasca, F., Eynard, D., Mannion, D., Bronstein, M.M.: Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673 (2019)
17. Ozbay, F.A., Alatas, B.: Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications **540**, 123174 (2020)
18. Reis, J.C., Correia, A., Murai, F., Veloso, A., Benevenuto, F., Cambria, E.: Supervised learning for fake news detection. IEEE Intelligent Systems **34**(2), 76–81 (2019)

3 NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise

3.1 NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise

- Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2022). NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise. Expert Systems with Applications, 208, 118063.
  - Journal: Expert Systems with Applications
  - Status: Published.
  - Impact Factor (JCR 2021): 8.665.
  - Category: Computer Science, Artificial Intelligence. Order 21/145 Q1.

NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise

J. Angel Diaz-Garcia[a,*], M. Dolores Ruiz[a], Maria J. Martin-Bautista[a]

[a] *Department of Computer Science and Artificial Intelligence, University of Granada, Daniel Saucedo Aranda, s/n, 18014 Granada, Spain*

Abstract

Social networks have taken an irreplaceable role in our lives. They are used daily by millions of people to communicate and inform themselves. This success has also led to a lot of irrelevant content and even misinformation on social media. In this paper, we propose a user-centred framework to reduce the amount of irrelevant content in social networks to support further stages of data mining processes. The system also helps in the reduction of misinformation in social networks, since it selects credible and reputable users. The system is based on the belief that if a user is credible then their content will be credible. Our proposal uses word embeddings in a first stage, to create a set of interesting users according to their expertise. After that, in a later stage, it employs social network metrics to further narrow down the relevant users according to their credibility in the network. To validate the framework, it has been tested with two real Big Data problems on Twitter. One related to COVID-19 tweets and the other to last United States elections on 3rd November. Both are problems in which finding relevant content may be difficult due to the large amount of data published during the last years. The proposed framework, called NOFACE, reduces the number of irrelevant users posting about the topic, taking only those

*Corresponding author

Email addresses: joseangeldiazg@ugr.es (J. Angel Diaz-Garcia), mdruiz@decsai.ugr.es (M. Dolores Ruiz), mbautis@decsai.ugr.es (Maria J. Martin-Bautista)

that have a higher credibility, and thus giving interesting information about the selected topic. This entails a reduction of irrelevant information, mitigating therefore the presence of misinformation on a posterior data mining method application, improving the obtained results, as it is illustrated in the mentioned two topics using clustering, association rules and LDA techniques.
Keywords: social media mining, pre-processing, vea, credibility, word embeddings

1. Introduction

Social networks have become an essential part of our lives. They are great sources of information, used daily by thousands of people to explore news and share their opinions about them. This great success has also led to the increasing spread of irrelevant information, hoaxes or misinformation, even interfering in electoral processes (Allcott & Gentzkow, 2017). Twitter, is one of the most successful social networks today, and undoubtedly the most used social network to share and comment on news around the world. Its character is mainly public so anyone can see that someone else is tweeting about a certain topic. This has led to Twitter gain on a relevant role, for example, to obtain relevant information in real time about events and disasters, but it has also made it a target for those who want to spread misinformation. But, what are relevant information and misinformation?

In our field of application, relevant information is understood as information that may contain valuable content in a certain context. For example, in a health topic, relevant information would be that issued by a doctor about prevention measures for a certain disease. That is, in the case of Twitter, for the proposed case of health, we will be interested in keeping those candidates (tweets) of relevance to medicine, discarding those samples (tweets) that are not related to this sector.

As for misinformation or disinformation, there are more and more papers that provide a description or new characteristics of this concept (Kar & Aswani,

2021), (Aswani et al., 2019). Specifically, in (Aswani et al., 2019) they provide a very interesting vision of misinformation seen in different ways such as willful misinformation, fictional discussions, and non-verifiable information or news. In any case, it is untruthful information that is disseminated for various purposes, such as to negatively influence political issues. In these cases where there is a clear intention to disseminate false content, we will speak of disinformation. Therefore, the difference between misinformation and disinformation lies in the intentionality or purpose. In the scope of our paper, we will focus on experience-driven misinformation reduction, since a person on Twitter may share false content, because of their beliefs without actually knowing whether that is false to a greater or lesser extent.

Being able to discern what is true or relevant in social networks and what is not, has taken up a great amount of literature in recent years (Shu et al., 2017), (Oehmichen et al., 2019), with artificial intelligence systems assuming a great importance in the process. The process of eliminating misinformation on social networks is, by its nature, closely linked to the processes of dimensionality reduction and instance selection, as both seek to eliminate data that is not interesting for subsequent data mining processes. This process can be guided by statistical methods such as features or instance selection algorithms (Olvera-López et al., 2010), or by objective credibility data in the case of misinformation detection.

Our goal is to design a framework to address the problem of relevant content selection in social networks. With this objective, we seek to obtain smaller and more cohesive datasets on which to obtain better knowledge with subsequent data mining processes. From this main objective, the elimination of misinformation can be derived. The framework will select only those accounts with experience and credibility, which will therefore eliminate to some extent the possible misinformation present in the dataset.

In this paper, we propose a framework based on iterative filters, word embeddings, user authority and credibility, to reduce the irrelevant content of data and, at the same time, increase the confidence of retrieved information coming

from social networks. With our framework, we achieve a more cohesive, clean and truthful dataset that can be used in subsequent processes with greater accuracy regarding the credibility of the source and the data. The system is based on the premise that if a user is credible, his or her content will also be credible. Therefore the proposed system will identify which users are credible and get their tweets according to a specific topic. The major contributions of the work to the state of the art are as follows:

- A new framework based on iterative filters, word embeddings and credibility is proposed for instance reduction in social media.
- A new method for filtering irrelevant information in social networks is proposed.
- A new algorithm is proposed for the selection of credible users in the social network Twitter based on the popularity and expertise of the user. To do this we focus on the user's biographies and process it with word embedding. As far as we know, this is the first work that applies word embeddings techniques to the biographies of the user on Twitter, using this to discern the user's expertise on a certain topic.
- The functionality and versatility of the proposed pre-processing system that can be used with a wide variety of data mining techniques, especially those sensitive to the amount of data. The framework has been tested with LDA, association rules and clustering techniques.

In order to validate the system, a set of experiments has been devised in which K-means Clustering (MacQueen et al., 1967), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Apriori algorithm (Agrawal et al., 1994) are applied to large sets of data from Twitter, one related to the COVID-19, and another one concerning the United States elections. The experiments show the differences in the results when NOFACE (NOise Filtering According Credibility and Expertise) framework is applied or not. The results illustrate the reduction of content, processing efficiency, achieving an improvement of the algorithms

used. This improvement will come in terms of a more manageable dataset and execution times.

The paper is organized as follows. In the following section, we study related work and different approaches for instance selection and dimensionality reduction according to credibility in social networks. In Section 3 we go into detail in the NOFACE framework describing each of the constituent modules. In Section 4, we explain the evaluation of the framework, which will be based on comparing the results of different data mining techniques, when NOFACE is applied or not. In Section 5, we discuss some of the challenges that this framework and others in the literature should face in the future. Finally, in Section 6, we examine the conclusions remarks and the future work of the research carried out.

2. Related work

The dimensionality reduction seeks to have clean data without noise, or loss of information, more understandable, and easily manageable and processable, something that highlights its relevance to Big Data problems. Dimensionality reduction usually focuses on the reduction of training variables (Solorio-Fernández et al., 2020), but there is also a branch dedicated to reduce or select the number of examples, which is where instance selection algorithms appear. By definition, the instance selection is framed within the techniques of data pre-processing and has been approached from multiple and different perspectives over the course of the years (Olvera-López et al., 2010), (Chandrashekar & Sahin, 2014) . In problems related to social networks, where misinformation, noise and massive amounts of data are always part of the problem, these dimensionality reduction techniques are paramount. In this context, we focus on credibility-based dimensionality reduction techniques. In credibility analysis, the amount of data related to a problem is also reduced, but this reduction is guided by factors inherent to the social network where it is applied. These factors can be, for example, the followers or the engagement or the expertise of a user in a certain topic (Canini et al., 2011). The analysis of credibility and the consequent reduc-

tion of noise and examples of the problem has been approached from different perspectives within Data Science and Artificial Intelligence. These perspectives depend on the techniques used and the granularity of the entity on which its credibility is studied. At this point, we have carried out a study of the state of the art of the credibility analysis in Twitter. We have classified the works according to their granularity in content (tweet or topic) level or user level.

Since the framework, by selecting relevant and credible content, can also help to eliminate misinformation, additionally we have conducted a literature review on this aspect.

2.1. Content level credibility

Castillo et al. (Castillo et al., 2011) have addresed the problem of credibility on Twitter. Their research is one of the most comprehensive and attempts to classify contents (tweets) based on whether they are credible or not. To evaluate and create the model, they use a large number of indicators that are closely linked to the analysis that a human would do to study the credibility of a tweet, such as whether an account is verified, whether the user has enough followers or whether the tweet uses appropriate hashtags, to name some of the features taken into consideration by the classification system. Concerning the user, it also obtains information but in a very simplistic way: for example, if it has biography or if it is empty. At this point NOFACE goes one step further, analysing the biography completely and obtaining knowledge of it to guide the process of content filtering and dimensionality reduction.

Kang et al. offers in (Kang et al., 2012) three different ways to obtain a credibility rating. The first proposal analyses the social graph of Twitter, by means of ratios between concepts like retweets or number of followers. The second one focuses on content, and finally the third model is a hybrid model that takes into consideration graphs and content. Being the first model, the one based only on graphs, which works best.

Finally, there is also a credibility-oriented dimension to event-related content. Hassan (Hassan, 2018) uses text mining techniques on event-related tweets.

The text mining techniques used are guided by the frequency of terms in different topics, and finally the algorithm is evaluated using different classifiers.

2.2. User level credibility

With regard to the analysis of user credibility, we find approaches such as those of Cognos or CredSaT. Cognos (Ghosh et al., 2012) offers a web solution for searching experts in a certain topic, for this, it uses Twitter lists. The lists on Twitter are user-managed lists, in which users add other users related to topics. Cognos exploits this potential, even improving the search for accounts in the native system of recommendation of Twitter. The CredSaT (Abu-Salih et al., 2019) system, is a Big Data solution that takes into consideration the content and the time stamp to create a ranking of expert and influential users in the social network. It also adds a semantic analysis layer with sentiment analysis on tweets and responses used to enrich the final corpus of experts.

Unlike these approaches, the NOFACE seeks to reduce the amount of data, that is, the aim is not to search for influential people but to guide content reduction of a social media dataset through expert users.

Finally, it is necessary to mention the papers proposed by Alrubaian et al. (Alrubaian et al., 2016), (Alrubaian et al., 2018). These papers also deal with the analysis of credibility on Twitter in a very exhaustive way and similar to NOFACE through 3 modules. These modules deal with content credibility, reputation and expertise. However, NOFACE obtains the expertise according to the user's biography, instead of according to the content as made in (Alrubaian et al., 2016), (Alrubaian et al., 2018). Our approach exploits the potential of biography in social networks such as Twitter, where it is very common to talk about professions. As far as we know, this is the first work that addresses and uses this option in addition to word embeddings, and with great results as we will see in future chapters. Additionally, our analysis of reputation is different to that of the above-mentioned papers focusing more on the engagement of user-generated content, which will give a value about how interesting is a user's content to his or her followers.

2.3. Misinformation detection

In the field of misinformation and fake news reduction, we find that supervised approaches are the most widespread. In (Ozbay & Alatas, 2020) Ozbay and Alatas, apply 23 different classification algorithms over a previously labelled fake news dataset coming from the political scene. With this same approach, we find the paper (Cordeiro et al., 2019) in which, the authors apply again a battery of different classification methods that go from the traditional decision trees to the neural networks, all of them with great results. Also in the field of classification, but using bio-inspired algorithms, we find the paper (Batra et al., 2021). In this paper, the misinformation, or worthless information, comes in the form of email spam. The authors create a classifier based on K nearest neighbours and bio-inspired algorithms to obtain the instances that best represent the problem domain according to three different distance metrics.

If we look at the branch of deep learning, many papers have been used to detect fake news or rumours. One of the first is the one proposed in (Ma et al., 2016). In this paper, they use an architecture based on three layers. The first layer uses as input the K most significant elements of the text based on the tf-idf ratio to train a Recurrent Neural Network (RNN). Then it uses a Long Short-Term Memory (LSTM) layer to model the dependencies along with the text and in a final step it uses a layer based on a Gated Recurrent Unit (GRU). The model improves on the performance of other base models. Although the model gives good results, it is necessary to train the network, so it is necessary to re-train it for its use in another domain, because it is necessary to have labelled databases. This is something that does not need to be taken into account in systems based on data and user characteristics such as NOFACE and other models such as the one proposed in (Castillo et al., 2011).

Also within the framework of deep learning there is a wide range of papers (Molina-Solana et al., 2018), (Kaliyar, 2018), (Monti et al., 2019). These papers have a similar focus. They use pre-labelled fake and non-fake databases to train classifiers based on neural networks. The proposal in (Kumari et al., 2021) is based on the use of concepts such as novelty and emotions to guide the

detection of misinformation. Its foundation is the premise that this type of news and information tends to be emotionally charged to favour its diffusion. To do so, they use a combination of BERT, LSTM and feed-forward networks. In (Nasir et al., 2021) authors also propose a combination of different neural network topologies. Specifically, they use Convolutional Neural Network (CNN) to obtain fake news features and applies in a later stage RNNs to store the sequential dependencies between terms. The output is then used for fake news classification.

Other papers use more novel approaches. For example, the work (Khoo et al., 2020) use Twitter conversations generated around fake news to early detect the spread of fake news using neural networks. The work (Liu & Wu, 2020) uses the concatenation of user-related features and user-generated text to use them as input in a rumour classification layer applying word embeddings.

These previous approaches, although novel, also require training. Our model uses word embeddings in an unsupervised way, helping to reduce the amount of irrelevant data and, to some extent, mitigating the problem of false information. Importantly, NOFACE also uses a conjunction of user-based (number of favourites or retweets) and text-based (experience-related words present in the biographies) features. The potential of this feature fusion has been also highlighted in papers such as (Liu & Wu, 2020).

In summary, the major differences of the NOFACE framework compared to the solutions seen in the literature are:

- The main aim of NOFACE is the reduction of content coming from social networks considering the credibility and expertise of the publisher. To achieve this main objective, the reduction is guided by credibility, engagement and expertise analysis. These tasks are an intermediate stage of the main purpose, being, therefore one of the first methods of this kind.

- NOFACE offers a more restrictive cascade approach than other approaches where credibility, expertise or engagement is computed for all examples. NOFACE discards those that do not pass the first filter, the second filter

and so on.

- NOFACE is, as far as we know, the first framework that applies word embeddings and text mining to the process of computing expertise through biographies.

- NOFACE offers an interpretable way to locate useful content in social networks and without having to train a classifier or neural network, so it can be used as a first stage of analysis on any dataset without the need to have a ground truth.

3. The NOFACE framework architecture

In this section we will go into detail in the NOFACE framework. In Figure 1 we can see a general diagram of our framework NOFACE (NOise Filtering in social media According to Credibility and Expertise). The system is based on Twitter databases, on which the different modules are applied in cascade. The first module, focused on expertise, is based on a filter that uses word embeddings, concretely FastText (Bojanowski et al., 2016), to obtain those descriptions of users with greater expertise about a certain topic according to their profession. As far as we know, this is the first work that applies word embeddings on the descriptions instead of on the tweets. Once filtered by the users' expertise, a new filtering step is necessary to discard those profiles without interaction or credibility on the network. Afterwards, the next two modules, about the engagement and the credibility, are applied over the selected users by the first module. This new filter is based on the network metrics, such as the number of followers, if they publish valuable content, the favourites or the retweets. The final result is a very reduced set of data where tweets have been generated by people who not only have credibility on the net, but also have experience in the topic under analysis.

The central core of our framework is to be able to know which people talking about a certain topic on Twitter really have a relationship in terms of experience

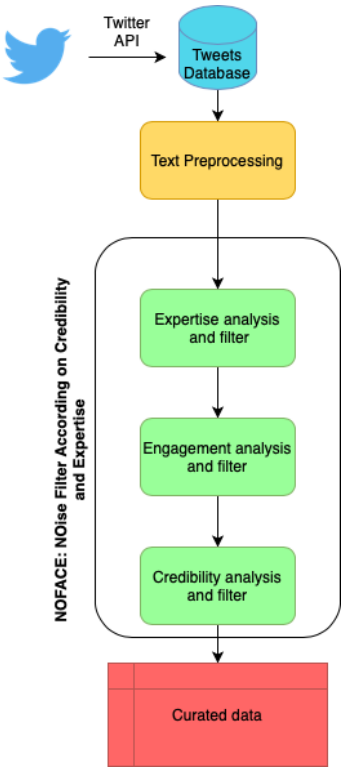

Figure 1: NOFACE framework.

and credibility with the topic under study. Modelling credibility and expertise is an arduous task, approached by other papers in a quite exhaustive and efficient way (Alrubaian et al., 2016). Our framework is based on the premise that if a Twitter user is credible, his or her content is also credible and therefore the user can be used to guide the selection of the content.

Our framework is based on 3 modules plus a pre-processing module that are applied in cascade, that is, the users that do not pass the first filter, will not pass to the second one, which implies that we reduce computation times and generate a reduced and trustworthy solution. The framework, first applies a pre-processing method. It then computes the expertise through biographical analysis, then focuses on content quality (engagement) and finally filters based on the user's credibility on the topic. In (Kumar et al., 2021) authors highlight

the value of text mining applied in the field of credibility and fake information. They mention some approaches that corroborate our premise as valid. For example, in (Barbado et al., 2019) authors detected a strong weight between several social features, such as lists or the number of followers, which are associated with more trustworthy content. The relation between misinformation and certain user-inherent components such as the longevity of the account, the presence of certain words in its content or the interaction generated with other users of social networks, has also been analysed in the state-of-the-art review in (Wu et al., 2019). Therefore, we find feasible to create a system based on user features in conjunction with expertise to remove irrelevant information from social networks.

3.1. Pre-processing module

The pre-processing module uses common cleaning techniques in addition to others specific to the Twitter domain. All the techniques have been modularized in a Python function that allows to select the language of the text corpus, to guide sections like the elimination and detection of stop words. The pre-processing is applied to the user's biography and to the tweet content. The techniques used and their order of application are:

1. Twitter domain related cleaning. For this purpose we have eliminated URLs, hashtags, mentions, reserved words from Twitter (RT, FAV...), emojis and smileys.
2. Cleaning of the usual text mining domain, removing numbers, additional spaces and punctuation marks.
3. Turning the text into lowercase letters.
4. Detecting the tweet language, all those tweets using a non-recognised language or from a language other than the one desired by the user are eliminated.
5. The stop words of the language introduced by the user in the pre-processing function are removed.

6. Any empty tweets (composed of items eliminated in previous stages of pre-processing) are removed.
7. Tokenization of the biography and the tweet.

After this process, we have achieved that the raw content coming from Twitter, can be easily processable in later stages. Although NOFACE only uses data related to the users (biography, followers, etc.) it also cleans data related to the tweet content, to generate a final corpus useful for posterior data mining processes.

3.2. Expertise filter module

The first filter and therefore the main filter and greatest contribution to the state-of-the-art of this approach is the expertise filter. This filter must be able to locate and eliminate those users who are not really related to the topic. To do this, the filter will exploit the biographies of the users to the extreme. The Twitter code itself, also takes advantage of the biographies in its service *Who To Follow*, so we can conclude that using biographical content can be of a great interest. NOFACE will use the biographies to create a filter and through the use of word embeddings, it will exploit the potential of biographies in a very exhaustive way.

The vector space representation using word embeddings corresponds to the current state-of-the-art in Natural Language Processing (Liu et al., 2015), (Levy & Goldberg, 2014). The underlying technique is to represent all the words within a given vocabulary in vector space as vectors. With these vectors, operators such as addition or substraction can be applied, so that the words *king - man + woman* would result in the word *queen*.

Our expertise filter exploits this potential by using a representation in which each word is represented by a vector in the vector space. We can see the pseudo-code in Algorithm 1. Our algorithm uses the power of semantic relations between words to increase the search space in Twitter biographies. Many works have demonstrated the power of word embeddings to expand search queries. In (Roy

et al., 2016) Roy et al. use the KNN algorithm on vector space generated by embbedigns to obtain which terms are most similar to others and expand the search query. With a similar point of view we find the works (Diaz et al., 2016) and (Kuzi et al., 2016). In (Diaz et al., 2016) Diaz et al. train locally embedding, namely GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013a), (Mikolov et al., 2013b), to improve search processes in information retrieval. In a very similar way but with Word2Vec+CBOW Kuzi et al. demonstrate in (Kuzi et al., 2016) how document retrieval actually improves with this technique. In our algorithm, we will use this potential of word embeddings, not for retrieving documents, but for locating users that are experts on a topic.

Our algorithm works as follows. We introduce a list of words related to the topic under study, for example, about medicine, we can introduce words: *medical, doctor*. The algorithm will start to train a word embedding model on the biographies using a part of a data partition (10% of the entire dataset), and in the first iteration it will obtain the 5 most similar words to *medical* and *doctor* among the corpus itself. The algorithm will use these 12 words, (5 more similar to medical, 5 more similar to doctor, besides doctor and medical), to find users whose biographies contain any of these terms and start creating the list of experts and topic-related users. In the next iteration, we will already have 12 words to search for their 5 similar ones, and so on. This leads to an exponential growth of words linked by word embeddings to the domain of the problem. To prevent this and thus leading to a degeneration in meaning relative to the input words, the stopping condition of the algorithm is 3 iterations. This will enrich the space of words obtaining very related and linked to the topic words that will guide our filter.

In each iteration, the algorithm checks if any user id is already present in the expert list to avoid processing it again, since its words and content are already in the search corpus, thus avoiding additional processing. The output of the algorithm is a clean set of data in the form of a data frame ready to be processed in the following modules. It should be noted that at the end of the computation of the algorithm we will have a new column in the dataset, where

we will see the words learned and used for filtering. In this way, a potential user can see in a readable and interpretable way what set of words the algorithm has used on each account to determine if it is a valuable account for analysis. The interpretability of the result of each step of the algorithm is therefore a value to be taken into account.

There are a multitude of models and representations for word embedding, so for fine-tuning our algorithm we have compared Word2Vec and FastText (Bojanowski et al., 2016), since they are the most widespread and relevant in terms of versatility and performance at present. The main difference between Word2Vec and FastText is that the latter decomposes each of the input words in the neural network into n-grams, for example for the word *matter*, and *n=3* we would have {*ma, mat, att, tte, ter, er* } and the final representation would be the sum of the vectors associated with each n-gram. This representation is very interesting to discern out-of-vocabulary words or words with low presence in the dataset. Word2Vec takes each word as a vector, therefore does not consume as much memory and resources as FastText (which for each word stores a vector per n-gram), although it is more sensitive to out-of-vocabulary words. For each of these embedding models, we have two representations, Skip-Gram and Continuous Bag of Words (CBOW). Skip-Gram tries to predict the context words surrounding the word, i.e. it predicts context based on a word. On the other hand, CBOW predicts a word based on the surrounding context words, i.e. it predicts a word based on the context. Regarding the embeddings parameters, it has been run with a window of 5 words, words with frequencies lower than 2 have been ignored and negative sampling of 10 has been done in the case of CBOW and a hierarchical softmax in the case of Skip-Gram. The dataset used to fine-tune the algorithm, is about COVID-19 (4.1.1) and is composed by 3 batches of 936.427, 1.062.900 and 1.319.912 tweets respectively (total 3.319.239 tweets). The results in the case of Word2Vec can be seen in Table 1. The results in the case of the experiments carried out with FastText are presented in Table 2.

In our problem, where we have few context words due to the fact that Twitter

Algorithm 1: Expertise filter algorithm

```
Result: Dataframe with experts in the topic
# pre-processing, initializing the variables and data structures
cleaned-dataset=preprocess(dataset)
expert_set=[]
finaldataframe=pd.dataframe()
split cleaned-dataset into batches
for batch in batches do
    #we check if any user of the batch is already located as an expert
    if user_id in expert_set then
        # For experts, we add all their tweets to the final data frame and do not
          process
        finaldataframe.extend(batch[id=user_id])
    else
        # We process the rest of the content to locate new experts
        # get the biographies tokens
        tokenized_tweet = batch['biographies_clean']
        # train the word2vec model
        model=train_word2vec(tokenized_tweet)
        # create a list with the words of the list present in the model
        final_words=[]
        for word in expert_words do
            if word in model then
                final_words.append(word)
            end
        end
        # create a data frame with the 5 most similar words to each expert word
        for word in final_words do
            most_similar.extend(find_5_mosts_similar(model, word))
        end
        # extend the expert word list with the most similar words in the embedding
        expert_words.extend(most_similar)
        # We locate all users who have in their biographies any of the words
        new_experts =find_users(batch['biographies_clean'], expert_words )
        # Extend the final data frame with the tweets of the located experts
        finaldataframe=finaldataframe.extend(batch[user_id in new_experts])
        # Extend the expert set with the new experts
        expert_set.extend(new_experts)
    end
end
```

	Word2Vec+CBOW	Word2Vec+Skip-Gram
Elapsed Time	Min: 11min 7s	Min: 13min 1s
	Max: 13min 36s	Max: 14min 22s
Words	Min: 209	Min: 45
	Max: 228	Max: 64
Users located	Min: 21 390	Min: 7 937
	Max: 23 377	Max: 10 044
Final dataset size	Min: 31 347	Min: 12 281
	Max: 34 107	Max: 15 239

Table 1: Minimum and maximum value for each variable in the Word2Vec experiments.

texts are not very large, this makes an important difference. It is easier for the algorithm to predict context words based on a single word (Skip-Gram), than to predict a single word based on several words (CBOW), since the search space and the window within each document (tweet) is very small. A priori it may seem that this does not influence, since Word2Vec+CBOW obtains great results, but if we compute the ratio of users found for each word, we can see how this value is 102 users per word on average in Word2Vec+CBOW, while this rises to 156 users per word in the case of Word2Vec+SkipGram. This ratio in the case of FastText+CBOW stands at 155, while in FastText+SkipGram the ratio reaches a value of 162. This leads us to conclude that the words located by FastText have a higher representation in the dataset, as well as a higher relationship with the topic under study. Therefore, the best option for our algorithm will be to use FastText+SkipGram, although more time-consuming. this increase is also linked to a higher match value for the selected words and their relation to the topic, as well as a better user selection ratio.

3.3. Engagement filter module

The next filter concerns engagement. Engagement could be defined as the capacity of a user to generate useful content that is appreciated by other users

	FastText + CBOW	FastText + Skip-Gram
Elapsed Time	Min: 16min	Min: 18min 20s
	Max: 17min 15s	Max: 20min 10s
Words	Min: 50	Min: 111
	Max: 79	Max: 125
Users located	Min: 10 975	Min: 18 128
	Max: 12 312	Max: 20 306
Final dataset size	Min: 16 748	Min: 26 547
	Max: 18 773	Max: 31 611

Table 2: Minimum and maximum value for each variable in the FastText experiments.

of the social network. In other words, it is a measure of how good is the content a user publishes on a social network. In the specific case of Twitter, interaction is usually measured in terms of RTs (Retweets) and FAVs (Favourites). A RT corresponds to a share, i.e. another user finds your content useful and shares it with their community. On the other hand, a FAV, corresponds to a 'like' on Facebook or Instagram, i.e. a way for users to indicate that they like a particular tweet.

At this point, we would like to make a distinction between users who have many followers and those who have few followers. A person with many followers, consequently will also have more interaction than one with few followers, but this does not imply that their content is better, therefore, we will define engagement as the arithmetic mean of the interaction variables: number of retweets and number of favourites, normalised by the number of followers of the user. The number of retweets and favourites used is the accumulated sum of user retweets and favourites in the dataset in question, i.e. the retweets or favourites that a user has received on the topic under study. This value, therefore, offers a contrast to other formulas seen in the literature (Baum, 2019), where engagement modelling is done globally for all user-generated content without taking into account that this content may belong to more than one topic. We find that

it is closer to reality to consider the engagement of a user for a certain topic, rather than the global engagement since a user can have experience in different areas. For example, a user can tweet about Artificial Intelligence (A.I.) with little success, and at the same time, about a sport that he practices, getting a lot of interaction in the tweets related to the sport. If we are analysing a topic related to A.I. and consider the global engagement of the user, we may have a bias that tells us that the user is relevant to A.I., when he or she is not.

Mathematically for each user $u \in U$, their engagement, denoted as ϵ, is calculated with the following formula:

$$\epsilon(u) = \frac{\frac{nFavsInTopic}{nFollowers} + \frac{nRtsInTopic}{nFollowers}}{2} \tag{1}$$

In this way, we achieve to increase the engagement of a user that has generated very useful content about the topic under analysis. For example, let's suppose the following simple example: *user1* with 20 Followers, 50 Favs, 30 Rts, and *user2*, with 1000 followers, 100 Favs and 120 Rts. If we apply the formula (1), we will have $\epsilon(user1) = 2$ and $\epsilon(user2) = 0.11$. So *user1*, will have more useful content than *user2*, because despite having less followers, they share more content in proportion to *user2*, who although having more followers, they do not interact as much. To pass the filter, we will select those accounts whose ϵ value is higher than the mean of all the engagement values.

It is necessary to mention that the modelling of engagement is very complicated, since the system can be susceptible to mark relevant users that interact a lot in the social network without caring about the content, although normally, the content that is relevant is shared. This is much more accentuated if we are in the professional field, where networks such as Twitter are often used to share and find research, results or studies. It is here where the cascade filter comes into play, because the previous expertise filter has already discarded non-professional accounts, so the system is less sensitive to this problem of sharing less valuable content.

3.4. Credibility filter module

The last filter is based on the credibility of the user. The user's credibility on a social network is intimately related to his or her popularity. That is, an account becomes popular because many other accounts believe it and therefore follow it and share its content. In other words, we can model credibility for our filter, based on an arithmetical mean of the Twitter values that are related to popularity. These values are: the number of followers, the number of public list in which the user appears, the number of retweets and the number of favourites. In the literature, other works closely related to the NOFACE framework use a standardized linear calculation of variables such as the number of followers, favourites, retweets and mentions. We have preferred to give importance to the lists, as opposed to the mentions, because the mentions are not necessary a good indicator as they can be mentions of anger or reproach, while the lists, have demonstrated in solutions like Cognos (Ghosh et al., 2012) offering good results. According to this, mathematically for each user $u \in U$, their credibility, denoted as ζ, is calculated with the next formula:

$$\zeta(u) = \frac{nFollowers + nLists + nRetweets + nFavs}{4} \tag{2}$$

To pass this last filter, the value must be above the mean of all ζ values. After applying this filter, the system will capture those user accounts related to the topic under study, whose content is usually interesting and who also have a wide popularity and credibility in social networks.

4. Framework evaluation

In this section we will go into detail in the experimentation carried out with the NOFACE framework as well as its application to a real problem. It is worth mentioning that all the code has been programmed in Python 3 and that the tests, the development and the application to a real problem have been carried out with the equipment whose specifications are shown in Table 3. The equipment is a non-professional laptop, which shows that the potential of

certain techniques such as word embeddings can be democratized, and that a useful system does not necessary have to use large processing clusters to obtain a meaningful result.

Component	Features
CPU	2 GHz Intel Core i5 with 4 cores
RAM	16 GB 3733 MHz LPDDR4X
VRAM	Intel Iris Plus Graphics 1536 MB
Hard Disk	SATA SSD de 512 GB

Table 3: Machine specifications.

4.1. Datasets

To check that the framework performs properly, it has been applied to two real problems, one relating to COVID-19 and the other to the US elections in November 2020. The datasets used for the experimentation have been released on (Diaz-Garcia et al., 2022). In this repository, the source code of the algorithms will also be released.

4.1.1. COVID-19 Dataset

The disease caused by the new Coronavirus (Sars-Cov-2) (Zhou et al., 2020), first reported in Wuhan (Huang et al., 2020) in December 2019, now affects the entire world and is considered one of the largest pandemics in the history of mankind. The virus is present in all inhabited areas of the world, and has caused millions of infections and millions of deaths. Europe is currently one of the epicentres of the pandemic and is likewise one of the territories that is allocating the most resources to research into the new disease. One of the ways of research related to the pandemic, lies in the automatic processing of information related to the virus, because it is necessary to have systems that allow us to obtain truthful, summarized and useful information from those channels where the disease is reported and talked about.

One of these channels is Twitter, where there are millions of people talking about COVID-19, associated pathologies, virus mitigation measures, prevention measures or means of propagation. Being able to process this data correctly involves dealing with a lot of irrelevant information. Currently, the tweet dataset (Lamsal, 2020) related to COVID includes more than 700 million entries, which makes it a perfect candidate for testing our algorithm.

Our problem, uses a part of that dataset, specifically the tweets of the first week of the pandemic, which goes from March 20, 2020 01:37 AM to March 26, 2020 12:46 PM. The total number of tweets that have been taken into consideration for the experiment is 7 293 933 with 34 features for each of the tweets.

4.1.2. November 2020 US elections Dataset

The elections on 3rd November 2020, pitted Democratic candidate Joe Biden against Republican, Donald Trump. During the days leading up to the election and up to election day, using Twitter's streaming API, we obtained a database of tweets. To filter the tweets related to the election, we saved those that used any of the following hashtags *#election2020*, *#november3*, *#2020election*, *#vote2020*, *#votebidenharris*, *#votedonaldtrump*, *#biden2020*, *#trump2020*, *#democrats*, *#republicans* or *#election*. For this use case, we have selected a part of the complete database. Specifically, the number of tweets taken into consideration for the experiments carried out in this paper is 2 118 180. These tweets are from 28 October at 01:55 PM to 30 October at 4:44 PM.

4.2. A use case in Big Tweet Datasets

Lets assume that we need to apply techniques such as clustering, association rules or LDA to obtain valuable information about our dataset. We know that much of the content in social networks is noisy, data without value. We are also aware of the volatility and speed of data generation in social networks. We see that the number of tweets generated on a topic in a day exceeds the million. If we extrapolate this to a week or a month, the volume of data begins

to be unmanageable and, in addition, the vast majority of these data will have no value for our analysis. This is where the NOFACE framework comes in, allowing the reduction of the data keeping only what really adds value to our analysis.

The objective of our use case is (among others) to apply data mining to obtain valuable information about COVID-19 or elections. It is about obtaining, for example, the most representative topics regarding virus containment measures, in the case of `COVID-19`, or clusters of tweets from independent, non-party biased sources of information in the case of `Elections`. The experimental design will deal with the comparison of results with the application of the NOFACE framework proposed in this paper and without its application. Therefore, the application of the framework in these use cases is intended:

- To reduce the amount of data thus favouring the computation time of the subsequent data mining algorithms.
- To help reduce the irrelevant information present on social media by maintaining only those instances with a high reputation and relation to the domain of the problem.
- To demonstrate that the content reduction of the algorithm improves the results in the subsequent data mining processes, in this case, clustering, association rules and LDA.
- To demostrate that the NOFACE framework obtains topic relevant accounts.

4.2.1. Robustness checks

For the robustness checks of the system and to verify that the filters work properly, we have checked various factors inherent to the algorithm such as time, percentage of content reduction, and localised expert and reliable users. These concepts are linked to the proper functioning of the algorithm, since what the algorithm seeks to do is to reduce the amount of data in subsequent analyses.

This explains why our approach obtains the same or better results in subsequent data mining processes.

For the evaluation of the results and improvements provided by NOFACE in conjunction with other data mining techniques, specific methods are used for each evaluation. Specifically the coherence in LDA (Röder et al., 2015) and the silhouette coefficient in clustering (Rousseeuw, 1987). The silhouette coefficient in clustering measures how well defined the clusters are, while the coherence coefficient in LDA measures the relationship of terms within the same topic. We consider that these two robustness measures are the most suitable for our experimentation since they are the most widespread in the literature with respect to clustering and LDA. Regarding association rules we will rely on the number of obtained rules for a given confidence threshold, as well as the time to create transactions and obtain rules.

In addition, a graphical interpretation of the results is carried out using visualisation techniques. The interpretation of the visualisation and the results will be detailed in the next section.

4.2.2. Experimental results

The framework starts taking as input a set of words some related to health or to independent journalists, depending on the dataset used. The exact words are: *doctor, medical, researcher, medicine, epidemiologist* and *clinical* in `COVID-19` use case, and *communicator, nonprofit, truth, journalist* and *analyst* in the `Elections` use case. The mean score for passing the engagement filter has been set at 4 and 6, in the case of `Elections` and `COVID-19` respectively, whilst for passing the credibility filter the mean of the values has been 0.01 in both cases. These thresholds are defined by the mean engagement value obtained by the engagement analysis module and the mean credibility value obtained by the credibility analysis module of the NOFACE framework as explained in Section 3.3 and Section 3.4. The results in terms of execution time, located users, final dataset size and percentage of content reduction can be seen in Table 4. Looking in detail Table 4, we have a comparison between the results when

Dataset	Configuration	Elapsed time	Located Users	Original size	Final size	% reduction
COVID-19	pre-processing	[29min 27s - 32min 54s]	1 047 496	7 293 933	1 652 975	77%
Elections	pre-processing	[5min 49s - 6min 41s]	388 688	2 118 180	568 640	73%
COVID-19	pre-processing + NOFACE	[58min 3s - 1h 1s]	[556 - 758]	7 293 933	[3 050 - 4 162]	**99%**
Elections	pre-processing + NOFACE	[17min 33s - 23min 32s]	[995 - 1 026]	2 118 180	[4 108 - 4 241]	**99%**

Table 4: Results and intervals for each configuration and dataset.

applying the pre-processing seen in Section 3.1 and the results when applying this pre-processing in conjunction with the NOFACE framework. The times are longer in the latter case, since we add one more layer of pre-processing, namely the NOFACE filter.

Looking at the content reduction, we see that in the cases where NOFACE is applied, the reduction is 99%. While in the cases in which the pre-processing is simply applied, we have a reduction of 77% which corresponds to the cleaning of retweets or tweets composed only of empty words, links or mentions.

We can conclude that the objective of content reduction is achieved and, moreover, the accounts that the algorithm considers relevant and interesting are actually correlated with the domain under study. Figures 2 and 3 contain some of the accounts that have been considered as relevant by the algorithm, showing that they are profiles with a great reputation in relation to the subject matter of health or journalism. According to `COVID-19` use case, it is interesting to mention, that as shown in the picture, the profiles, except for the first one, contain words related to health not introduced in the first step of the process. The algorithm, using word embeddings, has learned how relevant they are to the topic and therefore takes them into consideration to enrich the search space. On the other hand, if we look at the use case of `Elections`, it is very interesting to see how a large number of accounts selected by the algorithm as truthful are accounts that Twitter has already verified, which is a great quality marker for the accounts filtered by the NOFACE framework.

Regarding execution times, the complete NOFACE framework always takes values ranging from 17 minutes to 1 h. On the other hand, the standard pre-processing applied to the control dataset takes 5 to 32 minutes to complete.

Figure 2: Some anonymised profiles selected by NOFACE in the COVID-19 dataset.

Although it may seem that the time is slower with the NOFACE framework, we must considered that in the later stages that data mining algorithms could be applied (LDA, association rules and clustering in our case) we will have less amount of data to process, so the complete processing pipeline will take less time using the NOFACE framework. Table 5 shows the results in terms of time when the NOFACE framework is applied and not (where only pre-processing has been applied).

In the case of clustering, we can see very similar results to those seen in Table 4. In this case, when using NOFACE the time increase is in the order of milliseconds. Thus, the improvement is not very evident in the case of clustering, since the clustering algorithm used does not spend much time in the case of

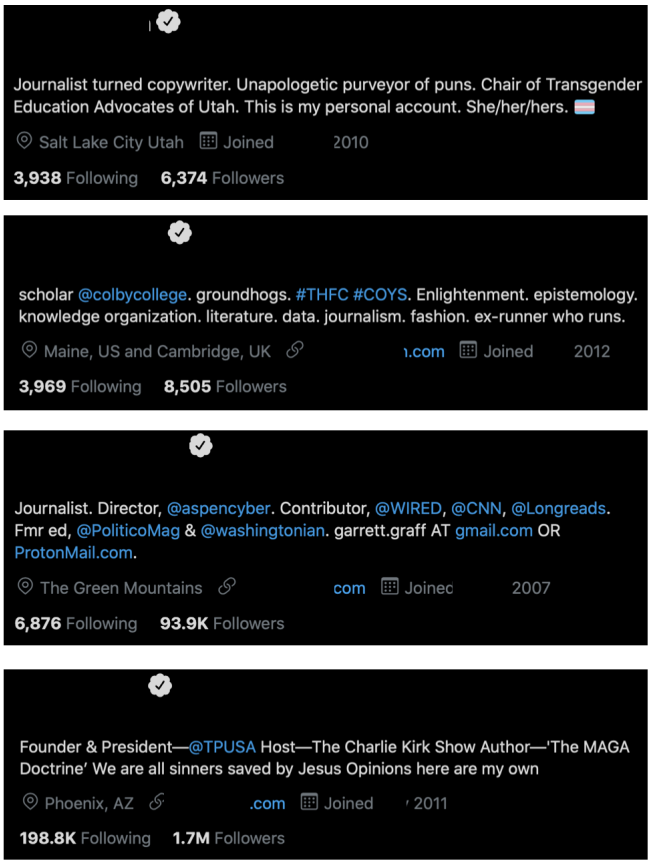

Figure 3: Some anonymised profiles selected by NOFACE in the elections dataset.

pre-processing either. Even so, in this case of using clustering, executions on the processed text took around 100 to 300 milliseconds, while in the case of unprocessed text, we are dealing with the range of 5 to 10 seconds of execution. The reduction in proportion is considerable, much more if we extrapolate it to a problem with larger datasets. At this point, it is necessary to note that the clustering algorithm has a parameter which is the number of features it will use. In all experiments and configurations, this number of features is set to 50 000. Therefore, in both cases (with NOFACE and without NOFACE) only a part of the data is taken into account to perform the clustering computation.

In the case of obtaining topics with LDA, we can see an improvement in the

Configuration \Dataset	COVID-19	Elections
Pre-processing	[29min 27s - 32min 54s]	[5min 49s - 6min 41s]
Pre-processing + Clustering	**[30min 24 - 33min 58s]**	**[6min 7s - 7min 1s]**
Pre-processing + LDA	[1h 34min 15s - 1 h 37min 32s]	[25 min 36s - 27min 39s]
Pre-processing + Apriori	[-]	[-]
Pre-processing + NOFACE	[58min 3s - 1h]	[17min 33s - 23min 32s]
Pre-processing + NOFACE + Clustering	[58min 4s - 1h 1s]	[17min 32s - 23min 33s]
Pre-processing + NOFACE + LDA	**[58min 15s - 1h 16s]**	**[17min 55s - 24min 1s]**
Pre-processing + NOFACE + Apriori	**[58min 42s - 1h 42s]**	**[18min 22s - 24min 30s]**

Table 5: Elapsed time of the experiments with and without NOFACE.

total execution pipeline time. The ranges of pre-processing, applying NOFACE and LDA will always be lower than those of pre-processing and applying LDA directly. The reduction and cleaning of input data to the LDA algorithm performed by the NOFACE framework leads to a reduction of the total execution time of more than 30 minutes in some cases. In Table 5 we can see that in both use cases, the best execution times for LDA are obtained when using the NOFACE filter.

One of the most interesting results can be obtained with regard to association rules mining. We have employed one of the most used algorithms for mining association rules, called Apriori. In this case, the algorithm cannot finish the execution due to the high amount of items and transactions to process. On the other hand, when filtering with the NOFACE framework, we obtain the rules in just a few seconds. This highlights the value of these filtering techniques for the use of subsequent algorithms in which the volume of data is a serious problem, like in the case of the Apriori algorithm.

Taking into consideration these results we can conclude that although a priori the NOFACE framework takes more time, if we considered the complete data processing pipeline, that is, in conjunction with other data mining techniques that could be interesting to apply for a complete analysis, it improves considerably the execution times.

4.2.3. Clustering results

Clustering, as far as texts are concerned, tries to find which documents are more similar to others, by placing them in the same cluster. In the Twitter domain, it tries to find out which tweets are more similar to others in terms of content, which has great implications in the process of summarising information, searching for influencers or categorising accounts, for example. Since it is one of the techniques widely used in text mining, we are going to apply K-means on the dataset filtered with NOFACE and the complete dataset. The characteristics that fed the clustering algorithm, correspond to a TF-IDF vectorization of the document. To choose the number of clusters to search for, we have carried out an analysis using the sum of quadratic error (SSE). The value of clusters (k) used for experimentation, given by the SSE value, was 11 in the `Elections` use case and 13 in the `COVID-19` use case.

To graphically compare the obtained results we have represented them by means of a t-distributed stochastic neighbour embedding (TSNE) graph (Maaten & Hinton, 2008).

In Figure 4 and Figure 6 the results of applying the clustering algorithm without filtering are shown and in Figure 5 and Figure 7, the results in the case of filtering using NOFACE.

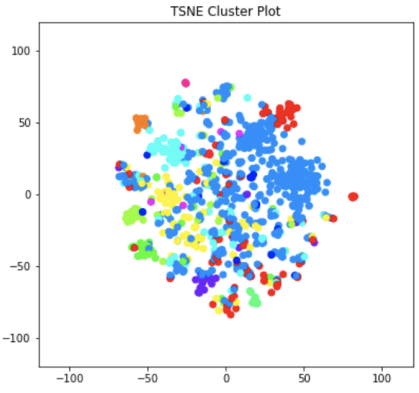

Figure 4: COVID-19 use case: TSNE plot without NOFACE

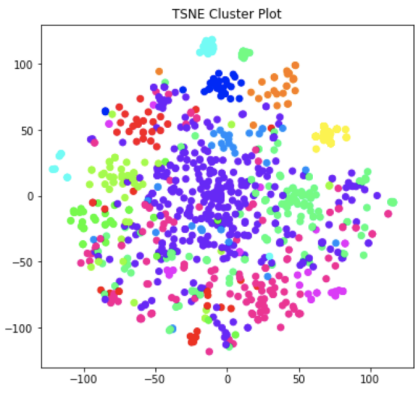

Figure 5: COVID-19 use case: TSNE plot with NOFACE

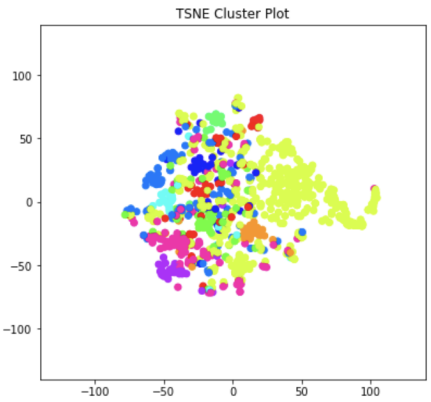

Figure 6: Elections use case: TSNE plot without NOFACE

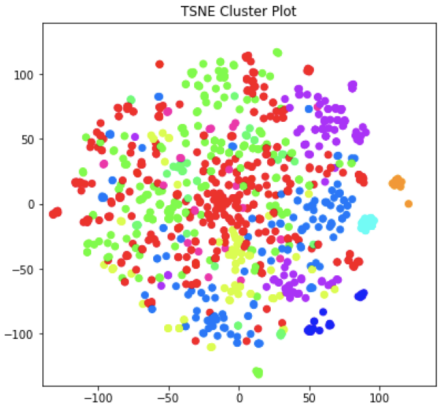

Figure 7: Elections use case: TSNE plot with NOFACE

One of the first things we can observe in the TSNE graph is that in the case of the NOFACE results, we have more dispersion between the clusters. This is a clear symptom that good accounts have been selected in which the features are very differentiated. In the case of not applying NOFACE, we have more overlap between clusters, and the silhouette coefficient is of a worse degree.

Following with the analysis of the TSNE graph for the COVID-19 use case, in Figure 4, the blue cluster is very dispersed over the whole area of the graph, while in Figure 5 we can see that the majority cluster (in this case the purple one) is quite well defined, as also are the green, light green, red, magenta, dark blue, light blue, yellow, pink and orange ones. In this way, we can see how the application of NOFACE, has greatly improved the execution of the clustering algorithm, because in Figure 4 it is complicated to identify more than 6 clusters (yellow, red, light blue, orange, red and magenta). This visual analysis is also corroborated by the calculation of the silhouette coefficient. Specifically, in the case of COVID-19, the silhouette coefficient is 0.0095 in the case of using the NOFACE framework and 0.0069 in the case of not using it. This coefficient gives us a value on how the clusters are differentiated from each other. This indicates that by applying the NOFACE framework, we have more

differentiated and higher quality clusters. Although the improvement is not of a high degree, we can conclude that we have managed to reduce the dataset to one of a better quality with less data, so less computation time and easier interpretation improve the results.

In the case of `Elections` an improvement in the number of identifiable clusters can also be observed, although in this case this improvement is less evident than in the case of `COVID-19`. In this case, we can conclude that with such similar results, there is no significant loss of information when applying the NOFACE framework. It is also necessary to point out that being a political dataset and already filtered by hashtags related to the elections, the dataset will contain very polarised opinions based on ideology. This fits perfectly with the background of clustering, so as we can see in the results, in both cases the outcomes are generally strong. On the other hand, Figure 7 shows how applying NOFACE makes the clusters more differentiated from each other. This analysis is reinforced by the silhouette coefficient, which is 0.12 when using NOFACE and 0.01 when not using it. Again, we see how the clustering algorithm offers better results if used in conjunction with the NOFACE filtering framework.

Another advantage that can be gained from using NOFACE is that by generating more cohesive clusters with less data, better content-based labelling of clusters can be carried out so that these can be used as classes to be applied to possible supervised classification problems.

4.2.4. LDA results

The LDA process seeks to obtain those topics that are being talked about in social networks. In our case, we tried to seek what are the main general topics about COVID-19 and elections in Twitter.

For example, a scenario for the application of the LDA on `COVID-19` would be to see if there is any kind of contradictory information, or relevant information for COVID-19 measures. In the case of the `Elections`, for example, it might be interesting to get related topics by state, to see what people are concerned in one state or another, or what is being discussed in the independent press. The

process of obtaining topics takes the information from a bag of words generated from the text of the tweets. The number of topics was set to 6.

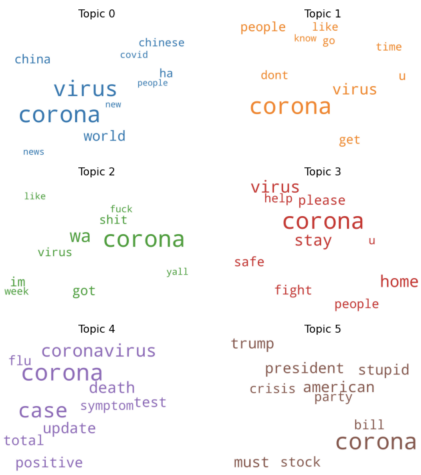

Figure 8: Topics from the COVID-19 dataset processed without NOFACE

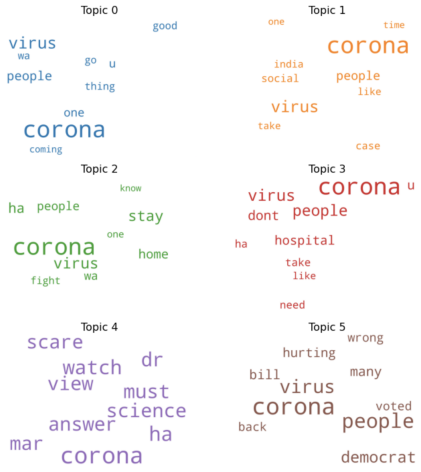

Figure 9: Topics from the COVID-19 dataset processed with NOFACE

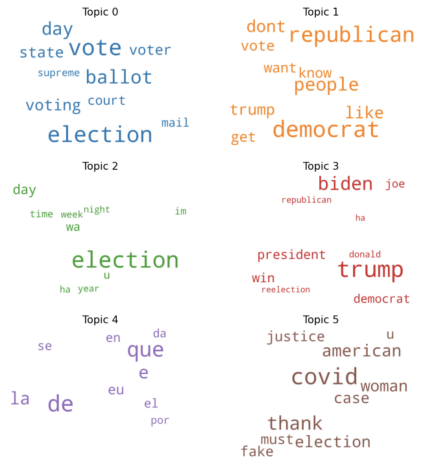

Figure 10: Topics from the elections dataset processed without NOFACE

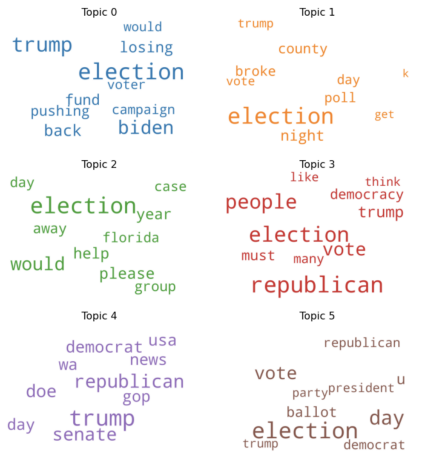

Figure 11: Topics from the elections dataset processed with NOFACE

In Figure 8 and 10 we can see the most representative words related to

the 6 topics obtained by the LDA algorithm on the dataset that does not use NOFACE. On the other hand, Figure 9 and 11 show the most representative topics obtained over the set of tweets filtered using NOFACE.

According to the figures, we can see how both outputs of the LDA algorithm contain very similar information, which shows that the NOFACE framework did not lose information and can therefore be very useful to keep those tweets and accounts that really add value. This analysis can be supported by coherence results, which give a value of 0.305 for COVID-19 using NOFACE, and 0.271 without using NOFACE. In the case of Elections, the improvement is less evident and the coherence results hardly fluctuate from one experiment to another. This leads us to reinforce the conclusion that there is no loss of important information, while there is a reduction of invaluable information and noise. Our analysis is focused on documents (tweets). Therefore, we have also experimented with the alpha hyper-parameter of the LDA algorithm, which can adjust sensitivity with respect to document-topic density or document-topic distribution. Specifically, we have used the symmetric and asymmetric values for each experiment. In the case of COVID, the best result in terms of coherence is obtained for the filtered dataset, with asymmetric alpha, obtaining a coherence value of 0.365. In the Elections dataset, the results are similar with both configurations.

In a more subjective analysis, we could even see words with more sense and relation with the COVID-19 or elections in Figure 9 and 11 respectively. For example in the Election use case of topic 2, the text filtered by NOFACE (Figure 11) contains related content about the state of Florida and the Democrat and Republican parties, which was a disputed and swing state until the last moment.

Finally, we have also added a display layer using graphics of topics according to (Chuang et al., 2012). This graph is useful to show how the topics would be distributed in a 2D graph using principal components. Figures 12 and 14 show the Intertopic Distance of results without applying NOFACE, whilst 13 and 15 are the results using the NOFACE framework. In the case of `COVID-19`, we see that there are clearly better results in the case of filtering using NOFACE, since

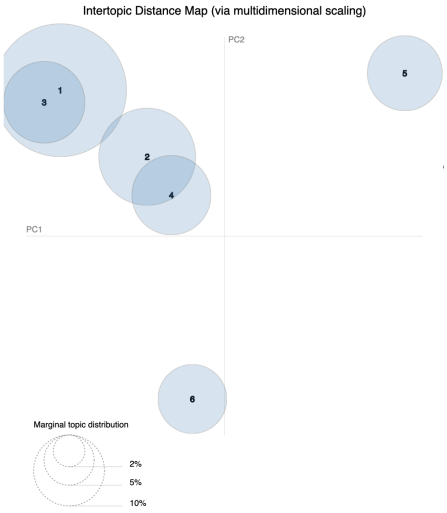

Figure 12: Intertopic Distance Maps from the COVID-19 dataset processed without NOFACE

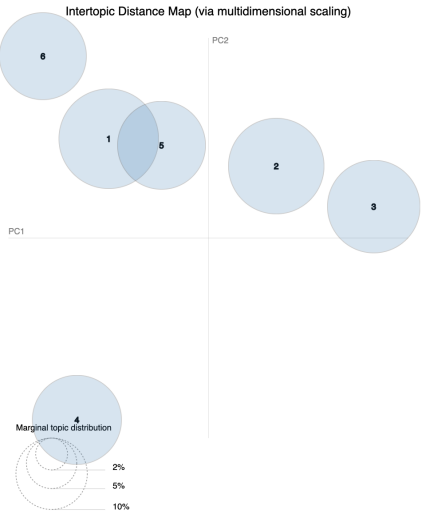

Figure 13: Intertopic Distance Map from the COVID-19 dataset processed with NOFACE

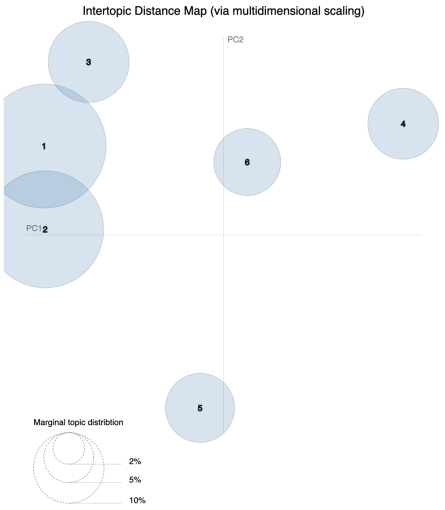

Figure 14: Intertopic Distance Maps from the elections dataset processed without NOFACE

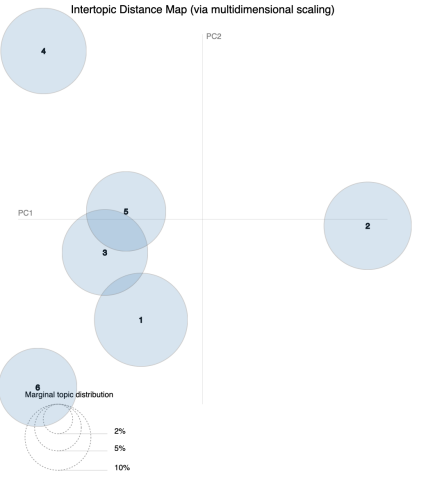

Figure 15: Intertopic Distance Map from the elections dataset processed with NOFACE

we have more differentiated and dispersed topics, i.e. we have less intratopic overlap. Another analysis that can be distilled from the graphs is how the topics

(circle size) are more homogeneous in the case of the NOFACE filtered datasets. This indicates that there is a better distribution of words between topics in this case, than if we compare it with the use cases without using NOFACE. On the other hand, in the `Elections` case we have a very similar overlapping and situation of the topics. Again we have a similar situation to the one we had in clustering. In this case we have obtained very similar results processing [4 108 - 4 241] accounts selected by NOFACE instead of 388 688 with traditional pre-processing. This brings a remarkable improvement in terms of performance, maintainability and analysis capability.

4.2.5. Apriori algorithm results

In the case of the association rules we cannot compare the results of both experiments because in the case of the unfiltered dataset the algorithm did not finish, as the explosion of frequent itemsets combinations is too high. In order to have some kind of comparison, we have carried out the experimentation by obtaining a random sample of the unfiltered dataset of the same size as the one resulting after filtering the dataset with NOFACE. Support values of 0.1 and 0.01 have been taken, for a confidence value of 0.6. For the support threshold value of 0.01 using the NOFACE filter, a total of 130 rules were obtained using the `COVID-19` dataset and 33 rules in the `Elections`. On the other hand, for the sample of the unfiltered dataset, this number is reduced to 105 in the case of `COVID-19` and 9 in the case of `Elections`. For the value of 0.1 for minimum support, only 1 rule was obtained in all cases except in the case of the random sample of the unfiltered `Elections` dataset where no rules were obtained. These results are interesting because they demonstrate how the filtering produces a more cohesive dataset, since the support value is a direct indicator of the co-occurrence of items in a dataset, and getting more rules in the case of applying NOFACE indicates that certain terms are more likely to appear together.

5. Discussion

After the experiment and analysis of the different use cases, it is necessary to review the results and check if the considered objectives set (Section 4.2) have been achieved.

One of the goals was to eliminate irrelevant content from the dataset. At this point, we can conclude that the NOFACE framework complies perfectly, perhaps even being too restrictive. As we have seen throughout the previous section, it reduces the content coming from Twitter to a great extent, in addition, it limits it to the scope of research that could be desired at a certain moment, health or journalism in our case. In other words, the framework, takes benefit of credibility, engagement and what is more important the user's experience in a domain, offering very representative content and profiles. The reduction in number of examples is very large, going from millions of tweets to just a few thousand. However, we must bear in mind that this reduction in terms of examples is appropriate for Big Data environments in social networks, where many of the content is noise or content generated by accounts with no value for the topic in question. Therefore, if we analyse the result from the point of view of users with experience for the topic, and in a period of 2 days, the algorithm detects from 500 to 1 000 relevant users among the available 10 000 users. Thus, analysing the content of those 1 000 users with experience in the topic will be more efficient, than analysing 10 0000 who have simply given their opinion on the topic.

Another objective was to compare the execution times and to check if the framework introduces any computational improvements. In this case, there are conflicting results, because in the case of LDA and association rules, the framework obviously improves the execution times. In the case of clustering, the extra time involved in running the NOFACE framework makes it behave worse in terms of time. In this case, it is necessary to mention that both the standard pre-processing experiment and the NOFACE experiment employed 50 000 characteristics for the TF-IDF vector. Therefore this result is biased by

this value, that at this point is a dimensionality reduction based on the characteristics of the TF-IDF. This means that at this point for both experiments (with and without NOFACE), the feature dataset used is considerably reduced, since the TF-IDF discards many features (words). Thus, we are not using the full dataset in the case of the experiment without NOFACE, just a selection of the best textual features guided by the TF-IDF. This makes the results more similar in terms of elapsed time. Even so, the clustering algorithm takes about 0,001 seconds to finish with the dataset processed with NOFACE, and from 19 to 64 seconds with the complete dataset, so the improvement achieved by the proposed filtering method is still evident.

The last goal was to demonstrate that the results of other data mining techniques on a dataset processed with NOFACE, would improve against a dataset that had not been processed with our framework. Throughout this section, we have seen how in the worst case, the result is very similar, which indicates that the framework really selects good profiles whose information is relevant, that is, there is no loss of relevant information. In other cases, we have seen how the framework improves considerably, as in the case of the clustering algorithm on the `COVID-19` dataset, where cohesive and differentiated clusters with fewer outliers were found. Finally, in the case of association rules, we have seen how the filter can help certain algorithms, sensitive to the amount of data, to function normally, improving the results in terms of obtained rules.

5.1. Challenges

One of the most important challenges of the NOFACE framework, as well as other similar systems, involves dealing with lies and noise in social media. We must bear in mind that nothing prevents a person from describing themselves in their biography as a doctor, researcher or engineer without actually being one. This makes the system very sensitive to these issues, and it is therefore a challenge to design an automatic system that is capable of discerning between real people and those who are not. The system proposed in this paper, as well as other proposals presented in the literature, offers more layers of anal-

ysis (engagement and credibility) trying to mitigate this problem, having as a premise that probably real Twitter followers will not share content from people of dubious biography or belonging.

Another of the great challenges to be faced by content-based systems and not by graph-based systems, as proposed in this paper, is that they can detect a fake influencer as relevant. The fake influencers are behind accounts in social networks with great statistics of interaction and content generation. This can lead a brand to think that it can be a great investment to hire this account to spread their products or services, but in reality it would be a waste of money because these accounts really generate interaction with accounts managed by bots and other non-real accounts, so there is no real interaction. Detecting these accounts requires network analysis, and according to (Tsapatsoulis et al., 2019), they are usually egocentric accounts easily identified by network analysis algorithms based on centrality.

5.2. Contributions to literature

The main contribution of this paper to the literature has been the creation of a framework for the selection of relevant content and users in social networks. Throughout the paper it has become clear that social networks play an irreplaceable role in our daily lives, and in particular in many business processes. The review (Kumar et al., 2021) shows how users use social networks to inform themselves about products and services of various kinds. It also mentions several studies on how companies of different sizes use social networks to obtain information from users in numerous aspects (Chatterjee & Kar, 2020). It is in these points, where our framework takes special relevance and can help companies or individuals to filter the content of social networks to favour their subsequent stages of data analysis. With this filtering, the algorithm also achieves a reduction of the dataset to be processed.

Another contribution to the state of the art relates to misinformation. By selecting credible users, with experience in the sector and with a certain impact, we are also ruling out the misinformation component to a certain extent. There-

fore the proposed framework helps to eliminate the misinformation present in social networks. In this sense, we are carrying out an elimination through user features (favourites, retweets) and information obtained from natural language processing of their biographies. Currently, there are other models specially designed for these tasks that use classification algorithms or deep learning models (Mahir et al., 2019) such as recurrent neural network models and LSTM, to classify whether something is misinformation or not. As we have seen throughout the paper, these models need prior training, something that makes them sensitive to changes. So we find that there is a need for systems, such as the one proposed throughout this paper or others in the literature, based on content features (Wu et al., 2016) and filters that allow to narrow down the amount of false content in a way that does not require large databases and training.

Finally, three comprehensive use cases of data mining in conjunction with the NOFACE framework have been provided to the literature. In the case of the paper, clustering, association rules and LDA have been used. It has been demonstrated how the framework improves clustering results in terms of silhouette and cohesion of the clusters. As for LDA, the topics obtained are of better quality in terms of coherence. The literature (Joung & Kim, 2021) highlights the need for these pre-processing techniques to improve the performance of algorithms such as topical detection algorithms (LDA). Also, in this paper it has been highlighted how the use of efficient pre-processing can help to improve the execution times of a complete data mining pipeline. It has been shown that the filter can be of special interest in those algorithms where the number of data can make them fail or the execution time is very inefficient, as in the case of the Apriori algorithm (Al-Maolegi & Arkok, 2014).

6. Conclusion and future work

The present work has proposed a new framework for filtering irrelevant content on social media, demonstrating its usefulness as a technique for preprocessing data before applying other data mining techniques such as cluster-

ing, association rules or LDA. Two of the most widespread robustness metrics in these techniques (silhouette and coherence) have been used on the datasets filtered by the NOFACE framework. Based on these metrics, it has been shown that the framework does not lose information and improves the results obtained. Additionally the proposed framework can be used with a wide range of data mining algorithms, being specially appropriate on those that may be limited by data size and those that may be sensitive to noise for a given type of analysis.

During the development and research, a study of the state of the art in the subject has been carried out. The use of advanced text mining techniques based on word embeddings has also been highlighted. This is, as far as we know, the first contribution that applies these techniques to compute expertise on a topic.

Also, the potential of the framework has been highlighted on two real problems of tweets relating to COVID-19 and the 2020 US elections, on which the consequent reduction of number of examples without loss of information has been demonstrated, even improving the results obtained using the complete dataset. In short, the paper:

1. Offers a new framework for irrelevant content reduction in social networks based on iterative filters. It also helps to reduce misinformation as it is usually issued by inexperienced or low credibility users in a particular sector, being these discarded by NOFACE. Iterative filters address the problem in a very strict way and can alleviate the problem of lies about professional experience on social networks.
2. It introduces an algorithm for locating experts in social networks through the use of word embedding. It has been demonstrated that the algorithm is feasible and can be used as a pre-processing step prior to other data mining applications.
3. It proposes an interpretable and easily understandable solution to the problem of detecting user-generated content useful for a given topic or analysis.
4. Provides two detailed interpreted use cases to support the use of the frame-

work in conjunction with other data mining tasks. In these use cases it has been demonstrated that there is no loss of information and improved results in terms of computation time and robustness.

The proposed framework opens up future channels of development that are closely linked to the challenges seen in Section 5, like the study of how sensitive is the system towards lies and egocentric networks generated by fake influencers or false credibility, so being able to identify these issues would considerably improve the system. It is also necessary to mention the opposite case to the one described above, since maybe the system is not considering users who are very influential in their field, but who have hardly any presence in social networks. That is, the framework in its current state is very restrictive, so being able to locate the low statistical but really good accounts would be a great improvement and a future path of development and research as well.

Although two unsupervised use cases have been provided in the use cases, the framework could also be used in supervised methods. A possible future application in this sense would be to filter a large dataset of topic-related data into useful or truthful information and non-relevant or fake information using the NOFACE framework. Then, using these resulting labelled datasets to train a classifier, for example based on deep learning, that allows us to determine whether a new tweet is truthful or not.

Finally, there is the possibility of extending the system to a purely streaming environment, where related words could be mutated by time windows and a list of expert users would be maintained over time, who could cease to be experts if their engagement or credibility levels drop.

Acknowledgment

Funding for open access charge: Universidad de Granada / CBUA. The research reported in this paper was partially supported by the COPKIT project under the European Union's Horizon 2020 research and innovation program (grant agreement No 786687), the Andalusian government and the FEDER

operative program under the project BigDataMed (P18-RT-2947 and B-TIC-145-UGR18). The paper is part of the NOFACEPS project (PPJIB2021-04) of the University of Granada's internal plan. Finally the project is also partially supported by the Spanish Ministry of Education, Culture and Sport (FPU18/00150).

References

References

Abu-Salih, B., Wongthongtham, P., Chan, K. Y., & Zhu, D. (2019). Credsat: Credibility ranking of users in big social data incorporating semantic analysis and temporal factor. *Journal of Information Science*, *45*, 259–280.

Agrawal, R., Srikant, R. et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (pp. 487–499). Citeseer volume 1215.

Al-Maolegi, M., & Arkok, B. (2014). An improved apriori algorithm for association rules. *arXiv preprint arXiv:1403.3948*, .

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*, 211–36.

Alrubaian, M., AL-Qurishi, M., Alrakhami, M., Hassan, M., & Alamri, A. (2016). Reputation-based credibility analysis of twitter social network users: Reputation-based credibility analysis of twitter social network users. *Concurrency and Computation: Practice and Experience*, *29*. doi:10.1002/cpe.3873.

Alrubaian, M., Al-Qurishi, M., Hassan, M. M., & Alamri, A. (2018). A credibility analysis system for assessing information on twitter. *IEEE Transactions on Dependable and Secure Computing*, *15*, 661–674. doi:10.1109/TDSC.2016.2602338.

Aswani, R., Kar, A. K., & Ilavarasan, P. V. (2019). Experience: managing misinformation in social media—insights for policymakers from twitter analytics. *Journal of Data and Information Quality (JDIQ)*, *12*, 1–18.

Barbado, R., Araque, O., & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, *56*, 1234–1244.

Batra, J., Jain, R., Tikkiwal, V. A., & Chakraborty, A. (2021). A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques. *International Journal of Information Management Data Insights*, *1*, 100006.

Baum, A. (2019). Scraping Twitter User Data Using Google and Tweepy. `https://towardsdatascience.com/use-google-and-tweepy-to-build-a-dataset-of-twitter-users-cbfd556493a9`. [Online; accessed 18-January-2020].

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, .

Canini, K. R., Suh, B., & Pirolli, P. L. (2011). Finding credible information sources in social networks based on content and social structure. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 1–8). IEEE.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675–684).

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*, 16–28.

Chatterjee, S., & Kar, A. K. (2020). Why do small and medium enterprises use social media marketing and what is the impact: Empirical insights from india. *International Journal of Information Management*, *53*, 102103.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 74–77).

Cordeiro, P. R. D., Pinheiro, V., Moreira, R., Carvalho, C., & Freire, L. (2019). What is real or fake?-machine learning approaches for rumor verification using stance classification. In *IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 429–432).

Diaz, F., Mitra, B., & Craswell, N. (2016). Query expansion with locally-trained word embeddings. `arXiv:1605.07891`.

Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2022). NOFACEPS source code and data repository. URL: `https://github.com/ugritlab/NOFACEPS`.

Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., & Gummadi, K. (2012). Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 575–590).

Hassan, D. (2018). A text mining approach for evaluating event credibility on twitter. In *2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* (pp. 171–174). IEEE.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet*, *395*, 497–506.

Joung, J., & Kim, H. M. (2021). Automated keyword filtering in latent dirichlet allocation for identifying product attributes from online reviews. *Journal of Mechanical Design*, *143*.

Kaliyar, R. K. (2018). Fake news detection using a deep neural network. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)* (pp. 1–7). IEEE.

Kang, B., O'Donovan, J., & Höllerer, T. (2012). Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (pp. 179–188).

Kar, A. K., & Aswani, R. (2021). How to differentiate propagators of information and misinformation–insights from social media analytics based on bio-inspired computing. *Journal of Information and Optimization Sciences*, *42*, 1307–1335.

Khoo, L. M. S., Chieu, H. L., Qian, Z., & Jiang, J. (2020). Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 8783–8790). volume 34.

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights*, *1*, 100008.

Kumari, R., Ashok, N., Ghosal, T., & Ekbal, A. (2021). Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. *Information Processing & Management*, *58*, 102631.

Kuzi, S., Shtok, A., & Kurland, O. (2016). Query expansion using word embeddings. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 1929–1932).

Lamsal, R. (2020). Coronavirus (covid-19) tweets dataset. URL: `https://dx.doi.org/10.21227/781w-ef42`. doi:`10.21227/781w-ef42`.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 302–308).

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). Topical word embeddings. In *Twenty-ninth AAAI Conference on Artificial Intelligence*. Citeseer.

Liu, Y., & Wu, Y.-F. B. (2020). Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, *38*, 1–33.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks, .

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (pp. 281–297). Oakland, CA, USA volume 1.

Mahir, E. M., Akhter, S., Huq, M. R. et al. (2019). Detecting fake news using machine learning and deep learning algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1–5). IEEE.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. `arXiv:1301.3781`.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Molina-Solana, M., Amador Diaz Lopez, J., & Gomez, J. (2018). Deep learning for fake news classification. In *I Workshop in Deep Learning, 2018 conference spanish association of artificial intelligence* (pp. 1197–1201).

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. arXiv:1902.06673.

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, *1*, 100007.

Oehmichen, A., Hua, K., Amador Dííaz López, J., Molina-Solana, M., Gómez-Romero, J., & Guo, Y. (2019). Not all lies are equal. a study into the engineering of political misinformation in the 2016 us presidential election. *IEEE Access*, *7*, 126305–126314. doi:10.1109/ACCESS.2019.2938389.

Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., & Kittler, J. (2010). A review of instance selection methods. *Artificial Intelligence Review*, *34*, 133–143.

Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and its Applications*, *540*, 123174.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). URL: http://www.aclweb.org/anthology/D14-1162.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399–408).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Roy, D., Paul, D., Mitra, M., & Garain, U. (2016). Using word embeddings for automatic query expansion. arXiv:1606.07608.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*, 22–36.

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, *53*, 907–948.

Tsapatsoulis, N., Anastasopoulou, V., & Ntalianis, K. (2019). The central community of twitter ego-networks as a means for fake influencer detection. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)* (pp. 177–184). IEEE.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, *21*, 80–90.

Wu, S., Liu, Q., Liu, Y., Wang, L., & Tan, T. (2016). Information credibility evaluation on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 30.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L. et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*, 270–273.

3.2 A Comparative Study of Word Embeddings for the Construction of a Social Media Expert Filter

- Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2021, September). A Comparative Study of Word Embeddings for the Construction of a Social Media Expert Filter. In International Conference on Flexible Query Answering Systems (pp. 196-208). Cham: Springer International Publishing.
  - Conference: International Conference on Flexible Query Answering Systems. FQAS2021, Bratislava.
  - Status: Published.

A comparative study of word embeddings for the construction of a social media expert filter

J. Angel Diaz-Garcia[ab] and M. Dolores Ruiz[ac] and Maria J. Martin-Bautista[ad]
[a]Department of Computer Science and A.I., University Of Granada, Daniel Saucedo Aranda, s/n, 18014 Granada
[b] joseangeldiazg@ugr.es
[c] mdruiz@decsai.ugr.es
[d] mbautis@decsai.ugr.es

Abstract

With the proliferation of fake news and misinformation on social media, being able to differentiate a reliable source of information has become increasingly important. In this paper we present a new algorithm for filtering expert users in social networks according to a certain topic under study. For the algorithm fine-tuning, a comparative study of results according to different word embeddings as well as different representation models, such as Skip-Gram and CBOW, is provided alongside the paper.

Keywords: Word Embeddings, Preprocessing, Expertise, Social Media Mining

1 Introduction

Nowadays, the world could not be understood without social networks. They have become an irreplaceable source of information, and are used daily by millions of people to obtain information on a wide range of topics such as investments, what their friends have done, travelling, etc. This great success has led to social networks also being used for dubious moral purposes, such as the spread of misinformation to influence various factors, such as the opinions of other users. Identifying this fake information, or information of poor value, is therefore a very important task.

In this paper, we present a new algorithm for preprocessing a Twitter dataset according to the user expertise on a given topic under study. The algorithm is tested with a dataset [9] of Tweets related to COVID-19 [17], [7], composed of millions of Tweets and with a large amount of noise, low-value or false information. Our topic of study, therefore, will be medicine, and on this experimental dataset the main objectives of our algorithm will be:

- Discover which users are relevant on the social network according to their Tweets about a topic.
- Dimensionality reduction of a large dataset by keeping only the information relevant to a given topic.

To achieve these objectives, the system will be based on the assumption that if a user has expertise in an area, his or her content will be useful and will contain relevant to topics related to that area. To obtain which users are relevant, we will focus on Twitter biographies, on which we will train word embedding. The word embedding corresponds to the current state-of-the-art in Natural Language Processing [11], [10]. The underlying technique is to represent all the words within a given vocabulary in a vector space as vectors. With these vectors, operators such as addition or subtraction can be applied, so that the words *king - man + woman* would result in the word *queen*. In brief, if *king* and *queen* are the words and **king** and **queen** their embeddings, the distance in vector space between **king** and **queen** is a quantitative indicator of the semantic relation between *king* and *queen*. In this particular case the difference in distances would be very small because only the gender changes.

There are a multitude of models and representations for word embedding, so for fine-tuning our algorithm we have compared Word2Vec [12], [13] and FastText [4], since they are the most widespread and relevant in terms of versatility and performance at present. The main difference between Word2Vec and FastText is that the latter decomposes each of the input words in the neural network into n-grams, for example for the word *matter*, and *n=3* we would have *<ma, mat, att, tte, ter, er>* and the final representation would be the sum of the vectors associated with each n-gram. This representation is very interesting to discern out-of-vocabulary words or words with low presence in the dataset. Word2Vec takes each word as a vector, therefore does not consume as much memory and re-

sources as FastText (which for each word stores a vector per n-gram), although it is more sensitive to out-of-vocabulary words. For each of these embedding models, we have two representations, Skip-Gram and Continuous Bag of Words (CBOW). Skip-Gram tries to predict the context words surrounding the word in question, i.e. it predicts context based on a word. On the other hand, CBOW predicts a word based on the surrounding context words, i.e. it predicts a word based on the context.

The main contribution of the paper to the state-of-the-art is: the comparison and detailed study of the performance of different word embedding algorithms and their internal representations for the task of retrieving similar search terms in social networks. We also consider the best of the options found in the experimentation for the definition of an algorithm to exploit the potential of word embedding for the retrieval of experts on Twitter.

The paper is organized as follows: Next section is devoted to related work. In Section 3 we go into detail in the algorithm. In Section 4, we provide a discussion and comparison of the performance of different word embeddings in our algorithm. Finally, in Section 5 we examine the conclusions and the future work.

2 Related work

Our algorithm uses the power of semantic relations between words to increase the search space in Twitter biographies by employing word embedding techniques. Many works have demonstrated the power of word embeddings to expand search queries. In [16] Roy et al. uses the KNN algorithm on vector space generated by embbedigns to obtain which terms are most similar to others and expand the search query. With a similar point of view we find the works [5] and [8]. In [5] Diaz et al. train locally embedding, namely GloVe [14] and Word2Vec, to improve search processes in information retrieval. In a very similar way but with Word2Vec+CBOW Kuzi et al. demonstrate in [8] how document retrieval actually improves with this technique. In our algorithm, we will use this potential of word embeddings, not for retrieving documents, but for locating users that are experts on a topic. As far as we know, this is the first work that addresses and uses this option in addition to word embeddings.

With regard to the expert users retrieval, we find approaches such as those of Cognos or CredSaT. Cognos [6] offers a web solution for searching experts in a certain topic, for this, it uses Twitter lists. The lists on Twitter are user-managed lists, in which users add other users related to topics. Cognos exploits this potential, even improving the search for accounts in the native recommendation system of Twitter. The CredSat [1] approach, is a Big Data solution that takes into consideration the content and the time stamp to create a ranking of expert and influential users in the social network. It also adds a semantic analysis layer with sentiment analysis on Tweets and responses used to enrich the final corpus of experts.

Finally, it is necessary to mention the works proposed by Alrubaian et al. [2], [3]. These papers also deal with the analysis of expertise on Twitter, although they also model other concepts such as credibility or engagement. In terms of experience, the papers [2], [3] approach the problem from the point of view of user-generated content, unlike our proposal, which addresses it from the point of view of biographies, which are widely used to add information related to the work experience of the user.

3 Expertise filter algorithm

In this section we go into detail in the expertise filter algorithm (Algorithm 1). The algorithm takes as input, the directory where we find the csv files with the Tweets, a list of searching related to the topic under studying and the language we are interested in.

The first step of the algorithm is a pre-processing module, in this stage the cleaning of every Tweet is carried out. For this, the algorithm eliminates URLs, hashtags, mentions, reserved words from Twitter (RT, FAV...), emojis, smileys, numbers, additional spaces and punctuation marks. Following this, all the textual terms are turned into lowercase letters. After this, the database language has been detected. During the process the system also eliminates stop-words and all those Tweets using a non-recognised language or from a language other than the one desired by the user. Finally, any empty Tweets (composed of eliminated items in previous stages of pre-processing) are removed the biography and Tweet text are tokenized.

Then, the core of the algorithm starts to operate. Let's imagine the case that concerns us related to COVID-19, where we want to obtain experts or people related to science and medicine. We introduce a list of words related to medicine, for example we can introduce: *medical, doctor*. The algorithm will start to train a word embedding model on the biographies of a part of a data partition (one of the input csv file), and in the first iteration it will obtain the 5 most similar words to medical and doctor among the corpus itself. Namely, in this case, the most similar words to *doctor* and *medical* are *itresearcher, medicine, researcher, physician, epidemic, pediatrician, epidemiologist, pe-*

Algorithm 1: Expertise filter algorithm

Result: Dataframe with experts in the topic

```
# preprocessing, initializing the variables and
  data structures
cleaned-dataset=preprocess(dataset)
expert_set=[]
finaldataframe=pd.dataframe()
split cleaned-dataset into batches
for batch in batches do
    #we check if any user of the batch is already
      located as an expert
    if user_id in expert_set then
        Add all their Tweets to the final data
         frame and do not process them
    else
        Process the rest of the content to locate
         new experts
        Get the description tokens for each Tweet
        Train the word embedding model over the
         description tokens
        Create a data frame with the input words
         and 5 most similar words to each input
         word
        Extend the input word list with the most
         similar words in the embedding
        Locate all users who have in their
         description any of the words
        Extend the final data frame with the
         Tweets of the located experts
        Extend the expert set with the new experts
    end
end
```

diatrics, postdoctoral and *toxicologist*. The algorithm will use these 12 words, (5 similar to medical, 5 similar to doctor, besides doctor and medical), to find users whose biographies contain any of these terms and start creating the list of experts and topic-related users. In the next iteration, the set of 12 words will be used to search for their 5 similar ones, and so we will have an exponential growth of words linked by the word embedding to the domain of the problem, which will make us have a space of words very linked to the topic and which will grow intelligently to guide our algorithm.

In each iteration, the algorithm checks if any user id is already present in the expert list to avoid processing it again, since its words and content are already in the search corpus, avoiding thus additional processing. The output of the algorithm is a clean set of data in the form of a data frame ready to be processed in the following filters.

4 Word Embedding comparison

For the experimentation and choice of the best word embedding and representation for our algorithm we have selected a part of the Tweets dataset related to COVID-19 composed by 3 batches of 936.427, 1.062.900 and 1.319.912 Tweets respectively (total 3.319.239 Tweets). The Tweets correspond to the first days of the pandemic, specifically from March 23, 2020 09:11 AM to March 26, 2020 12:46 PM. The code has been developed in Python 3, with the Gensim [15] word embeddings library. Both development and experimentation have been carried out on the machine whose specifications can be found in Table 1.

Component	Features
CPU	2 GHz Intel Core i5 with 4 cores
RAM	16 GB 3733 MHz LPDDR4X
VRAM	Intel Iris Plus Graphics 1536 MB
Hard Disk	SATA SSD de 512 GB

Table 1: Machine specifications.

Regarding the embeddings parameters, it has been run with a window of 5 words, words with frequencies lower than 2 have been ignored and negative sampling of 10 has been done in the case of CBOW and a hierarchical softmax in the case of Skip Gram.

Below we can see the average results of the experiments for the different factors of interest in our algorithm. Figure 1 shows the average time consumed for each of the algorithms. Figure 2 shows the average number of words found to be similar to those introduced in the execution of the algorithm. In Figure 3, it can be seen the average number of users that the algorithm found to be relevant. Finally, Figure 4, contains the average dataset's final size. All figures are in lineal scale. Additionally Table 2 and Table 3, contain the intervals of results in the range of minimum and maximum values obtained during the experimentation.

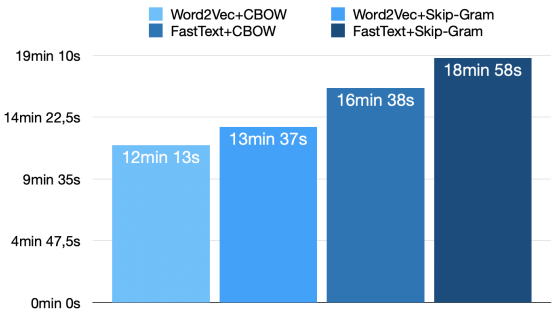

Figure 1: Mean value of elapsed time in experiments for each of the algorithms.

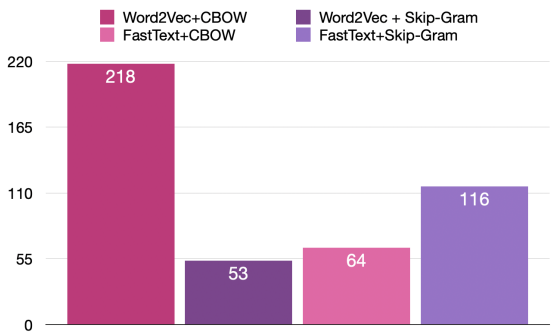

Figure 2: Mean value of words located in experiments for each of the algorithms.

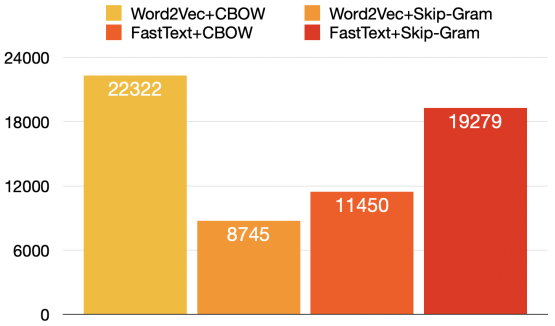

Figure 3: Mean value of users taken as relevant in experiments for each of the algorithms.

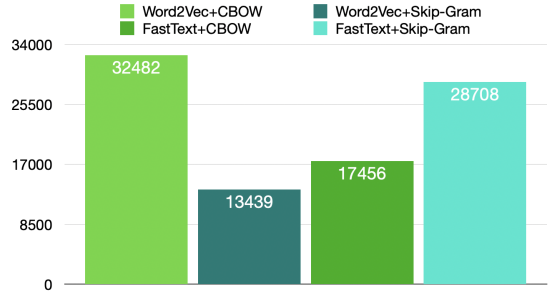

Figure 4: Mean value of the final dataset size for each of the experiments and algorithms.

If we analyse the results, we can easily see how FastText is more time-consuming than Word2Vec, due to the n-gram decomposition that the algorithm performs. In terms of manageable datasets, such as the ones we are dealing with, this value is neither worrying nor prohibitive, but the execution time would undoubtedly increase a significant amount for larger datasets.

Considering the results of users and words found, we can see a divergence between the models. There

	W2V+CBOW	W2V+Skip-Gram
Elapsed Time	Min: 11min 7s	Min: 13min 1s
	Max: 13min 36s	Max: 14min 22s
Words	Min: 209	Min: 45
	Max: 228	Max: 64
Users located	Min: 21390	Min: 7937
	Max: 23377	Max: 10044
Final dataset size	Min: 31347	Min: 12281
	Max: 34107	Max: 15239

Table 2: Minimum and maximum value for each variable in the Word2Vec experiments.

	FastText+CBOW	FastText+Skip-Gram
Elapsed Time	Min: 16min	Min: 18min 20s
	Max: 17min 15s	Max:20min 10s
Words	Min: 50	Min: 111
	Max: 79	Max: 125
Users located	Min: 10975	Min: 18128
	Max: 12312	Max: 20306
Final dataset size	Min: 16748	Min: 26547
	Max: 18773	Max: 31611

Table 3: Minimum and maximum value for each variable in the FastText experiments.

are two that we can immediately discard, because they offer a very restrictive behaviour, and find fewer words related to the domain and therefore fewer results, these would be Word2Vec+Skip-Gram and FastText+CBOW.

In terms of the majority of users and words found, Word2Vec+CBOW and FastText+SkipGram, are the best options. This leads us to conclude that each of the algorithms performs better with a different inference model for their words. In our problem FastText performs better at predicting context words on a one-word basis, while Word2Vec is better at predicting one word based on several context words. In our problem, where we have few context words due to the fact that Twitter texts are not very large, this makes an important difference. It is easier for the algorithm to predict context words based on a single word (Skip-Gram), than to predict a single word based on several words (CBOW), since the search space and the window within each document (Tweet) is very small. A priori it may seem that this does not influence, since Word2Vec+CBOW obtains great results, but if we obtain the ratio of users found for each word obtained, we can see how this value is 102 users per word on average in Word2Vec+CBOW, while this rises to 167 users per word in the case of FastText+SkipGram. This leads us to conclude that the words located by FastText have a higher representation in the dataset, as well as a higher relationship with the topic under study. This ratio is also improved in the case of FastText+CBOW, but after a manual check we have been able to ver-

ify that there are words that are very representative of medicine, such as *surgeon, pharmacology* or *toxicologist* among many others that are not localised by the latter option.

Therefore, the best option for our algorithm will be to use FastText+SkipGram, although more time-consuming, this increase is also linked to a higher match value for the selected words and their relation to medicine, as well as a better user selection ratio. A complete result of words found by the final algorithm is: *research, doctor, surgeon, researcher, medicine, medical, lecturer, clinical, physician, epidemic, pediatrics, epidemiologist, exclinical, postdoctoral, toxicologist, epidemiolog, nonmedical, medicinal, radiotherapy, exmedical, cally, surgery, psychologist, institute, professor, ecologist, searched, paediatric, regarding, postdoc, pediatric, musicologist, chronically, oncosurgeon, clinician, paramedicine, biomedicine, postdocs, medicina, lyricologist, smedical, telemedicine, lagosmedical, cancer, marched, baemedical, medicity, laparoscopy, toxicology, labmedicine, endoscopic, issue, depressed, health, phd, fellowship, institut, chronic, ecology, organically, faculty, electoral, tanto, biomedical, infectious, biome, graphologist, clinic, surge, medica, bariatric, physically, lapar, mdspediatric, orthopaedic, treatment, technically, topical, ecological, cliched, lucina, telemark, unironically, guarding, biomed, proffícinal, musicology, untouched, laparoscopic, psychologer, endemic, harding, opioid, geriatrician, dermatology, biomedic, neurobiology, logically, exdoctor, mycology, ethically, endoscopy, bandemic, mycologist, ironically, paediatrics, trichologist, psicologia, scorched* and *dicovered*; where we can see that the vast majority are words closely related to the domain of medicine and science. These terms are then used by the system to filter our dataset and keep only high value users as we can see in Figure 5.

We can observe that these users are very relevant to the topic of medicine. The algorithm even selects users verified by Twitter as in the case of the second account in Figure 5. The vast majority of the accounts located by the algorithm are similar to those seen in Figure 5 where the user's experience in the topic is more than evident. The power of the algorithm to automatically extend the search words is also evident, as we can see, none of the accounts shown above have the words we entered as input to the algorithm.

In contrast, a not so good result was obtained for the case of Word2Vec+Skip-Gram where the words found are very poorly relative to medicine and have hardly any meaning in many cases. The words for one of the experiments are: *doctor, researcher, medicine, medical, clinical, obstetrics, hayatabad, epidemiologist,*

Figure 5: Some anonymised profiles selected by the best configuration of the algorithm.

bionerd, jeenal, traumatology, maternal, anaesthetic, traumacare, survivorship, anovus, mielipiteet, trainee, multifand, veeda, philologist, activistig, amira, ksomlive, khoutv, gmcian, teamfcb, disparity, nan, veterinary, nephrology, neurologist, learing, gastro, infant, antiaging, cruff, trialist, lowes, mdspediatric, lovehate, adequacy, kashmirim, finewine, tonga, underpinning, gmers, kaggle, psychiatric, kinnaird, sefako, maxilofacial, rheumatology, poindi, hematology, rosier, paediatrics, compounder, fetal, neetfailure, utilization, inpatient, lordgod, gynecologic, and *muadhin*. These words lead the algorithm to get users like the ones we can see in Figure 6.

5 Conclusion and future work

Throughout the paper it has been demonstrated that the algorithm works adequately, and that it meets the objectives set out in Section 1. Relevant users have been obtained, and the dimensionality reduction has been of great importance. This shows that the algorithm can be very useful in Big Data problems related to Twitter, as it can select the content and users relevant to a given topic under study.

As for word embedding, it has been proven to be very useful for expanding search terms, which in conjunction with Twitter biographies is very useful for locating relevant accounts. It has been shown that in terms

Figure 6: Some anonymised profiles selected by the worst configuration of the algorithm.

of efficiency-accuracy ratio, FastText+SkipGram offers the best solution although n-gram decomposition can be a bottleneck in larger datasets.

In terms of future possibilities and challenges, it is worth mentioning that the algorithm is very sensitive to lies, because if someone lies in their biography it is very difficult to dismiss them as not relevant. Being able to more accurately discern lies on social networks would improve the algorithm considerably and is certainly a very promising direction of research.

Acknowledgement

The research reported in this paper was partially supported by the COPKIT project under the European Union's Horizon 2020 research and innovation program (grant agreement No 786687), the Andalusian government and the FEDER operative program under the project BigDataMed (P18-RT-2947 and B-TIC-145-UGR18). Finally the project is also partially supported by the Spanish Ministry of Education, Culture and Sport (FPU18/00150).

References

[1] B. Abu-Salih, P. Wongthongtham, K. Y. Chan, D. Zhu, Credsat: Credibility ranking of users in big social data incorporating semantic analysis and temporal factor, Journal of Information Science 45 (2) (2019) 259–280.

[2] M. Alrubaian, M. AL-Qurishi, M. Alrakhami, M. Hassan, A. Alamri, Reputation-based credibility analysis of twitter social network users: Reputation-based credibility analysis of twitter social network users, Concurrency and Computation: Practice and Experience 29.

[3] M. Alrubaian, M. Al-Qurishi, M. M. Hassan, A. Alamri, A credibility analysis system for assessing information on twitter, IEEE Transactions on Dependable and Secure Computing 15 (4) (2018) 661–674.

[4] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606.

[5] F. Diaz, B. Mitra, N. Craswell, Query expansion with locally-trained word embeddings, arXiv preprint arXiv:1605.07891.

[6] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, K. Gummadi, Cognos: crowdsourcing search for topic experts in microblogs, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012, pp. 575–590.

[7] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in wuhan, china, The Lancet 395 (10223) (2020) 497–506.

[8] S. Kuzi, A. Shtok, O. Kurland, Query expansion using word embeddings, in: Proceedings of the 25th ACM international on conference on information and knowledge management, 2016, pp. 1929–1932.

[9] R. Lamsal, Coronavirus (covid-19) tweets dataset (2020).
URL `https://dx.doi.org/10.21227/781w-ef42`

[10] O. Levy, Y. Goldberg, Dependency-based word embeddings, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, pp. 302–308.

[11] Y. Liu, Z. Liu, T.-S. Chua, M. Sun, Topical word embeddings, in: Twenty-ninth AAAI Conference on Artificial Intelligence, Citeseer, 2015.

[12] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems 26 (2013) 3111–3119.

[14] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
URL http://www.aclweb.org/anthology/D14-1162

[15] R. Řehřek, P. Sojka, Gensim statistical semantics in python, Retrieved from genism. org.

[16] D. Roy, D. Paul, M. Mitra, U. Garain, Using word embeddings for automatic query expansion, arXiv preprint arXiv:1606.07608.

[17] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (7798) (2020) 270–273.

3.3 Improving text clustering using a new technique for selecting trustworthy content in social networks

- Diaz-Garcia, J. A., Fernandez-Basso, C., Gutiérrez-Batista, K., Ruiz, M. D., & Martin-Bautista, M. J. (2022, July). Improving Text Clustering Using a New Technique for Selecting Trustworthy Content in Social Networks. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 275-287). Cham: Springer International Publishing.
  - Conference: Information Processing and Management of Uncertainty in Knowledge-Based Systems: 19th International Conference, IPMU 2022, Milan, Italy, July 11–15, 2022, Proceedings, Part II
  - Status: Published.

Improving text clustering using a new technique for selecting trustworthy content in social networks

J. Angel Diaz-Garcia[1 [0000−0002−9263−1402]], Carlos Fernandez-Basso[1 [0000−0002−8809−8676]], Karel Gutiérrez-Batista[2 [0000−0003−2711−4625]], M. Dolores Ruiz[1 [0000−0003−1077−3173]], and Maria J. Martin-Bautista[1 [0000−0002−6973−477X]]

Department of Computer Science and A.I., University of Granada, Spain {jagarcia, cjferba, karel, mdruiz, mbautis}@decsai.ugr.es

Abstract. Today's information society has led to the emergence of a large number of applications that generate and consume digital data. Many of these applications are based on social networks, and therefore their information often comes in the form of unstructured text. This text from social media also tends to contain a high level of noise and untrustworthy content. Therefore, having systems capable of dealing with it efficiently is a very relevant issue. In order to verify the trustworthiness of the social media content, it is necessary to analyse and explore social media data by using text mining techniques. One of the most widespread techniques in the field of text mining is text clustering, that allows us to automatically group similar documents into categories. Text clustering is very sensitive to the presence of noise and so in this paper we propose a pre-processing pipeline based on word embedding that allows selecting trustworthy content and discarding noise in a way that improves clustering results. To validate the proposed pipeline, a real use case is provided on a Twitter dataset related to COVID-19.

Keywords: Clustering · pre-processing · social media mining

1 Introduction

Nowadays the world in we live, are very influenced by social networks. Every day, we are consuming o generating social networks data. These data, usually come in form of user-generated content, or in other words, unstructured data. Being able to process and analyse the data properly is a very arduous task in which Artificial Intelligence (A.I.) is taking a leading role. Among the A.I. techniques aimed at obtaining relevant information about these unstructured data are supervised techniques such as classification [19], or descriptive techniques such as association rules [17] [16] or clustering. Clustering is one of the most widespread A.I. techniques, with great results in various fields of application such as energy [38], health [6], or economics [23]. Due to its potential to obtain hidden groups in data without prior labelling, clustering is also very relevant in

social media analysis problems. Some of its most relevant applications have been in community analysis [7] or text mining through document clustering [8]. Due to its importance for these sectors, it is necessary to have increasingly reliable clustering techniques.

In this paper, we propose a new text filtering technique to improve clustering results on microblog social networks. Social networks contain a lot of noise and unhelpful or untrustworthy content. This content does not contribute anything, but it is capable of causing traditional data mining techniques to underperform [14]. Therefore, detecting and eliminating it is a relevant task. Our approach is based on selecting trustworthy accounts for a given analysis as well as their content. On the other hand, we eliminate those that are not related to our topic of study, as these accounts have a high probability of generating noise and untrustworthy content. To do this, the system uses on Twitter biographies, as these are typically used to report on professions and interests, so we can create a system that can automatically select which accounts are useful for a topic and which are not in the way a human could. A user on Twitter, to decide whether to believe a certain content or not, would visit the account of the user issuing that tweet and contrast the content of the tweet with the information provided by that person in his or her biography. For example, if a user is reading a tweet about COVID-19, the first thing he or she will do is to check who has issued the tweet. If the author of the tweet provides in his biography information about his profession, such as whether he is a doctor or works in a certain hospital, these will be valuable arguments to give more credibility to the tweet. In this paper, we propose an automatic technique to perform this task of account contrasting, so that we can remove noise and untrustworthy content from the analysis and improve the clustering results. To do this, the system uses the potential of Word Embbedings (W.E.).

Two of the most famous W.E. are Word2Vec [29], [30] and FastText [12]. Both are based on the representation of each of the words present in a vocabulary by means of vectors in a vector space so that the distance between words can be operated mathematically. The distance between one word and another will correspond to their semantic relationship. That is, words with a greater vector distance will have less relation and words with a very similar vector distance can be considered synonyms. In this paper, we will use this potential to find similarities between words to guide an automatic and incremental filtering of users in relation to a given topic. The proposed system will be validated in a real problem related to COVID-19, in which we seek to obtain opinion clusters in the social network Twitter.

The paper is organized as follows: Next section focuses on the study of related work. In Section 3 we go into detail in the proposed text filtering for content-based. In Section 4, we provide the results of the experimentation as well as the parameters used. Finally, in Section 5 we examine the conclusions and the future work.

2 Related works

Regarding the improvement of clustering results, most of the existing studies in the literature are based on the optimization of the algorithm [9], [35], [37], [32]. These approaches are based on improving the selection of the initial parameters of the algorithms such as the initial centroids, the calculation of distances between the centroids and the examples of each of the clusters. While optimisation of algorithms is undoubtedly one of the vital parts of improving today's A.I.-based systems, we should not forget that proper data pre-processing can have even better implications on the final result.

Many of the approaches to pre-processing data for clustering are based on feature selection. By means of this technique, they try to solve one of the biggest problems in text clustering: the sparsity in document-term matrices. Before applying clustering on a document, it must be represented by a matrix in which the rows are the documents and the columns are the terms present in it. Each cell of the matrix will have the frequency of a term in a document. Using this structure a vector representation can be obtained, usually based on TF-IDF. A detailed explanation of the calculation of the TF-IDF and the creation of the vector space for clustering can be studied in [36]. The problem with these matrices is that they are very sparse. The fact that they are sparse is useful for calculating distances and correctly locating clusters, but in many cases this process is heavily biased by words that introduce noise and are not really meaningful. The detection and elimination of these non-useful words in order to achieve a dataset with the best possible characteristics to improve clustering, has been tackled by several A.I. techniques, such as principal component analysis techniques [13], using evolutionary algorithms [2], or even using algorithms based on harmony search [3]. Recently, there is also an attempt to reduce the noise present in the matrices by means W.E. based techniques [10], [34].

All of these approaches aim to keep all documents present in the analysis and remove certain features from each of them. Our approach, in contrast to what has been studied in the literature, is to keep all those words but only from those documents that are really useful for a given analysis. In this way, we achieve more cohesive matrices, without having to eliminate characteristics (words), since one of the problems of clustering is usually that it does not have enough words to correctly categorise a document. To select the documents (tweets) that should be deleted, we rely on the trustworthiness of the user issuing them; if a user is trustworthy, we will take the document into account, if not, we will not. The study of the experience or location of trustworthy sources of information, has been approached from multiple perspectives. Some of the most widespread are based on analysing user content and concepts such as engagement (level of interaction with the community) to determine if content is credible or not [5], [4]. In our case, we focus on detecting accounts that denote expertise in a sector, and therefore generate useful content related to that sector or domain. In this way, there are also systems such as Cognos [18] or CredSat [1]. These systems offer recommendations of expert users on a topic, making use of lists or again, of the content generated. The ultimate goal of these systems is to retrieve reliable

or valuable users, as opposed to our system that seeks to filter a large set of data using the concept of expertise in order to obtain better data structures and improve downstream data mining processes. To the best of our knowledge, this is the first paper that proposes a trustworthy-based filtering approach to reduce the dimensionality of a problem and improve clustering results.

3 Trustworthy content selection pipeline

In this section our proposal is detailed.

3.1 System architecture

In Figure 1 we can see a complete scheme of our proposal. The different elements that comprise the proposal shown in the figure will be analysed in detail throughout this section. We start from a database of tweets. We have to bear in mind that from all the metadata that the APIs and tweet databases offer us, we will only keep those relating to the user, such as their username, location, number of followers, favourites, lists, friends, biography and content of the tweet. From all these elements, the really interesting ones are the biography, which we will use for the trustworthy content filter, and the content, which we will use for later stages of data mining, in this case clustering.

The power of biographies lies in the fact that normally are used by people to inform about their professions or interests. therefore, we propose a system capable of automatically detect words related to professions and jobs, and create a list of subject matter experts. At the same time selecting their content as trustworthy for further stages of data mining.

The system is composed of a pre-processing module, a module based on the selection of trustworthy content based on user biographies and finally a layer dedicated to data mining through clustering. Given the importance of these modules in the system, we will see them in detail in the following sections.

3.2 Text pre-processing

The pre-processing uses common cleaning techniques for user-generated text. This step is applied to the user's biography and to the tweet content. The techniques used are:

1. Twitter domain related cleaning. For this purpose we have eliminated URLs, hashtags, mentions, reserved words from Twitter (RT, FAV...), emojis and smileys.
2. Text mining domain cleaning: This comprises removing numbers, additional spaces and punctuation marks.
3. Turning the text into lowercase letters.
4. Detecting the tweet language. All those tweets using a non-recognised language or from a language other than the one desired by the user are eliminated.

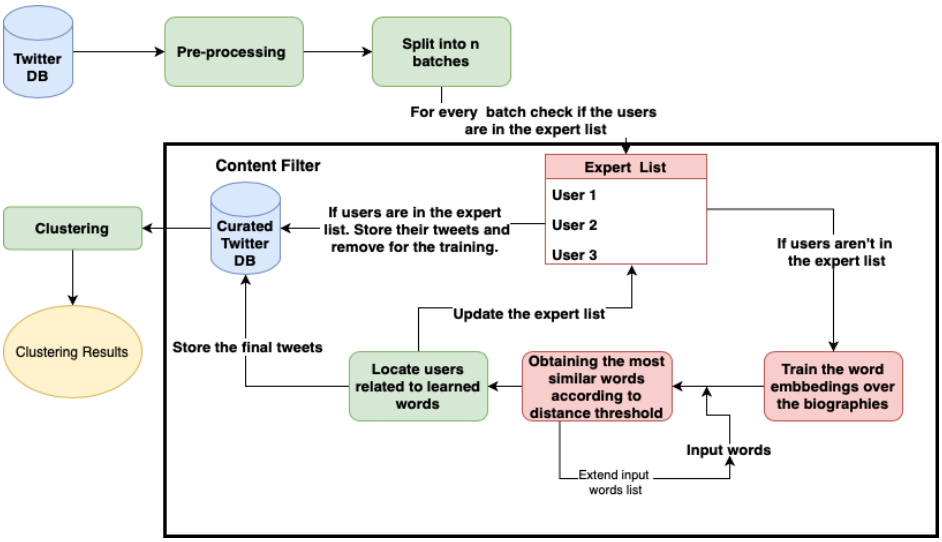

Fig. 1. Data flow and architecture of our proposal.

5. Stop words removal.
6. Any empty tweets are removed.
7. Tokenization of the biography and the tweet.

3.3 Content filter

For the content filter, we have based it on the premise that if a person has knowledge on a subject, the content that this particular user generates on a topic related to his or her area of expertise will be of high value for the analysis. In contrast, content generated by a user with no knowledge of the topic will generally correspond to noise. As an example, if we look at the COVID-19 use case, it will be more useful to analyse 10 000 tweets created by doctors and researchers than 1 000 000 tweets generated by users who are simply talking about the topic without any real background in it. Therefore, the content filter is based on locating users who are experts in the subject matter and its content. To do this, we will focus on their biographies. The system works as follows (in Figure 1, we can find the data flow of this filter):

1. The filter method receive as input a list of words related to the topic of analysis, e.g. if we want relevant information about COVID-19 and its symptoms, the initial word list could be *medical, doctor* and *epidemiology*.
2. Using an adaptation of the divide and conquer approach, the dataset resulting from the pre-processing stage is split into different batches.
3. In the first batch, the structures are empty so there are no experts located. Therefore, word embedding is trained on the biographies. The W.E. used for this stage was FastText.

4. On the resulting vector space, we obtain those words that exceed the threshold of 0.6 in similarity over the initial list. These words are added to the search list so that at each iteration the list grows and improves covering more aspects of the problem domain. In this case, in each iteration, the system learns words related to medicine.
5. Using these words we filter out users who use some of these words in their biographies and add them to a list of credible or expert users.
6. The content of these localised users is reported as relevant and added to the final database.
7. In the second and subsequent batches. The first step is to check if the user is located in the expert list, if so, all the content related to the user is passed to the final dataset avoiding redundant processes.
8. For all content remaining after those already identified as relevant, steps 3, 4, 5 and 6 are repeated.

After this stage, we will have a clean dataset consisting only of content that is actually related to the topic of analysis. The algorithm can be applied to any other problem. For example, if the dataset is related to economics or stock market surveillance, we might not be interested in analysing everything that is talked about, we would only need what is talked about by people related to the sector. We could use this method with the words *economist* and *economy* and we would get a cohesive dataset on the subject.

3.4 Clustering

In the final stage of the pipeline we use clustering on the filtered dataset. We must take into account that the level of dimensionality reduction is around 90%, so we will have fewer documents with, also, fewer words. Therefore, we have a more cohesive matrix to apply the K-Means clustering algorithm.

The k-means clustering algorithm is widely known for clustering data using different types of distances [27]. It is based on a calculation of the distance between different data using a number of base centroids that are determined by a parameter.

The main advantages of the k-means method are that it is a simple and fast method [24]. In addition, this algorithm works well with large or small data sets, is efficient and performs well.

Specifically, this clustering algorithm has been widely used for tex mining. Thanks to its advantages such as being able to use different distance measures, to use it with large datasets and to parameterise the number of clusters to extract, it has been used in many applications such as sentiment analysis [26], recommender systems [11] and text analysis [21].

4 Experimentation

For the experimentation, we have proposed a hypothesis test having as a starting hypothesis that the filtering of our dataset improves the clustering result. To

demonstrate this, a comparative study has been carried out on the results of the K-Means algorithm. The comparison is made between the case of using the pre-processing seen in Section 3.2 together with the content filtering seen in Section 3.3 and the case of only using the pre-processing seen in Section 3.2, i.e. without filtering the content but pre-processing the text.

4.1 Dataset

The disease caused by the new Coronavirus (Sars-Cov-2) [39], first reported in Wuhan [20] in December 2019, now affects the entire world and is considered one of the largest pandemics in the history of mankind. News and information about the pandemic is generated on a daily basis. Analysing them automatically is an important task. One of the most used channels for disseminating information about COVID-19 is Twitter, so our datasets come from real data from this social network. Currently, the tweet dataset [25] related to COVID-19 includes millions of entries, which makes it a perfect candidate for our use case. Our problem, uses random samples of that dataset comprising a total of 2 626 275 tweets.

4.2 Evaluation metrics

In order to prove the feasibility of the proposed approach, we have used two automatic metrics: Silhouette coefficient [31], and Davies-Bouldin score [15]. Both metrics will allow the evaluation of the obtained results using our approach and without using it. The reason for using these evaluation techniques is that they are two of the best known, most widely used and complementary evaluation techniques. With the silhouette coefficient we can check how good the separation between clusters is, and with the Davies-Bouldin score, we can check the goodness of the examples within each cluster.

Silhouette coefficient. The Silhouette coefficient has been widely used in clustering problems because it determines the quality of separation and cohesion of the obtained clusters. One of the main usages of this metric is to obtain the optimal number of clusters for which a clustering algorithm shows a better performance [31].

The coefficient is computed using the mean intra-cluster similarity $a(i)$ and the mean nearest-cluster similarity $b(i)$ for each sample (Eq.1). The overall Silhouette coefficient is the mean Silhouette coefficient of all samples. The values are in the range [-1,1], being the best results those values that are close to 1.

$$S(i) = \frac{a(i) - b(i)}{max\left\{a(i), b(i)\right\}} \tag{1}$$

Davies-Bouldin score. Like the Silhouette coefficient, the Davies-Bouldin score allows the evaluation of the results of the clustering algorithms [15]. The Davies-Bouldin score is the average similarity measure of each cluster with its

most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances (Eq.3).

$$r_k = \frac{1}{|A_k|} \sum_{x_i \in A_k} d(x_i, C_k) \tag{2}$$

$$DB = \frac{1}{N} \sum_{i}^{N} \sum_{j}^{N} \max_{i \neq j} \frac{r_i + r_j}{d(C_i, C_j)} \tag{3}$$

where r_i and r_j are represented in Eq.2 (intra-cluster distance), and $d(C_i, C_j)$ is the distance between the centroids C_i and C_j. .

It means that clusters that are farther apart and less dispersed will have a better score. The minimum possible value is zero, with lower values indicating better clustering results, unlike the Silhouette coefficient.

4.3 Clustering results

In this section we present the final results of clustering on the filtered dataset and compare it with the unfiltered dataset. In order to take data so that the results are biased by the number of tweets, a random sample of 10000 tweets was obtained from the original data and the filtered one. So we have 10000 tweets from the filtered dataset and 10000 from the original dataset. In this way we try to mitigate possible biases that might arise from the size of the dataset, since the unfiltered dataset is much larger. These 10000 tweets are obtained randomly for each of the executions, trying to mitigate also that the random choice corresponds to an optimal solution, and favouring the generalisation of the experimentation. So the comparison can be made on the same grounds, and a better result in this case will indicate that the filtering favours the result.

Table 1 shows the average results of 10 runs in terms of Silhouette Score and Davies Bouldin Score. We can see how the results in both metrics are better in the case of filtering the text using our proposal. We must take into account that, although the results may not look good, we are facing a clustering problem in short texts [22]. In these problems, there are few words in each of the documents, in which case, the clustering algorithms do not usually have very high values in terms of fit measures.

Table 1. Average result of the runs on the dataset filtered with our proposal and without filtering.

Metric	Filtered	Unfiltered
Silhouette Score	**0,0229**	0,00606
Davies Bouldin	**4,74957**	5,65733

Apart from the goodness of fit measures, to graphically compare the obtained results, we have represented them by means of a t-distributed stochastic

neighbour embedding (TSNE) graph [28]. Figure 2 shows the clustering results in the case of filtering the data while Figure 3 shows the results in the case of no filtering. If we analyse the results we can see that in the first case, we can easily locate 11 clusters, while in the case of not filtering by content, we can only identify 7 clusters.

If we take into account that the final objective of our filtering process is to favour the subsequent stages of data mining, clustering in this case, we see a clear improvement in the case of filtering versus not filtering. In a text mining problem, it is very likely that a large number of clusters will appear, depending on the documents. Specifically in these cases, a battery of tests has been carried out to determine the number of clusters that minimises the error, which is around 26. Therefore, being able to manually locate a greater number of clusters is an indicator that the clustering process is better, and therefore, the preprocessing applied actually works.

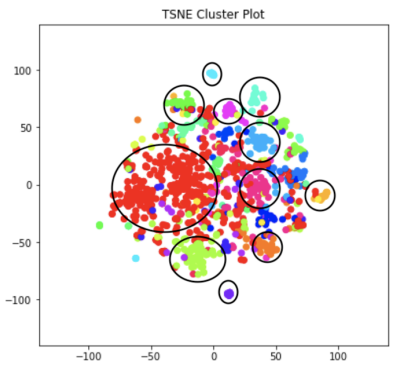

Fig. 2. TSNE plot using the content filter

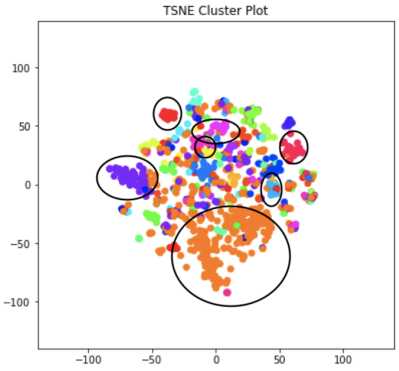

Fig. 3. TSNE plot without filter

For a more detailed analysis and to support the obtained results, we have conducted a statistical analysis to determine whether there are any significant differences among the obtained values for the two approaches (filtered and unfiltered). Figures 4 and 5, depict the boxplots for the Silhouette coefficient and Davies-Bouldin score respectively, regarding the two approaches. At first sight, we can see that the best results are yielded for both metrics when we use our approach.

We can see how the distribution of results, in both cases, is better when using our filtering method. The box is always wider in the case of filtering because the same accounts and content are not always selected. There is a certain random component to word embedding filtering. This causes the result to fluctuate more in the filtering case than in the non-filtering case where it always runs on similar terms. To justify that the improvement is not due to randomness, we have performed numerous runs and statistical tests.

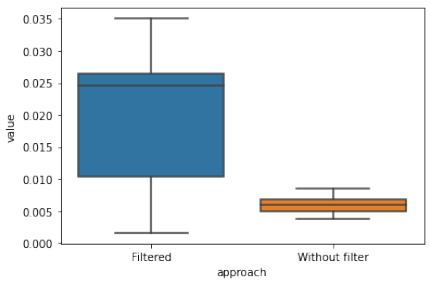

Fig. 4. Boxplot of the Silhouette coefficient taking into account both approaches

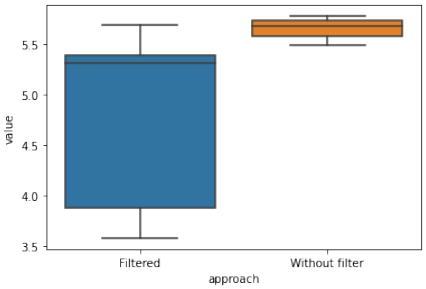

Fig. 5. Boxplot of the Davies-Bouldin score taking into account both approaches

For the statistical analysis, we have used the Wilcoxon's test [33] as there are only two groups (**Filtered** and **Unfiltered**). Considering the $p - values$ shown in Table 2, we can conclude that there are significant differences between both approaches, where the approach that applies the filter offers the best results.

Table 2. $P - values$ for the statistical analysis using both approaches (filtered an unfiltered) and both metrics (Silhouette and Davies-Bouldin).

Measure	p-value
Silhouette coefficient	0.02182
Davies-Bouldin score	0.00691

5 Conclusions and future work

In this paper we have proposed a useful data filtering technique to improve the textual clustering result on user-generated content in social networks. The values obtained and the statistical tests carried out show that the results are improved compared to not using the proposed filter. These conclusions are also obtained by visualising the clusters generated on a real problem related to COVID-19, where it can be seen how the clusters are of a better quality at a glance.

Is worth to highlight that the system can be very sensitive to lies in the biographies and to possible groups of well-informed users but directed with the intention of misinforming other users or social groups. As future work, a multidimensional solution to mitigate these possible problems of lies should be studied.

Acknowledgment

The research reported in this paper was partially supported by the Andalusian government and the FEDER operative program under the project BigDataMed

(P18-RT-2947 and B-TIC-145-UGR18) and grant PLEC2021-007681 funded by MCIN / AEI / 10.13039 / 501100011033 and by the European Union NextGenerationEU / PRTR. Finally the project is also partially supported by the Spanish Ministry of Education, Culture and Sport (FPU18/00150).

References

1. Abu-Salih, B., Wongthongtham, P., Chan, K.Y., Zhu, D.: Credsat: Credibility ranking of users in big social data incorporating semantic analysis and temporal factor. Journal of Information Science **45**(2), 259–280 (2019)
2. Abualigah, L.M., Khader, A.T., Al-Betar, M.A.: Unsupervised feature selection technique based on genetic algorithm for improving the text clustering. In: 2016 7th international conference on computer science and information technology (CSIT). pp. 1–6. IEEE (2016)
3. Abualigah, L.M., Khader, A.T., AlBetar, M.A., Hanandeh, E.S.: Unsupervised text feature selection technique based on particle swarm optimization algorithm for improving the text clustering. In: 1st EAI International Conference on Computer Science and Engineering. p. 169. European Alliance for Innovation (EAI) (2016)
4. Alrubaian, M., Al-Qurishi, M., Hassan, M.M., Alamri, A.: A credibility analysis system for assessing information on twitter. IEEE Transactions on Dependable and Secure Computing **15**(4), 661–674 (2018). https://doi.org/10.1109/TDSC.2016.2602338
5. Alrubaian, M., AL-Qurishi, M., Alrakhami, M., Hassan, M., Alamri, A.: Reputation-based credibility analysis of twitter social network users: Reputation-based credibility analysis of twitter social network users. Concurrency and Computation: Practice and Experience **29** (01 2016). https://doi.org/10.1002/cpe.3873
6. Alshabeeb, I.A., Ali, N.G., Naser, S.A., Shakir, W.M.: A clustering algorithm application in parkinson disease based on k-means method. Computer Science **15**(4), 1005–1014 (2020)
7. Arenas, A., Danon, L., Diaz-Guilera, A., Gleiser, P.M., Guimera, R.: Community analysis in social networks. The European Physical Journal B **38**(2), 373–380 (2004)
8. Arpaci, I., Alshehabi, S., Al-Emran, M., Khasawneh, M., Mahariq, I., Abdeljawad, T., Hassanien, A.E.: Analysis of twitter data using evolutionary clustering during the covid-19 pandemic. Computers, Materials & Continua **65**(1), 193–204 (2020)
9. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. Tech. rep., Stanford (2006)
10. Asyaky, M.S., Mandala, R.: Improving the performance of hdbscan on short text clustering by using word embedding and umap. In: 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA). pp. 1–6 (2021). https://doi.org/10.1109/ICAICTA53211.2021.9640285
11. Berry, M.W., Castellanos, M.: Survey of text mining. Computing Reviews **45**(9), 548 (2004)
12. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
13. Chaudhary, G., Kshirsagar, M.: Enhanced text clustering approach using hierarchical agglomerative clustering with principal components analysis to design document recommendation system. Advanced Research in Computer Engineering. Research Transcripts in Computer, Electrical and Electronics Engineering **2**, 1–18 (2021)

14. Dave, R.N.: Characterization and detection of noise in clustering. Pattern Recognition Letters **12**(11), 657–664 (1991)
15. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (1979)
16. Diaz-Garcia, J.A., Fernandez-Basso, C., Ruiz, M.D., Martin-Bautista, M.J.: Mining text patterns over fake and real tweets. In: Lesot, M.J., Vieira, S., Reformat, M.Z., Carvalho, J.P., Wilbik, A., Bouchon-Meunier, B., Yager, R.R. (eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems. pp. 648–660. Springer International Publishing, Cham (2020)
17. Diaz-Garcia, J.A., Ruiz, M.D., Martin-Bautista, M.J.: Non-query-based pattern mining and sentiment analysis for massive microblogging online texts. IEEE Access **8**, 78166–78182 (2020). https://doi.org/10.1109/ACCESS.2020.2990461
18. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., Gummadi, K.: Cognos: crowdsourcing search for topic experts in microblogs. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 575–590 (2012)
19. Godara, N., Kumar, S.: Twitter sentiment classification using machine learning techniques. Waffen-Und Kostumkunde Journal **11**(8), 10–20 (2020)
20. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al.: Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. The Lancet **395**(10223), 497–506 (2020)
21. Jalil, A.M., Hafidi, I., Alami, L., Ensa, K.: Comparative study of clustering algorithms in text mining context (2016)
22. Jin, C., Zhang, S.: Micro-blog short text clustering algorithm based on bootstrapping. In: 2019 12th International Symposium on Computational Intelligence and Design (ISCID). vol. 2, pp. 264–266. IEEE (2019)
23. Jin, Y., Liu, Y., Zhang, W., Zhang, S., Lou, Y.: A novel multi-stage ensemble model with multiple k-means-based selective undersampling: An application in credit scoring. Journal of Intelligent & Fuzzy Systems (Preprint), 1–14 (2021)
24. Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in k-means clustering. International Journal **1**(6), 90–95 (2013)
25. Lamsal, R.: Coronavirus (covid-19) tweets dataset (2020). https://doi.org/10.21227/781w-ef42, https://dx.doi.org/10.21227/781w-ef42
26. Li, N., Wu, D.D.: Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision support systems **48**(2), 354–368 (2010)
27. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means clustering algorithm. Pattern recognition **36**(2), 451–461 (2003)
28. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(Nov), 2579–2605 (2008)
29. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems **26**, 3111–3119 (2013)
31. Rousseeuw, P.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**(1), 53–65 (Nov 1987)
32. Shi, K., Li, L., He, J., Zhang, N., Liu, H., Song, W.: Improved ga-based text clustering algorithm. In: 2011 4th IEEE International Conference on Broadband Network and Multimedia Technology. pp. 675–679. IEEE (2011)
33. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945)

34. Xingliang, M., Fangfang, L.: Clustering of short text in micro-blog based on k-means algorithm. In: 2018 IEEE International Conference of Safety Produce Informatization (IICSPI). pp. 812–815 (2018). https://doi.org/10.1109/IICSPI.2018.8690507
35. Yedla, M., Pathakota, S.R., Srinivasa, T.: Enhancing k-means clustering algorithm with improved initial center. International Journal of computer science and information technologies **1**(2), 121–125 (2010)
36. Yuan, S., Wenbin, G.: A text clustering algorithm based on simplified cluster hypothesis. In: 2013 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA). pp. 412–415 (2013). https://doi.org/10.1109/IMSNA.2013.6743303
37. Zhang, G., Zhang, C., Zhang, H.: Improved k-means algorithm based on density canopy. Knowledge-based systems **145**, 289–297 (2018)
38. Zhang, G., Li, Y., Deng, X.: K-means clustering-based electrical equipment identification for smart building application. Information **11**(1), 27 (2020)
39. Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al.: A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature **579**(7798), 270–273 (2020)

4 A flexible Big Data system for credibility-based filtering of social media information according to expertise

- Jose A. Diaz-Garcia, Karel Gutiérrez-Batista, Carlos Fernandez Basso, M. Dolores Ruiz & Maria J. Martin-Bautista. (2023) A flexible Big Data system for credibility-based filtering of social media information according to expertise. International Journal of Computational Intelligence Systems.
  - Status: Submitted to International Journal of Computational Intelligence Systems.
  - Impact Factor (JCR 2021): 2.259
  - Subject Category:Computer Science, Artificial Intelligence. Order 102/145 Q3.

A flexible Big Data system for credibility-based filtering of social media information according to expertise

Jose A. Diaz-Garcia[a,c], Karel Gutiérrez-Batista[a,f], Carlos Fernandez-Basso[a,b,g], M. Dolores Ruiz[a,e], Maria J. Martin-Bautista[a,d]

[a]*Department of Computer Science and A.I, University of Granada, C. Periodista Daniel Saucedo Aranda, s/n, Granada, 18014, Granada, Spain*
[b]*Causal Cognition Lab Division of Psychology and Language Sciences University College London London United Kingdom*
[c]*Corresponding author: jagarcia@decsai.ugr.es*
[d]*mbautis@decsai.ugr.es*
[e]*mdruiz@decsai.ugr.es*
[f]*karel@decsai.ugr.es*
[g]*cjferba@decsai.ugr.es*

Abstract

Nowadays, social networks have taken on an irreplaceable role as sources of information. Millions of people use them daily to find out about the issues of the moment. This success has meant that the amount of content present in social networks is unmanageable and, in many cases, fake or non-credible. Therefore, a correct pre-processing of the data is necessary if we want to obtain knowledge and value from these data sets. In this paper, we propose a new data pre-processing technique based on Big Data that seeks to solve two of the key concepts of the Big Data paradigm, data validity and credibility of the data and volume. The system is a Spark-based filter that allows us to flexibly select credible users related to a given topic under analysis, reducing the volume of data and keeping only valid data for the problem under study. The proposed system uses the power of word embeddings in conjunction with other text mining and natural language processing techniques. The system has been validated using three real use cases.

Keywords: social media mining, pre-processing, big data, expertise, credibility

1. Introduction

Today's world is completely influenced by social media [1]. They are used daily by millions of people to share information or obtain information about current topics. This use generates an immeasurable amount of data, which, when correctly analysed, can be of great value to companies, organisations, or even the users of social networks themselves [2]. In the social media context, most of the data generated that may be of interest comes in the form of unstructured text, such as opinions about a product, reviews about a restaurant or simply conversation threads on a given topic. Due to its nature, this type of data has an associated problem since, being user-generated content, we can find incomplete data, false or irrelevant content, colloquial uses of words, syntactic errors, and use of emoticons or simplifications of words [3]. In addition to those problems, social media data usually has issues related to the volume of the dataset. Daily, millions of tweets, posts, news or images are posted to social networks. Traditional data mining techniques can not process properly that volume of data or have serious problems doing it [4].

Therefore, to derive value from this user-generated content, we must apply efficient data pre-processing techniques that allow us to have quality datasets from which to derive value and knowledge. In this paper, we offer a new technique that addresses two issues related to processing user-generated textual content in social networks: data validity and data volume reduction. The proposed technique is essentially a pre-processing technique designed to obtain a reduced and valid set of data for a given topic from a big dataset of data from social networks. On this dataset, other data mining techniques can be applied, with special value for those in which the volume of data can derive efficiency issues.

The premise of our paper focuses on the fact that many professionals use social networks such as Twitter to disseminate their knowledge and opinions about their field of expertise. The data from these professionals will be those that contribute the most value to the desired analysis or use cases, and therefore, they are the ones that should be located. Given that biographies on social networks allow us to provide information about our work, likes or interests, our system will focus on using this textual element to discern who might be interesting to take into account as an expert and who might not. To achieve this, the core of the system uses word embeddings and their potential to obtain semantic relations between words. For example, two words of similar meaning, such as *doctor* and *surgeon*, will have a very small

distance in the low-dimensional vector space so that by distance substraction, we can obtain those words that are similar to a given word. In this way, by comparing the words with the topic under analysis, we will have a broad relationship to the topic and, therefore, can be used to expand the search query and perform flexible filtering by topic of analysis. This paper is the final result of the work [5] where a first approach to the expertise filter was conducted and in which a comparative battery of experiments was generated to discern the underlying word embedding of the final algorithm. In the research presented in this paper, the algorithm has evolved to a final flexible version under the big data paradigm. The major contributions of our paper to the state-of-the-art are:

- The proposal of a flexible system capable of generalising and adapting to a user's information needs at any given time. The system is based on filtering content based on experience, so it obtains content issued by people with knowledge in a given field.
- The development of the complete system under the Big Data paradigm using Spark [6], highlighting its great usefulness in problems involving large data sets from social networks.
- The proposal of a pre-processing system that can be used to solve two of the most plaguing problems of Big Data: the validity and the volume of data.

To validate the system, a series of use cases and experiments have been proposed on a real dataset from Twitter. The filtering system proposed in this paper has been applied to different information needs, one related to COVID and health, one to politics and one to sports. In all cases, it is shown how the system adapts flexibly to the user needs, obtaining valid data for that topic according to the users' experience in it, considerably reducing the volume of data and achieving a data sample which represents the topic with higher credibility, augmenting thus, their quality and validity. For each use case, we demonstrate that our algorithm works properly using Named Entity Recognition (NER) [7, 8] and visualization based on tag clouds. In the visualization of the last layer of the process, we conduct an analysis that corroborates how the system can obtain valuable information starting from a non-quality dataset in an unsupervised way.

The paper is organized as follows: Next section focuses on the study of related work. In Section 3, we explain the algorithm proposed for experience-based filtering. In Section 4, we explain the Big Data architecture for the proposed system. In Section 5, we provide a detailed explanation of the experimental process carried out. Finally, in Section 6, we illustrate and validate the system with three real use cases. Final remarks and possible extensions are discussed in Section 7.

2. Related Work

As mentioned above, in this research, we propose a new data pre-processing technique based on Big Data and the power of word embeddings to solve two of the Vs regarding the Big Data paradigm: validity and volume. Many studies have addressed the problem of filtering information from social networks, most based on credibility-based dimensionality reduction techniques, either at content level [9, 10, 11], or user level [12, 13, 14, 15]. Studies can be found that tackle the problem of filtering information through other approaches, such as fake news detection [16, 17, 18].

Considering that one of the main goals of the proposal is to develop a system under the distributed paradigm in order to be able to deal with large datasets from social networks, the literature on this topic has been reviewed.

2.1. Credibility-based information filtering

In [9], the authors propose a supervised method to analyse the credibility of a given set of tweets from Twitter in an automatic way. In order to create the model, the authors extract relevant features from tweets considered trending topics. The features are extracted from the tweet, users' behaviour, and citation to external sources. Finally, using all the features mentioned above, the model can classify the tweets as credible or not. It should be noted that, unlike our proposal, the authors use users' information, like biography, in a straightforward way.

Hassan [11] develops a text mining-based approach to assess the credibility of events in social networks automatically. The author uses a set of popular Twitter events manually annotated with different credibility ratings to build a model based on the Decision Tree classifier.

Canini et al. [10] address the problem of filtering credible information in social networks based on its content and structure. In this research, the authors detect (through experimental results) different factors affecting explicit

and implicit credibility judgments in online social networks. Based on these results, they propose an approach to automatically identify and rank social network users according to their relevance and expertise for a given topic.

As stated, some approaches tackle the challenge of filtering credible information through fake news detection or similar tasks. That is the case in [18], where the authors propose a method for spam detection in emails based on the KNN classifier jointly with bio-inspired optimization techniques. In [17], a two-step-based method for fake news detection in social media is proposed. The first step comprises several pre-processing techniques in order to transform unstructured into structured data. The news in the transformed data is represented using the old-fancy Term Frequency (TF) feature representation. Finally, twenty-three classification algorithms are used to classify the news into fake or real.

Most of the approaches in this direction are based on supervised machine-learning techniques. Such is the case in [16], where the authors apply classic machine learning and deep learning algorithms for stance classification and rumour verification tasks. The obtained results conclude that classic algorithms outperform deep learning models for both tasks, and the information on stance does not enhance the rumour verification task.

Some works address the problem of filtering irrelevant content considering user-level information. In [12], Alrubaian et al. present a credibility analysis system for assessing information credibility on Twitter in order to avoid the increase of fake or malicious information. The proposal is based on four components that allow to analyse and assess the credibility of Twitter tweets and users (taking into account their reputation and experience). The authors use four machine learning algorithms to showcase the feasibility of the system achieving a significant balance between recall and precision.

The same authors propose in [13] a novel approach that analyses the user's reputation in the social network for a given topic. The proposal allows measuring the user's sentiment in order to recognise suitable and credible sources of information. The performance of the proposed reputation metric is evaluated with two machine-learning algorithms, concluding that the approach can identify credible Twitter users.

The main difference between our proposal and the other proposals is how the information is filtered. The approaches mentioned above tackle the problem using user-generated content, while our proposal addresses it using users' biographies, which are far more related to the user's work experience.

2.2. Big Data and credibility-based information filtering

Credibility-based information filtering has also been addressed from a Big Data perspective. Big data-based approaches include those works that deal with a massive amount of data and aim to reduce the amount of misinformation. Although the problem of filtering misinformation from large amounts of data has not yet received the attention it deserves [19], we can find some research in this direction.

That is the case in [14], where a Big Data-based solution is proposed. The system, called CredSaT (Credibility incorporating Semantic analysis and Temporal factor), considers the content and the temporal factor in order to build a ranking of proficient and significant users in the social network. It also adds a semantic analysis layer with sentiment analysis on tweets and responses used to enrich the final corpus of experts.

Diaz et al. [15] present a user-centred framework for filtering irrelevant content in social networks to facilitate data mining techniques post-usage. The proposed system, called NOFACE (NOise Filtering According Credibility and Expertise), also helps to reduce misinformation in social networks since it identifies credible and renowned members. The proposal relies on the fact that if a user is credible, then its content is credible too. The authors use word embeddings to capture the semantics of the texts in the user's profile. Then, these word embeddings enable the creation of a group of relevant users based on their expertise. In order to validate the framework, the experimentation was conducted using two datasets from Twitter (one related to COVID-19 and the other related to the United States elections) in a distributed environment (Big Data). In both cases, the system considerably reduces the number of irrelevant users, just considering users with higher expertise.

There are also other approaches to address credibility-related problems. A framework for managing extracted knowledge from big social media data for decision-making is proposed in [20]. The framework allows to extract relevant knowledge from social media by comparing different social media knowledge. In [21], the authors present an ontology-based approach for identifying the credibility domain in Social Big Data. They make use of ontologies in order to catch domain knowledge and enrich the semantics of the texts at both: entity and domain levels. Zhang et al. [22] develop a Scalable and Robust Truth Discovery (SRTD) scheme to tackle the misinformation spread and data sparsity challenges in social media in a distributed environment. The

approach quantifies both the trustworthiness of sources and the credibility of claims.

As stated before, this research is an extended version of the work presented in [5]. Unlike previous approaches, our proposal provides a Big Data-based flexible system capable of generalising and adapting to the user's needs. It also allows filtering content based on the user's experience. In this sense, the proposed system addresses two Big Data problems: the validity and volume of data.

3. Credibility filtering framework

Our credibility system is based on the premise that many professionals use Twitter biographies to inform about their jobs. These professionals are more suitable to be credible for a specific use case. For example, an economist could be more credible for a topic related to stock markets than a doctor, so we use word embedding to deal with that in an automatic way. This section focuses on details about our credibility filtering framework.

3.1. Twitter biographies

Nowadays, social media has an irreplaceable role in our lives. It's a fact that millions of professionals use social media apps daily, like Twitter, to disseminate their work or to be informed. In [23], Oksa et al. analysed social media usage from a work perspective, concluding that the usage of social networks for a professional is not only a fact but also brings improvements to the daily lives of workers in terms of networking, social support. According to [24], Twitter is one of the most whispered platforms to discuss recent advances or research in the field of medicine, but this success is also associated with more presence of misinformation or irrelevant content. Suppose we are browsing Twitter and we see some information that interests us. Probably the first thing we will do is to check that the account is reliable. For this, we will look at other data in the biographies since these are normally used to inform about jobs, likes or professions. So, in this paper, we propose an automatic system that exploits the potential of user biographies on Twitter and automatically discards accounts that do not fit with our desired analysis.

For example, suppose that we are interested in obtaining topics or clusters according to realistic medical information about COVID-19 from Twitter. For that, we need to obtain Tweets from people with some level of expertise in medicine. If we crawl Twitter data in order to apply a data mining

algorithm, we will filter according to the hashtag #COVID19. Many users have tweeted about that topic using that hashtag, but only a few of them are interesting because they have expertise in the topic. In Figure 1, we can see real accounts that have Tweeted about COVID that could be interesting for our topic. In contrast, in Figure 2, we have people who tweeted about COVID in a trivial form. Our algorithm will select and discard those in a massive and automatic way. These figures are examples of some accounts that our algorithm considers relevant and others that our algorithm discards.

Figure 1: Reliable users for medical topics

Figure 2: Unreliable users for medical topics

3.2. Expertise filter

In this section, we go into detail about the expertise filter algorithm. In Figure 3, a complete graph of the data flow throws our algorithm can be seen. The algorithm takes, as input, the directory where we find the CSV files with the Tweets, a list of searches related to the topic under study and the language we are interested in. The dataset has been divided into small parts in order to apply an adaptation of the divide-and-conquer methodology.

The first step of the algorithm is a pre-processing module. In this stage, the cleaning of every Tweet is carried out. For this, the algorithm eliminates

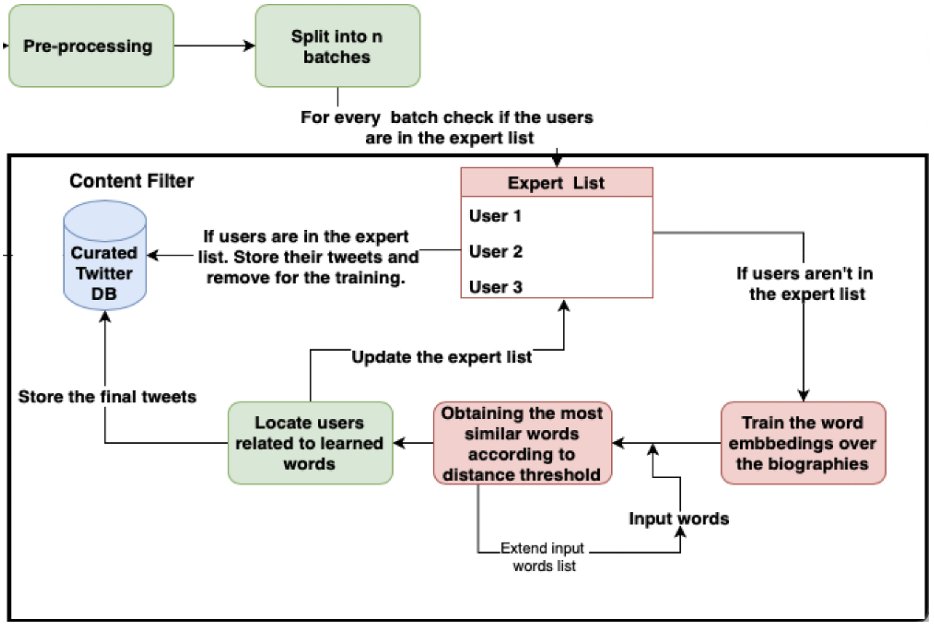

Figure 3: Algorithm data flow.

URLs, hashtags, mentions, reserved words from Twitter (RT, FAV...), emojis, smileys, numbers, additional spaces and punctuation marks. Following this, all the textual terms are turned into lowercase letters. After this, the dataset language is detected, and according to that language, we obtain the related stopwords. After that, the system eliminates these stop-words and all those Tweets using a non-recognised language or another language different to the one desired by the user. Finally, any empty Tweet (composed of eliminated items in previous stages of pre-processing) is removed, and Tweet texts are tokenized.

Then, the core of the algorithm starts to operate. The functionality of the algorithm has been introduced using a use case. Let's imagine a case related to COVID-19, where experts or people related to science and medicine are wanted. The process starts with a list of words related to medicine introduced by the user; for example *medical, doctor*. The algorithm will start to fit a word embedding model on the biographies of a part of a data partition (one of the input CSV files), and in the first iteration, it will obtain the most similar words to *medical* and *doctor* among the corpus itself according to a threshold. In the first version of our algorithm [5], the system obtained

the 5 most similar words in each iteration, but we realized that this could be derived from the exponential growth of the words. Also, in each iteration, the distance between the input and the selected words increases, so the words become worse. To mitigate that problem, a similarity threshold has been introduced in the algorithm, so the user can set a value for the similarity between words and iterate over all the batches. If any word has a higher similarity than the threshold, then it is selected.

Back to our use case, in the first iteration, most similar words to *doctor* and *medical* were *itresearcher, medicine, researcher, physician, epidemic, paediatrician, epidemiologist, paediatrics, postdoctoral* and *toxicologist*. The algorithm will use these 12 words (the most similar to medical and the most similar to doctor, besides doctor and medical) to find users whose biographies contain any of these terms and start creating the list of experts and topic-related users. In the next iteration, the retrieved words will be used to search for the most similar ones exceeding the threshold. This is one of the most flexible components of the system. Based on a very small set of one or two words, the algorithm automatically expands the query and flexibly locates words that could be very useful to filter the users according to their expertise in the related topic under study.

With the retrieved words, the algorithm selects the users who contain some of the words in their biographies, following the premise that biographies usually contain words related to professions. The selected users are people with some level of expertise in the topic, so we called them experts. We store the experts in the expert list. In each iteration, the algorithm checks if any user id is already present in the expert list to avoid processing it again since its words and content are already in the search corpus, thus avoiding additional processing. The output of the algorithm is a clean set of data in the form of a data frame ready to be processed in the following stages of the data mining pipeline. The complete pseudocode of the algorithm is in Algorithm 1.

4. Big Data architecture

With the increasing size of the data generated and stored, traditional Data Mining and data pre-processing techniques are facing a great challenge to process large data sets efficiently. For this reason, the use of distributed computing has been used as a solution even before the Big Data phenomenon.

This type of massive data processing does not only require adapting existing algorithms but also proposing new ones to handle Big Data problems.

MapReduce (MR) is one of the first distributed computing paradigms that allowed the generation and processing of Big Data datasets in an automatic and distributed way. MR has become the benchmark in distributed computing paradigms because of its simplicity and fault tolerance. By implementing two functions, *Map* and *Reduce*, users can process large amounts of data without worrying about technical issues such as data partitioning, fault recovery, or job communication. The algorithm we propose has been designed to enhance and use the full capacity of any data cluster using Spark. For this purpose, it has been implemented using several processes in which we use MapReduce.

To design the distributed algorithm for the credibility filter, some primitive Spark functions would be necessary. It is explained here:

- *Map*: Applies a transformation function to each element of RDD and returns a transformed RDD.
- *FlatMap*: Similar to Map, but each input item can be mapped to 0 or more output items.
- *Reduce*: Aggregates the elements of the dataset using an aggregation function.

Additionally, the algorithm uses broadcast variables to enable access to global variables in every node of the cluster, i.e. broadcast variables are available in every partition performed by the Map functions. The shared variables are stored in a hash table format, allowing direct and fast access to the query and insertion of the expert.

The algorithm has two main steps (see Algorithm 1). The first one consists of loading the dataset and filtering the experts from each dataset using the FlatMap function. The second step consists of aggregating all these data using a reduce function and grouping them by each found expert.

In the first stage of the algorithm, the algorithm needs the users and tweets with the information that will be processed by the *FindExpert()* function. For this purpose, a *FlatMap* function is used (see line 5 of the Algorithm 1 and the first column of Figure 4).

The FindExpert function is described in Algorithm 2, which filters the experts and returns the important information for each of them. The function

Algorithm 1 Main Spark procedure for expertise filter algorithm

1: **Input:** *Data:* RDD transactions: $\{t_1, \ldots, t_n\}$
2: **Input:** *similarity_threshold:* Similarity between words
3: **Output:** *Global_expert_set*: Expert discover in each $FlatMap$
4: Distributive computing in q chunks of transactions: $\{S_1, \ldots, S_q\}$
5: $\quad \{< Expert\_1 > \ldots < Expert\_m >\} \leftarrow S_i$.**FlatMap**(FindExpert())
6: $\quad FinalDataframe \leftarrow$ **ReduceByKey** $(getInfo())$
7: End distributive computation
8: **return** $FinalDataframe$

starts training a word embedding model on the biographies of the data partition finding in the first iteration, the 5 most similar words that have been passed as parameters (e.g. democratic, republican). The algorithm will use these 12 words (5 most similar to democratic, 5 most similar to republican, in addition to democratic and republican), to find users whose biographies contain any of these terms and start creating the list of experts that will be stored in the distributed variable *Global_expert_set* (line 3 of Algorithm 1). In the next iteration, these 12 words will be used to search for their 5 similar ones, and so on.

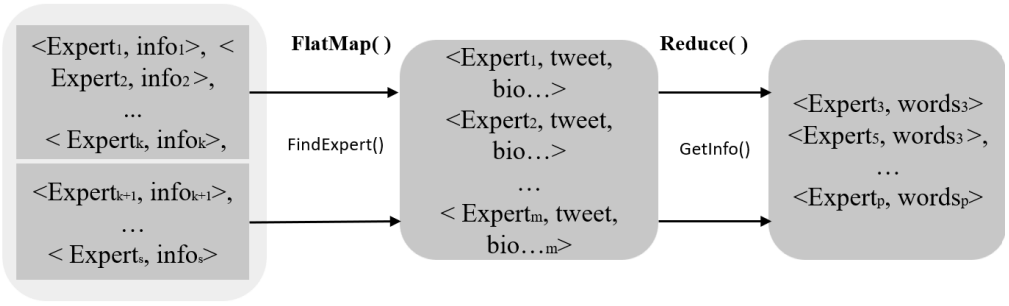

Figure 4: Workflow of Big Data process in Spark

The FindExpert function will return a list of peers containing the expert and its expert words and some other information. To aggregate and obtain all the information generated in a distributed manner, a *Reduce* function (see line 6 of Algorithm 1 and Figure 4) is used. As a result, a set of words associated with each expert is obtained.

Algorithm 2 FindExpert

Input: *datainput:* Tweet data and topics
Input: *similarity_threshold:* Similarity between words
Output: $KeyValuePair$: Topic and Expert
while $ExpertList \neq null$ **do**
    $Expert \leftarrow ExpertList.pop()$
    if $ExpertID \in Global\_expert\_set$ **then**
        # For experts, we add all their tweets to the final data frame
        $Expert_{final} \leftarrow < Expert_{id}, tweets >$
        $output.append(Expert_{final})$
    else
        # process the rest of the content to locate new experts
        $tokenized\_tweet = data[`biographies\_clean']$
        $model = train(tokenized\_tweet)$
        # create a list with the words of the list present in the model
        $final\_words = []$
        while $wordinexpert\_words$ **do**
            if $wordinmodel$ **then**
                $final\_words.append(word)$
            end if
        end while
        # create a data frame with the most similar words to each expert word according to the similarity threshold
        while $wordinfinal\_words$ **do**
            if $word.similarity > similarity\_threshold$ **then**
                $most\_similar.append(find\_most\_similar(model, word))$
            end if
        end while
        # expert word list extension with the most similar words
        $expert\_words.extend(most\_similar)$
        # location of all users having in their biographies any of the words
        $output.append(< Expert_{id}, [expert\_words, biographies\_clean, tweets] >)$
    end if
end while
return *Output*

5. Experiments

In this section, the experimentation has been carried out to analyse the system efficiency on 3 different use cases (sport, health and politics) in detail. The interpretation of the results will be discussed in Section 6.

5.1. Dataset

The system has been performed on 3 different real datasets without any processing of the social network Twitter according to the 3 mentioned use cases. To do so, the Twitter streaming API was used, without keywords, filtering only by the English language during the from 5 to 10 of August 2022. Taking into account the restrictions of the API in terms of time windows and that the data collection script was used intermittently during these days to avoid biases of topics that monopolised the network on a single day, the total number of tweets obtained is 5,000,000 in 20 different splits of 250,000 tweets.

The content of the dataset is not filtered in any way, i.e. it contains any tweet that tweeted in the time windows in which the data was collected. Therefore, it can be concluded that it is a noisy and very low-quality dataset.

5.2. Results

With the objective of validating the results and the improvement in terms of efficiency of the proposed big data system, in this section, comparative experimentation has been carried out between the sequential execution of the algorithm and the execution according to the big data paradigm proposed in Section 4. The technical specifications of the computer on which our experiments were run can be found in Table 1.

Component	Features
CPU	4 x 3 GHz Intel Xeon E5 with 8 cores
RAM	240 GB 3200 MHz LPDDR4X
Hard Disk	SATA SSD de 6 TB

Table 1: Machine specifications.

In the first version of the paper [5], has concluded that the best configuration in terms of a ratio between expert located and time, was Fast-Text [25] + Skip Gram. So, for the final version of our algorithm, this configuration has been chosen. Regarding the embedding parameters, it has been run with

a window of 5 words, words with frequencies lower than 2 have been ignored, and also having hierarchical softmax. The results using the mentioned seen configurations, for a mean score of 5 different executions for each use case, can be found in Table 2.

This table shows the experimental results on the entire dataset. As can be seen, the improvement in time of the sequential version versus the distributed version is remarkable. It is discussed in more detail by analysing the performance and efficiency. As for the localised experts, it can be seen how the system proposed is capable of extracting a subset by performing a 90% reduction of the dataset. As for the localised experts, how the system proposed is capable of extracting a subset by performing a 90% reduction of the dataset.

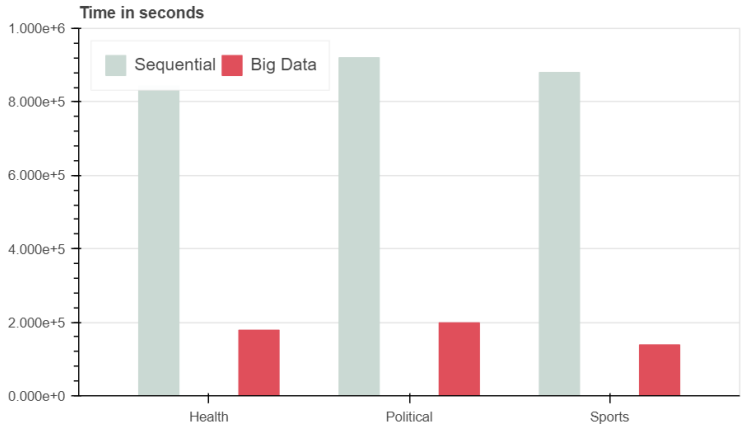

Figure 5: Ejemplo 5M

Use case	Time	Experts located	Final dataset size	% Reduction
Health	842350	72290	97342	98.05
Political	920235	98228	110293	97.7
Sports	879817	123342	185323	96.29
BigData-Health	178320	76170	89746	98.20
BigData-Political	198320	90674	137827	97.24
BigData-Sports	138320	110493	167425	96.65

Table 2: Comparison of results between sequential and big data executions

Figure 6 shows the memory usage of each algorithm for every use case. It can be observed that the distributed algorithm does not outperform the sequential case in all cases. This is due to the treatment of the data as they are replicated in different machines, which duplicates information.

With the purpose of measuring the efficiency of our proposal and comparing it to the existent approaches, it has been analysed the *speed up* and the *efficiency* [26, 27] according to the number of cores. For that, we have computed the well-known measure of speed up defined as [28]

$$S_n = T_1/T_n \tag{1}$$

where T_1 is the time of the sequential algorithm and T_n is the execution time of the distributed algorithm using several cores. The efficiency [26, 27, 28] is defined in a similar way as

$$E_n = S_n/n = T_1/(n \cdot T_n). \tag{2}$$

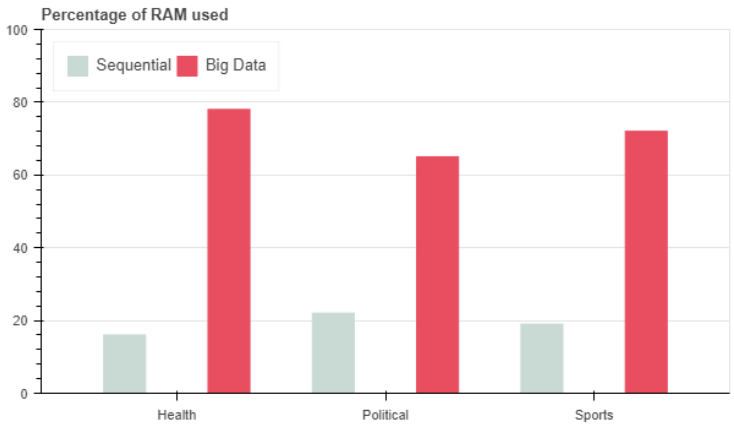

Figure 6: Percentage of memory used for the entire dataset

Figures 7 and 8 show the results obtained for the database with 2 million of Tweets. In them, it can be seen that the efficiency and speed improve as the number of cores increases, although they are not optimal. This is due to the workloads of the cores and the congestion of the network used for network communication between the cores. In addition, Figure 7 shows that the speed-up increases over the number of processors used, although it is not proportional as the resources increase (see also how the efficiency does not

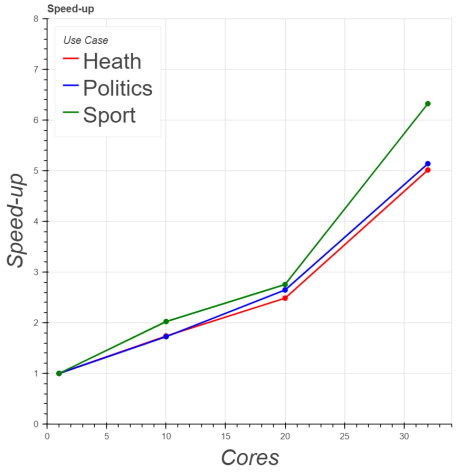

Figure 7: Speed-up of different use cases for 2 million of records

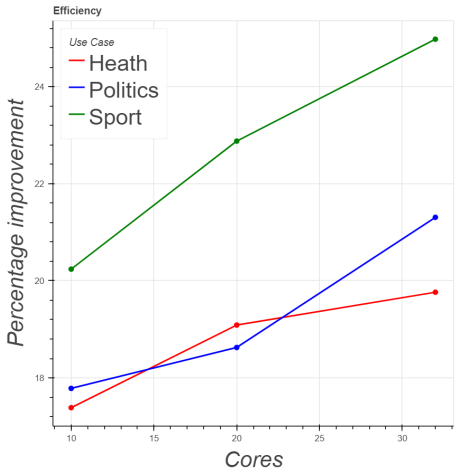

Figure 8: Efficiency of different use cases measured by the percentage of improvement for 2 million of records

increase in Figure 8). This same behaviour can be observed in other studies of speed up and efficiency in distributed algorithms, where the efficiency is not improved in a proportional way, as desired, with more processing cores [28].

Another advantage of the proposed algorithm is that it is based on a technology that allows the use of large data clusters in a simple way. Thanks to this, Larger clusters or cloud computing systems such as AWS or Google Cloud can be used if we need more capacity to process larger data sets.

In the results obtained, high-quality expert users on a specific topic have been identified using Big Data. This process is highly relevant in this context because one of the characteristics of Big Data is the volume of data. Using this filter, higher quality data can be validated within massive datasets. In the following section, we will analyse and validate some of the results obtained.

6. System validation

In order to validate the proper functionality of the system, we have applied the algorithm to our dataset with three different sets of input words. Each set of input words corresponds with a different use case, one related to health, another related to sports and another related to politics. These

different use cases demonstrate that the system is flexible and valid in multiple domains. Also, we demonstrate that our algorithm is suitable, in a very flexible way, to transform a non-quality and noisy dataset into a big-quality dataset. In Table 3, the input words and the top 20 learned words for each use case can be seen. In this point, we can derive one of the major contributions of our paper on the topic of flexible systems. If we pay attention to the learned words, it is easy to see how the algorithm retrieves closely related words to each domain in a semi-supervised way, only using a pair of input words. The system adjusts perfectly to each domain in a simple and effective way for the user who sees how his query is flexibly expanded according to the domain and with little effort on the user's part since only one or two words are needed as input.

Also, the proposed algorithm is able to mitigate one of the traditional problems in user-generated content, misspelling. Twitter biographies are user-generated text and usually contain typos or domain-related user-generated words, i.e. hashtags. Our algorithm can link those words with the topic. For example, in the case of sports, some interesting words that the algorithm used to filter the content are *fuball* or *fotball*, two misspelling words of *football*. Also, in the case of politics, we find *conserv* or *repub*, misspelling words or *resistbiden* and *jailtrump*, two user-generated words for political discussions in Twitter. That result led us to argue that the system is flexible and robust.

Using the set of words learned for each use case, the algorithm can filter the Twitter accounts and select only those users that fit with the topic. With that, from a non-quality and hyper-generic dataset, we can obtain valuable information. In order to prove that, we obtain word clouds over the filtered datasets. We try to prove the premise that if the system performs properly, visualization techniques over the filtered dataset must obtain closely-related topic words. We must bear in mind that from this moment on, we always refer to the textual content of tweets. The biographies have been used by the algorithm to obtain the most reliable accounts according to their expertise on the topic. Now, we only focus on the content of the Tweet, trying to demonstrate that the filtering process proposed in this paper improves the quality of the dataset. The word clouds have been created using two different corpus representations, one based on TF-IDF [29] and the other one based on traditional term frequency. To create the word clouds, first, we obtained the named entities of the content of the Tweet, and for that, we used two different models. Due to the specificity of the medical domain for the health use case, we have used the *ner_bionlp13cg_md* model included in SciSpacy

Use case	Input words	Top 20 earned words
Health	doctor, researcher	research, doctor, surgeon, researcher, medicine, medical, lecturer, clinical, physician, epidemic, paediatrics, epidemiologist, exclinical, postdoctoral, toxicologist, biologist, physiologist, cardiology, aesthetician, virologist
Sports	football, baseball	football, baseball, softball, redsox, hockey, basketball, football, sportsnet, volleyball, rugby, jockey, celtic, coach, intense, whitesox, paintball, handbasket, cricket, lebron, sport
Political	democratic, republican	democrat, socialist, trump, democratic, secular, republican, protrump, repubs, liberal, conservative, humanist, socialism, joebiden, exlabour, ecosocialist, freethinker, autocratic, exrepublican, demsocialist, conservatism

Table 3: Learned words for each use case

[30]. That model was trained on the BIONLP13CG corpus. Regarding the political and sports, which are use cases with more generic domains, we have used the *en_core_web_sm* included in SpaCy [31]. That was trained using web blogs, news and comments.

In Figure 9 and Figure 10, we show the results for the health use case. In Figure 11 and Figure 12, show the results for the political use case. Finally, Figure 13 and Figure 14 contain the results for the sports use case.

If we pay attention to the word clouds, we can see how we can easily locate topic-related words in each example. There are a few examples that appear in all the figures, if we take into account that all the results come from the same dataset, the low value of common words and the topic-related words present in each use case led us to argue that our filter works in a very

proper way and it can discard noise and irrelevant information from a big data set. Also, some words can be present in all the domains, but this is due to the bias of the data acquisition stage, and because these words like, for instance, *people*, or *american* or some verbs are very common in Twitter's conversations.

Figure 9: Word cloud for the health use-case using the frequency of words

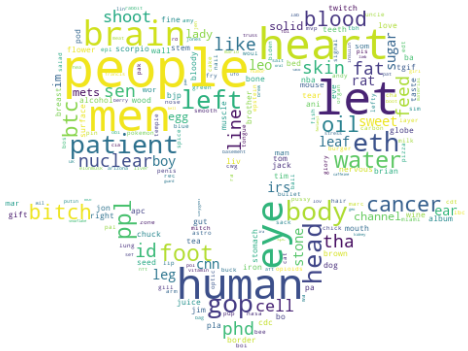

Figure 10: Word cloud for the health use-case using the TF-IDF value

Another interesting analysis that we can obtain is related to word frequency. In the case of health, Figures 9 and 12 are very similar due to the specification of the NER process in this topic. In the other uses cases, the frequency and TF-IDF figures are too different, and maybe in the first look at Figure 11 and Figure 13, we can think that the algorithm does not work properly. In a deeper analysis, we browsed the social network Twitter trying to find the meaning of that words, and we realized that they were usernames. The vast majority of these user names come from users related to sports brands, basket players and people related to sports in the case of sports. And users who are politicians, companies or journalists in the case of politics. In Figure 15 and , we can see the most representative users located in the conversations of Twitter for each use case.

Figure 11: Word cloud for the political use-case using the frequency of words

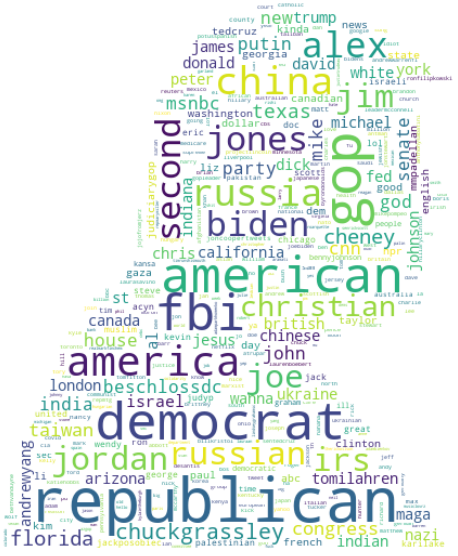

Figure 12: Word cloud for the political use-case using the TF-IDF value

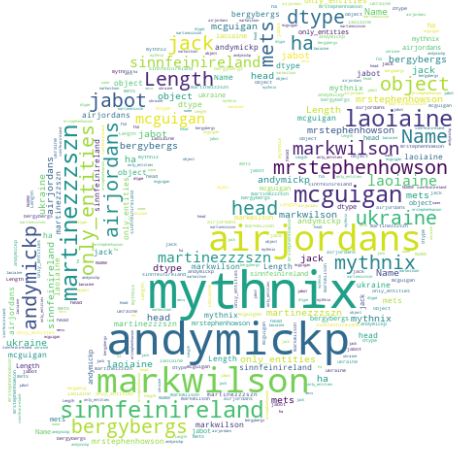

Figure 13: Word cloud for the sports use-case using the frequency of words

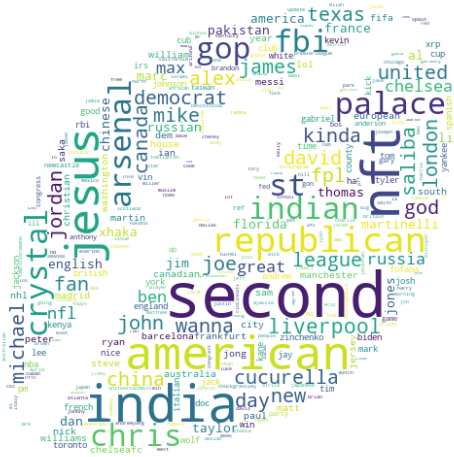

Figure 14: Word cloud for the sports use-case using the TF-IDF value

Considering that the dataset was obtained without any filter having as results data sets in which we can easily obtain closely related topic informa-

tion and find people that usually participate in the conversations about each topic is a very important result. It is interesting to note how some accounts located in the Twitter conversations by the entire pipeline do not have any information in the user names or biographies about the topic, but yes in the images or main user image. This again led us to argue the potential of the framework to locate interesting information for hundreds of analyses in a flexible and non-supervised way. Based on that findings, we can conclude that our filter works properly and can get value and information regarding interesting topic-related conversations and users from a non-quality and noisy dataset.

7. Conclusions and future work

In this paper, we have proposed a flexible, easy-to-use, easy-to-understand and easy-to-replicate big data system for filtering the content coming from Twitter based on the credibility of the user account based on his or her expertise in a certain topic. We have conducted exhaustive experimentation, in which we corroborate how the big data architecture proposed can improve in terms of time the results of the traditional processing. We also have conducted three different uses cases, in which we have demonstrated the following:

- The proposed system can help to acquire more valuable datasets in a flexible and unsupervised way.
- The proposed system can help to reduce the dimensionality of a big data set, maintaining only the interesting examples for the desired use case.
- The proposed system can help to improve the validity of the data in terms of expertise.

The system is sensitive to bias in data acquisition. Sometimes time windows can be more interesting than other windows, depending on which moment we collect the data to obtain better o worse results. In this paper, we wanted to demonstrate how our algorithm can obtain a curated dataset over a noisy dataset, so we create a real word dataset using time windows without content filters. But, real word applications, usually the data acquisition, are made with filters, for example, using hashtags, so these applications are

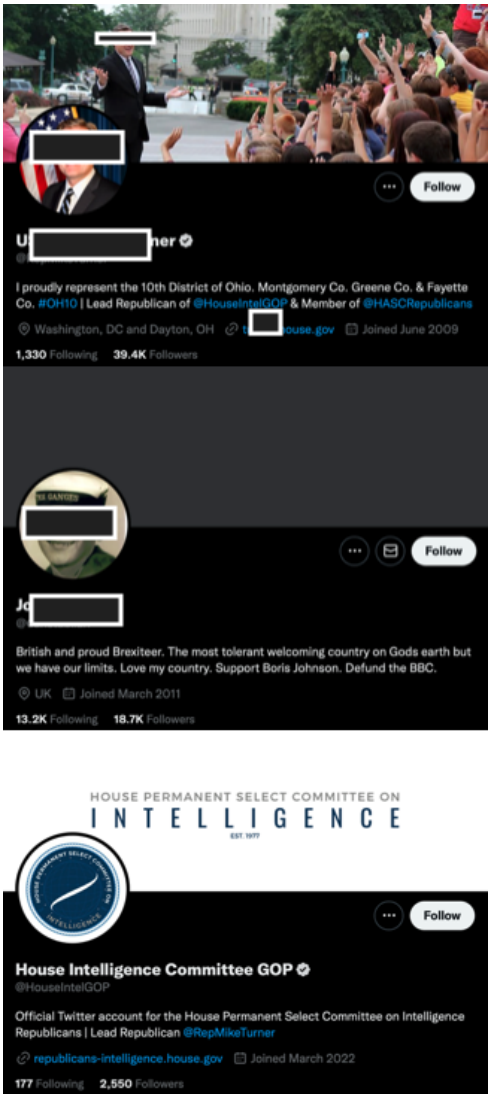

Figure 15: Some users mined by the filter and the pipeline for the political use-case

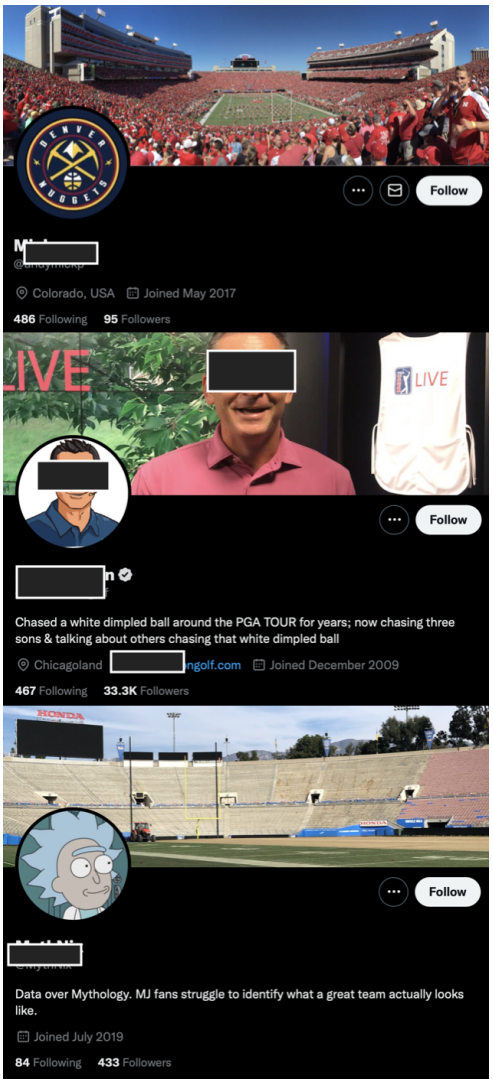

Figure 16: Some users mined by the filter and the pipeline for the sports use-case

independent of the time windows. In these problems, our system can also improve the subsequent data mining process [15]. In these real-world applications with filters in data capture, the systems obtain even good results, discarding people without expertise in the topic and maintaining very useful accounts in which we can perform more accurate analysis. Regarding the

limitations of the system, since the core of the system (Algorithm 1) is based on Twitter biographies, it is very sensitive to lies in biographies.

In terms of future work, the system can be adapted to a streaming application that can be used in an incremental learning pipeline filtering the irrelevant content for the topic under study and helping to obtain more credible results. This will be very useful in real-world applications for facing the misinformation spread on social networks like Twitter.

8. Declarations

Abbreviations:

- COVID: Coronavirus Disease.
- NLP: Natural Language Processing.
- KNN: K Nearest Neighbours.
- TF: Term Frequency.
- TF-IDF: Term Frequency–Inverse Document Frequency
- NOFACE: NOise Filtering According Credibility and Expertise.
- SRTD: Scalable and Robust Truth Discovery.
- FAV: Favourite.
- RT: ReTweet.
- URL: Uniform Resource Locator.
- CSV: Comma-Separated Values
- MR: MapReduce.
- RDD: Resilient Distributed Datasets
- API: Application Programming Interface
- AWS: Amazon Web Services.

Funding: The research presented in this paper has received funding from:

- The BIGDATAMED projects with references B-TIC-145-UGR18 and P18-RT-2947.
- The European Union NextGenerationEU / PRTR, grant PLEC2021-007681 funded by MCIN / AEI / 10.13039 / 501100011033.
- The DESINFOSCAN project. Ministerio de Ciencia e Innovacion and by the European Union NextGenerationEU (Grant TED2021-1289402B-C21).
- The NOFACEPS project (PPJIB2021-04) of the University of Granada's internal plan.
- Carlos Fernandez-Basso was supported by the Ministry of Universities through the EU-funded Margarita Salas Programme.
- Jose A. Diaz-Garcia was supported by the Spanish Ministry of Education, Culture and Sport (FPU18/00150).
- Karel Gutiérrez-Batista was supported by the Administration of the Junta de Andalucía.

Conflict of interest/Competing interests: The authors declare that they have no conflict of interest.

Ethics approval No ethical approval is required for this study.

Consent to participate Not applicable

Availability of data and material Data will be available on request.

Authors' contributions:

- Jose A Diaz-Garcia: Supervision, Investigation, Project administration, Writing – original draft & editing
- Carlos Fernandez-Basso: Investigation, Software, Writing – original draft.
- Karel Guitierrez-Batista: Investigation, Software, Writing – original draft.
- M. Dolores Ruiz: Funding acquisition, Project administration, Writing – review & editing

- Maria J. Martin-Bautista: Funding acquisition, Project administration, Writing – review & editing

9. Acknowledgments

The research reported in this paper was partially supported by the Andalusian government and the FEDER operative program under the project BigDataMed (P18-RT-2947 and B-TIC-145-UGR18) and grant PLEC2021-007681 funded by MCIN / AEI / 10.13039 / 501100011033 and by the European Union NextGenerationEU / PRTR. Also, the research is part of DESINFOSCAN project, founded by Ministerio de Ciencia e Innovacion and by the European Union NextGenerationEU (Grant TED2021-1289402B-C21). The paper is part of the NOFACEPS project (PPJIB2021-04) of the University of Granada's internal plan. Finally, the project is also partially supported by the Spanish Ministry of Education, Culture and Sport (FPU18/00150).

References

[1] A. Perrin, Social media usage, Pew research center 125 (2015) 52–68.

[2] B. Batrinca, P. C. Treleaven, Social media analytics: a survey of techniques, tools and platforms, Ai & Society 30 (1) (2015) 89–116.

[3] S. Li, F. Liu, Y. Zhang, B. Zhu, H. Zhu, Z. Yu, Text mining of user-generated content (ugc) for business applications in e-commerce: A systematic review, Mathematics 10 (19) (2022). doi:10.3390/math10193554.

[4] M. Assefi, E. Behravesh, G. Liu, A. P. Tafti, Big data machine learning using apache spark mllib, in: 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3492–3498. doi:10.1109/BigData.2017.8258338.

[5] J. A. Diaz-Garcia, M. D. Ruiz, M. J. Martin-Bautista, A comparative study of word embeddings for the construction of a social media expert filter, in: International Conference on Flexible Query Answering Systems, Springer, 2021, pp. 196–208.

[6] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica, Apache spark: A unified engine for big data processing, Commun. ACM 59 (11) (2016) 56–65. doi:10.1145/2934664.

[7] M. Honnibal, I. Montani, Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, Unpublished software application. https://spacy. io (2017).

[8] R. Sharnagat, Named entity recognition: A literature survey, Center For Indian Language Technology (2014) 1–27.

[9] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th international conference on World wide web, 2011, pp. 675–684.

[10] K. R. Canini, B. Suh, P. L. Pirolli, Finding credible information sources in social networks based on content and social structure, in: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, IEEE, 2011, pp. 1–8.

[11] D. Hassan, A text mining approach for evaluating event credibility on twitter, in: 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), IEEE, 2018, pp. 171–174.

[12] M. Alrubaian, M. Al-Qurishi, M. M. Hassan, A. Alamri, A credibility analysis system for assessing information on twitter, IEEE Transactions on Dependable and Secure Computing 15 (4) (2016) 661–674.

[13] M. Alrubaian, M. Al-Qurishi, M. Al-Rakhami, M. M. Hassan, A. Alamri, Reputation-based credibility analysis of twitter social network users, Concurrency and Computation: Practice and Experience 29 (7) (2017) e3873.

[14] B. Abu-Salih, P. Wongthongtham, K. Y. Chan, D. Zhu, Credsat: Credibility ranking of users in big social data incorporating semantic analysis and temporal factor, Journal of Information Science 45 (2) (2019) 259–280.

[15] J. A. Diaz-Garcia, M. D. Ruiz, M. J. Martin-Bautista, Noface: A new framework for irrelevant content filtering in social media according to credibility and expertise, Expert Systems with Applications (2022) 118063.

[16] P. R. D. Cordeiro, V. Pinheiro, R. Moreira, C. Carvalho, L. Freire, What is real or fake?-machine learning approaches for rumor verification using stance classification, in: IEEE/WIC/ACM International Conference on Web Intelligence, 2019, pp. 429–432.

[17] F. A. Ozbay, B. Alatas, Fake news detection within online social media using supervised artificial intelligence algorithms, Physica A: Statistical Mechanics and its Applications 540 (2020) 123174.

[18] J. Batra, R. Jain, V. A. Tikkiwal, A. Chakraborty, A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques, International Journal of Information Management Data Insights 1 (1) (2021) 100006.

[19] M. Viviani, G. Pasi, Credibility in social media: opinions, news, and health information—a survey, Wiley interdisciplinary reviews: Data mining and knowledge discovery 7 (5) (2017) e1209.

[20] W. He, F.-K. Wang, V. Akula, Managing extracted knowledge from big social media data for business decision making, Journal of Knowledge Management (2017).

[21] P. Wongthongtham, B. A. Salih, Ontology-based approach for identifying the credibility domain in social big data, Journal of Organizational Computing and Electronic Commerce 28 (4) (2018) 354–377.

[22] D. Zhang, D. Wang, N. Vance, Y. Zhang, S. Mike, On scalable and robust truth discovery in big data social media sensing applications, IEEE transactions on big data 5 (2) (2018) 195–208.

[23] R. Oksa, M. Kaakinen, N. Savela, N. Ellonen, A. Oksanen, Professional social media usage: Work engagement perspective, New media & society 23 (8) (2021) 2303–2326.

[24] Y. Pershad, P. T. Hangge, H. Albadawi, R. Oklu, Social medicine: Twitter in healthcare, Journal of clinical medicine 7 (6) (2018) 121.

[25] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.

[26] V. P. Kumar, A. Gupta, Analyzing scalability of parallel algorithms and architectures, Journal of parallel and distributed computing 22 (3) (1994) 379–391.

[27] A. Y. Grama, A. Gupta, V. Kumar, Isoefficiency: Measuring the scalability of parallel algorithms and architectures, IEEE Parallel & Distributed Technology: Systems & Applications 1 (3) (1993) 12–21.

[28] C. Barba-González, J. García-Nieto, A. Benítez-Hidalgo, A. J. Nebro, J. F. Aldana-Montes, Scalable inference of gene regulatory networks with the spark distributed computing platform, in: J. Del Ser, E. Osaba, M. N. Bilbao, J. J. Sanchez-Medina, M. Vecchio, X.-S. Yang (Eds.), Intelligent Distributed Computing XII, Springer International Publishing, Cham, 2018, pp. 61–70.

[29] S. Qaiser, R. Ali, Text mining: use of tf-idf to examine the relevance of words to documents, International Journal of Computer Applications 181 (1) (2018) 25–29.

[30] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 319–327. doi:10.18653/v1/W19-5034.

[31] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python (2020). doi:10.5281/zenodo.1212303.

Bibliography

[AAHA18] Alrubaian M., Al-Qurishi M., Hassan M. M., and Alamri A. (2018) A credibility analysis system for assessing information on twitter. *IEEE Transactions on Dependable and Secure Computing* 15(4): 661–674.

[AAQA+16] Alrubaian M., AL-Qurishi M., Alrakhami M., Hassan M., and Alamri A. (01 2016) Reputation-based credibility analysis of twitter social network users: Reputation-based credibility analysis of twitter social network users. *Concurrency and Computation: Practice and Experience* 29.

[ADEZ18] Abu Daher L., Elkabani I., and Zantout R. (09 2018) Identifying influential users on twitter: A case study from paris attacks. *Applied Mathematics and Information Sciences* 12: 1021–1032.

[AIS93] Agrawal R., Imieliński T., and Swami A. (1993) Mining association rules between sets of items in large databases. In *Acm sigmod record*, volumen 22, pp. 207–216. ACM.

[AKI19] Aswani R., Kar A. K., and Ilavarasan P. V. (2019) Experience: managing misinformation in social media—insights for policymakers from twitter analytics. *Journal of Data and Information Quality (JDIQ)* 12(1): 1–18.

[AMA14] Al-Maolegi M. and Arkok B. (2014) An improved apriori algorithm for association rules. *arXiv preprint arXiv:1403.3948* .

[AOGD+16] Adedoyin-Olowe M., Gaber M. M., Dancausa C. M., Stahl F., and Gomes J. B. (2016) A rule dynamics approach to event detection in twitter with its application to sports and politics. *Expert Systems with Applications* 55: 351–360.

[AS+94] Agrawal R., Srikant R., *et al.* (1994) Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. very large data bases, VLDB*, volumen 1215, pp. 487–499.

[ASWCZ19] Abu-Salih B., Wongthongtham P., Chan K. Y., and Zhu D. (2019) Credsat: Credibility ranking of users in big social data incorporating semantic analysis and temporal factor. *Journal of Information Science* 45(2): 259–280.

[AZ12] Aggarwal C. C. and Zhai C. (2012) A survey of text clustering algorithms. *Mining text data* pp. 77–128.

[BAI19] Barbado R., Araque O., and Iglesias C. A. (2019) A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management* 56(4): 1234–1244.

[BB06] Budgen D. and Brereton P. (2006) Performing systematic literature reviews in software engineering. In *Proceedings of the 28th international conference on Software engineering*, pp. 1051–1052.

[BCO14] Bing L., Chan K. C., and Ou C. (2014) Public sentiment analysis in twitter data for prediction of a company's stock price movements. In *2014 IEEE 11th International Conference on e-Business Engineering*, pp. 232–239. IEEE.

[BDV00] Bengio Y., Ducharme R., and Vincent P. (2000) A neural probabilistic language model. *Advances in neural information processing systems* 13.

[BGJM16] Bojanowski P., Grave E., Joulin A., and Mikolov T. (2016) Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

[BJTC21] Batra J., Jain R., Tikkiwal V. A., and Chakraborty A. (2021) A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques. *International Journal of Information Management Data Insights* 1(1): 100006.

[CBS23] (Month 2023) Twitter rebrand: What the new name and elon musk's role mean for the social media giant. `https://www.cbsnews.com/news/twitter-rebrand-x-name-change-elon-musk-what-it-means/`. CBS News.

[CF12] Cagliero L. and Fiori A. (2012) Analyzing twitter user behaviors and topic trends by exploiting dynamic rules. In *Behavior Computing*, pp. 267–287. Springer.

[CF13] Cagliero L. and Fiori A. (01 2013) Discovering generalized association rules from twitter. *Intelligent Data Analysis* 17.

[CMP11] Castillo C., Mendoza M., and Poblete B. (2011) Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pp. 675–684.

[CPM+19] Cordeiro P. R. D., Pinheiro V., Moreira R., Carvalho C., and Freire L. (2019) What is real or fake?-machine learning approaches for rumor verification using stance classification. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 429–432.

[CS14] Chandrashekar G. and Sahin F. (2014) A survey on feature selection methods. *Computers & Electrical Engineering* 40(1): 16–28.

[CSP11] Canini K. R., Suh B., and Pirolli P. L. (2011) Finding credible information sources in social networks based on content and social structure. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 1–8. IEEE.

[CU21] Culmer K. and Uhlmann J. (2021) Examining lda2vec and tweet pooling for topic modeling on twitter data. *Wseas Trans. Inf. Sci. Appl.* 18: 102–115.

[DB79] Davies D. L. and Bouldin D. W. (1979) A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1(2): 224–227.

[DCLT18] Devlin J., Chang M.-W., Lee K., and Toutanova K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

[DGFBRMB20] Díaz-García J. A., Fernandez-Basso C., Ruiz M. D., and Martin-Bautista M. J. (2020) Mining text patterns over fake and real tweets. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 648–660. Springer.

[DGRMB20] Diaz-Garcia J. A., Ruiz M. D., and Martin-Bautista M. J. (2020) Non-query-based pattern mining and sentiment analysis for massive microblogging online texts. *IEEE Access* 8: 78166–78182.

[DGRMB22] Diaz-Garcia J. A., Ruiz M. D., and Martin-Bautista M. J. (2022) Noface: A new framework for irrelevant content filtering in social media according to credibility and expertise. *Expert Systems with Applications* 208: 118063.

[DGRMB23] Diaz-Garcia J. A., Ruiz M. D., and Martin-Bautista M. J. (2023) A survey on the use of association rules mining techniques in textual social media. *Artificial Intelligence Review* 56(2): 1175–1200.

[DMC16] Diaz F., Mitra B., and Craswell N. (2016) Query expansion with locally-trained word embeddings.

[DMJS14] Dehkharghani R., Mercan H., Javeed A., and Saygin Y. (2014) Sentimental causal rule discovery from twitter. *Expert Systems with Applications* 41(10): 4950–4958.

[DSRO20] Da'u A., Salim N., Rabiu I., and Osman A. (2020) Recommendation system exploiting aspect-based opinion mining with deep learning method. *Information Sciences* 512: 1279–1292.

[EBBJ16] Erlandsson F., Bródka P., Borg A., and Johnson H. (2016) Finding influential users in social media using association rule learning. *Entropy* 18(5): 164.

[Fac23] (July 2023) Introducing threads: A new app for text sharing. `https://about.fb.com/news/2023/07/introducing-threads-new-app-text-sharing/`. Facebook Newsroom.

[FBFAMBR19] Fernandez-Basso C., Francisco-Agra A. J., Martin-Bautista M. J., and Ruiz M. D. (2019) Finding tendencies in streaming data using big data frequent itemset mining. *Knowledge-Based Systems* 163: 666–674.

[FBRMB16] Fernandez-Basso C., Ruiz M. D., and Martin-Bautista M. J. (2016) Extraction of association rules using big data technologies. *International Journal of Design & Nature and Ecodynamics* 11(3): 178–185.

[FJUA18] Farooq A., Joyia G. J., Uzair M., and Akram U. (2018) Detection of influential nodes using social networks analysis based on network metrics. In *2018 international conference on computing, mathematics and engineering technologies (icomet)*, pp. 1–6. IEEE.

[FKR+20] Fayaz M., Khan A., Rahman J. U., Alharbi A., Uddin M. I., and Alouffi B. (2020) Ensemble machine learning model for classification of spam product reviews. *Complexity* 2020: 1–10.

[FPSS96] Fayyad U., Piatetsky-Shapiro G., and Smyth P. (1996) The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39(11): 27–34.

[GSB+12] Ghosh S., Sharma N., Benevenuto F., Ganguly N., and Gummadi K. (2012) Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 575–590.

[Has18] Hassan D. (2018) A text mining approach for evaluating event credibility on twitter. In *2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pp. 171–174. IEEE.

[HK17] Hahsler M. and Karpienko R. (2017) Visualizing association rules in hierarchical groups. *Journal of Business Economics* 87(3): 317–335.

[HPY00] Han J., Pei J., and Yin Y. (2000) Mining frequent patterns without candidate generation. In *ACM sigmod record*, volumen 29, pp. 1–12. ACM.

[JK21] Joung J. and Kim H. M. (2021) Automated keyword filtering in latent dirichlet allocation for identifying product attributes from online reviews. *Journal of Mechanical Design* 143(8): 084501.

[KA21] Kar A. K. and Aswani R. (2021) How to differentiate propagators of information and misinformation–insights from social media analytics based on bio-inspired computing. *Journal of Information and Optimization Sciences* 42(6): 1307–1335.

[KAGE21] Kumari R., Ashok N., Ghosal T., and Ekbal A. (2021) Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. *Information Processing & Management* 58(5): 102631.

[Kal18] Kaliyar R. K. (2018) Fake news detection using a deep neural network. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pp. 1–7. IEEE.

[KCQJ20] Khoo L. M. S., Chieu H. L., Qian Z., and Jiang J. (2020) Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volumen 34, pp. 8783–8790.

[KJ20] Kumar A. and Jaiswal A. (2020) Systematic literature review of sentiment analysis on twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience* 32(1): e5107.

[KJMH+19] Kowsari K., Jafari Meimandi K., Heidarysafa M., Mendu S., Barnes L., and Brown D. (2019) Text classification algorithms: A survey. *Information* 10(4): 150.

[KKI21] Kumar S., Kar A. K., and Ilavarasan P. V. (2021) Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights* 1(1): 100008.

[KM13] Kodinariya T. M. and Makwana P. R. (2013) Review on determining number of cluster in k-means clustering. *International Journal* 1(6): 90–95.

[KOH12] Kang B., O'Donovan J., and Höllerer T. (2012) Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pp. 179–188.

[KRA20] Kamaran H. M., Ramadhan R., and Amin P. (2020) Twitter sentiment analysis on worldwide covid-19 outbreaks. *Kurdistan Journal of Applied Research* 5: 54–65.

[KSK16] Kuzi S., Shtok A., and Kurland O. (2016) Query expansion using word embeddings. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pp. 1929–1932.

[KYK+19] Khan A., Younis U., Kundi A. S., Asghar M. Z., Ullah I., Aslam N., and Ahmed I. (2019) Sentiment classification of user reviews using supervised learning techniques with comparative opinion mining perspective. In *Science and Information Conference*, pp. 23–29. Springer.

[Lam20] Lamsal R. (2020) Coronavirus (covid-19) tweets dataset.

[LG14] Levy O. and Goldberg Y. (2014) Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 302–308.

[LLCS15] Liu Y., Liu Z., Chua T.-S., and Sun M. (2015) Topical word embeddings. In *Twenty-ninth AAAI Conference on Artificial Intelligence*. Citeseer.

[LVV03] Likas A., Vlassis N., and Verbeek J. J. (2003) The global k-means clustering algorithm. *Pattern recognition* 36(2): 451–461.

[LW20] Liu Y. and Wu Y.-F. B. (2020) Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)* 38(3): 1–33.

[LZ12] Liu B. and Zhang L. (2012) A survey of opinion mining and sentiment analysis. In *Mining text data*, pp. 415–463. Springer.

[MBSSV04] Martin-Bautista M., Sánchez D., Serrano J., and Vila M. (2004) Text mining using fuzzy association rules. In *Fuzzy logic and the internet*, pp. 173–189. Springer.

[MCCD13] Mikolov T., Chen K., Corrado G., and Dean J. (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

[MFE+19] Monti F., Frasca F., Eynard D., Mannion D., and Bronstein M. M. (2019) Fake news detection on social media using geometric deep learning.

[MGM+16] Ma J., Gao W., Mitra P., Kwon S., Jansen B. J., Wong K.-F., and Cha M. (2016) Detecting rumors from microblogs with recurrent neural networks.

[MH08] Maaten L. v. d. and Hinton G. (2008) Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov): 2579–2605.

[MK19] Mohammed A. and Kora R. (2019) Deep learning approaches for arabic sentiment analysis. *Social Network Analysis and Mining* 9(1): 52.

[MLD19] Medvedev A. N., Lambiotte R., and Delvenne J.-C. (2019) The anatomy of reddit: An overview of academic research. *Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches 10* pp. 183–204.

[MPM16] Mamgain N., Pant B., and Mittal A. (2016) Categorical data analysis and pattern mining of top colleges in india by using twitter data. In *2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 341–345. IEEE.

[MSADLG18] Molina-Solana M., Amador Diaz Lopez J., and Gomez J. (2018) Deep learning for fake news classification. In *I Workshop in Deep Learning, 2018 conference spanish association of artificial intelligence*, pp. 1197–1201.

[MSB+14] Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., and McClosky D. (2014) The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System demonstrations*, pp. 55–60.

[MSC+13] Mikolov T., Sutskever I., Chen K., Corrado G. S., and Dean J. (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26: 3111–3119.

[MSSS20] Mehta R. P., Sanghvi M. A., Shah D. K., and Singh A. (2020) Sentiment analysis of tweets using supervised learning algorithms. In *First International Conference on Sustainable Technologies for Computational Intelligence*, pp. 323–338. Springer.

[MT13] Mohammad S. M. and Turney P. D. (2013) Nrc emotion lexicon. *National Research Council, Canada* 2: 234.

[NKV21] Nasir J. A., Khan O. S., and Varlamis I. (2021) Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights* 1(1): 100007.

[OA20] Ozbay F. A. and Alatas B. (2020) Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and its Applications* 540: 123174.

[OHA+19] Oehmichen A., Hua K., Amador Dííaz López J., Molina-Solana M., Gómez-Romero J., and Guo Y. (2019) Not all lies are equal. a study into the engineering of political misinformation in the 2016 us presidential election. *IEEE Access* 7: 126305–126314.

[OLCOMTK10] Olvera-López J. A., Carrasco-Ochoa J. A., Martínez-Trinidad J. F., and Kittler J. (2010) A review of instance selection methods. *Artificial Intelligence Review* 34(2): 133–143.

[OZP+97] Ogihara Z. P., Zaki M., Parthasarathy S., Ogihara M., and Li W. (1997) New algorithms for fast discovery of association rules. In *In 3rd Intl. Conf. on Knowledge Discovery and Data Mining*. Citeseer.

[PMNL21] Pradeep R., Ma X., Nogueira R., and Lin J. (2021) Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2066–2070.

[PNH18] Phan H. T., Nguyen N. T., and Hwang D. (2018) A tweet summarization method based on maximal association rules. In *International Conference on Computational Collective Intelligence*, pp. 373–382. Springer.

[RCE+11] Ritter A., Clark S., Etzioni O., *et al.* (2011) Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 1524–1534.

[Rob04] Robertson S. (2004) Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation* .

[Rou87] Rousseeuw P. (Noviembre 1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20(1): 53–65.

[RPMG16] Roy D., Paul D., Mitra M., and Garain U. (2016) Using word embeddings for automatic query expansion.

[SA97] Srikant R. and Agrawal R. (1997) Mining generalized association rules. *Future Generation Computer Systems* 13(2): 161 – 180.

[SGAO14] Stahl F., Gaber M. M., and Adedoyin-Olowe M. (2014) A survey of data mining techniques for social network analysis. *Journal of Data Mining and Digital Humanities* 18.

[SSW+17] Shu K., Sliva A., Wang S., Tang J., and Liu H. (2017) Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19(1): 22–36.

[TAN19] Tsapatsoulis N., Anastasopoulou V., and Ntalianis K. (2019) The central community of twitter ego-networks as a means for fake influencer detection. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pp. 177–184. IEEE.

[TF20] Tao J. and Fang X. (2020) Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data* 7(1): 1–26.

[Wil45] Wilcoxon F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin* 1(6): 80–83.

[WLP+09] Wu H., Lu Z., Pan L., Xu R., and Jiang W. (2009) An improved apriori-based algorithm for association rules mining. In *2009 sixth international conference on fuzzy systems and knowledge discovery*, volumen 2, pp. 51–55. IEEE.

[YOXS13] Yuan M., Ouyang Y., Xiong Z., and Sheng H. (2013) Sentiment classification of web review using association rules. In *International Conference on Online Communities and Social Computing*, pp. 442–450. Springer.

[ZLG20] Zulli D., Liu M., and Gehl R. (2020) Rethinking the "social" in "social media": Insights into topology, abstraction, and scale on the mastodon social network. *New Media & Society* 22(7): 1188–1205.

[ZVTAS22] Zhang D., Vakili Tahami A., Abualsaud M., and Smucker M. D. (2022) Learning trustworthy web sources to derive correct answers and reduce health misinformation in search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2099–2104.

[ZWL18] Zhang L., Wang S., and Liu B. (2018) Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(4): e1253.

[ZXW+16] Zaharia M., Xin R. S., Wendell P., Das T., Armbrust M., Dave A., Meng X., Rosen J., Venkataraman S., Franklin M. J., *et al.* (2016) Apache spark: a unified engine for big data processing. *Communications of the ACM* 59(11): 56–65.