IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 2, 2020, accepted April 16, 2020, date of publication April 27, 2020, date of current version May 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2990461

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

JOSE ANGEL DIAZ-GARCIA[1], M. DOLORES RUIZ[2], AND MARIA J. MARTIN-BAUTISTA[1], (Member, IEEE)

[1]Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain
[2]Department of Statistics and OR, University of Granada, 18071 Granada, Spain

Corresponding author: Jose Angel Diaz-Garcia (jagarcia@decsai.ugr.es)

This work was supported in part by the COPKIT Project, through the European Union's Horizon 2020 Research and Innovation Programme, under Grant 786687, in part by the Spanish Ministry for Economy and Competitiveness through a project under Grant TIN2015-64776-C3-1-R, in part by the Andalusian Government, through Data Analysis in Medicine: from Medical Records to Big Data Project, under Grant P18-RT-2947, in part by the Spanish Ministry of Education, Culture, and Sport, under Grant FPU18/00150, and in part by the University of Granada, through a program of research initiation for the master's degree students.

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

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Cusano.

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]–[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.

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

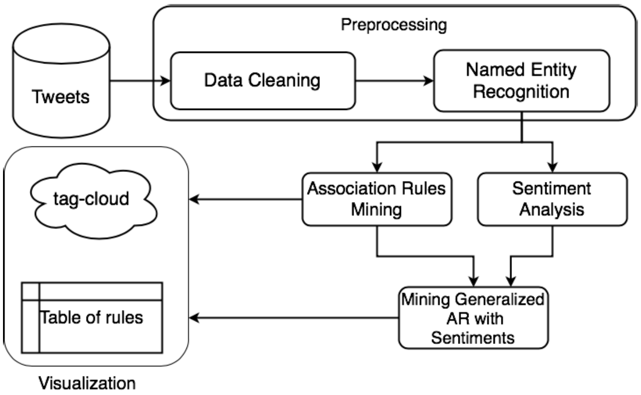

FIGURE 1. **Methodology flow.**

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.

A. 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.

B. 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

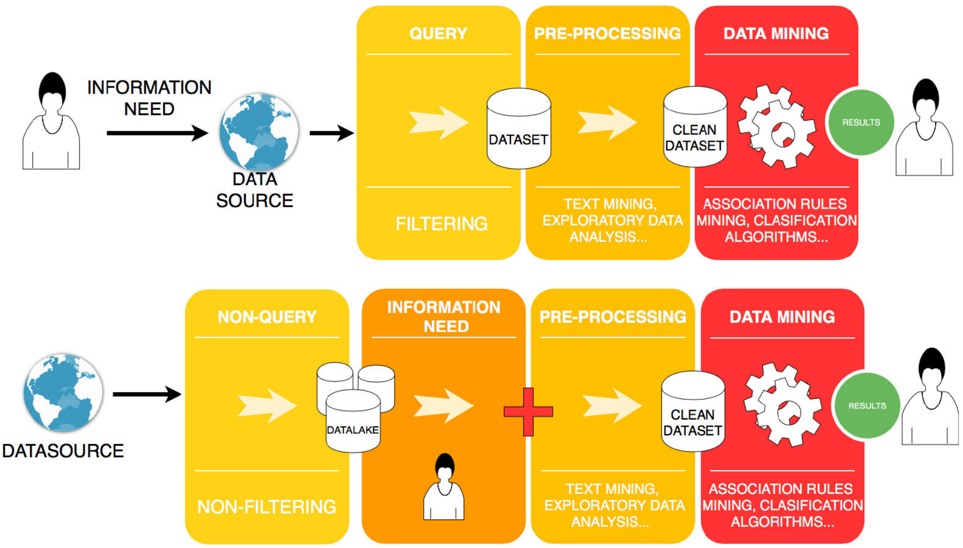

FIGURE 2. **Comparison between query and non-query based systems.**

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 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 → 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→Y)* and represents the proportion of transactions containing item X which also contains Y. The equation is:

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

- 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:

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

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 *et al.* [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.

C. 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.

A. 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]–[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.

B. 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.

C. 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 [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.

D. 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.

E. 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

[1] Internet Movie Database

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

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

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 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 preprocessing stage is necessary to normalize and clean the data. With this, we will get better results in future stages.

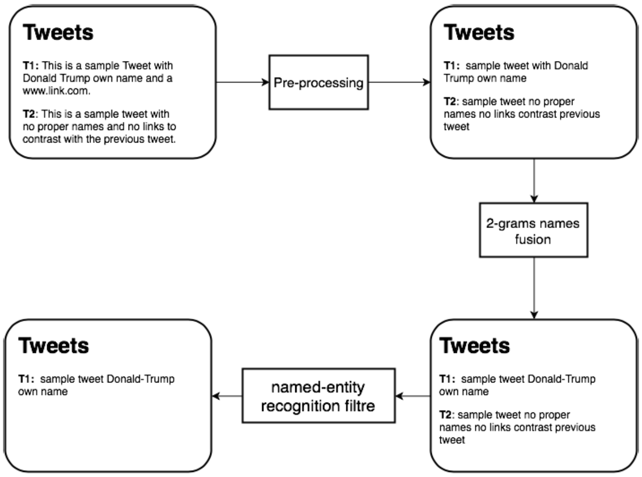
Tweets
T1: This is a sample Tweet with Donald Trump own name and a www.link.com.
T2: This is a sample tweet with no proper names and no links to contrast with the previous tweet.
Pre-processing
Tweets
T1: sample tweet with Donald Trump own name
T2: sample tweet no proper names no links contrast previous tweet
2-grams names fusion
Tweets
T1: sample tweet Donald-Trump own name
T2: sample tweet no proper names no links contrast previous tweet
named-entity recognition filtre
Tweets
T1: sample tweet Donald-Trump own name

FIGURE 3. Preprocessing and named entity recognition flow.

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 Twitter, 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*.
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 hashtags 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 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.

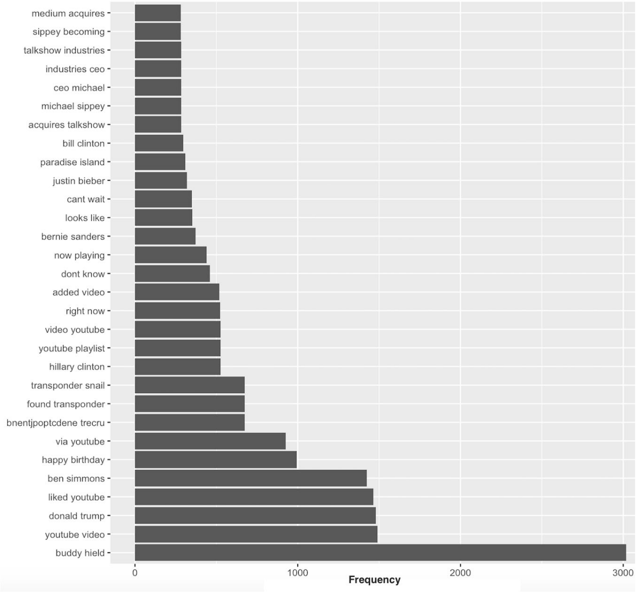

FIGURE 4. Most frequent bigrams.

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.

[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.

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.

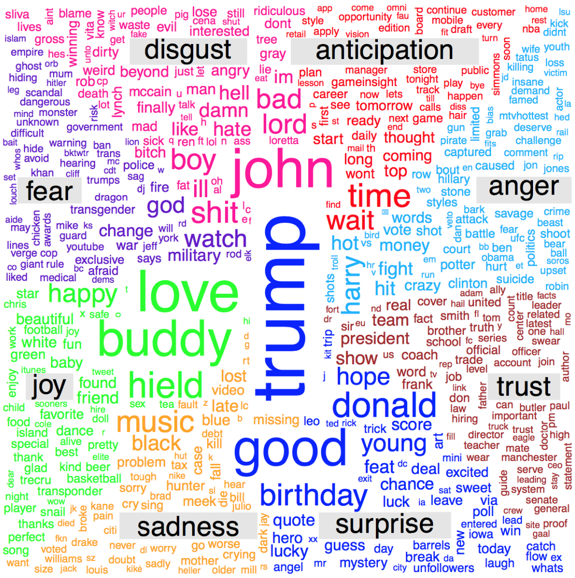

FIGURE 5. Words associated to emotions.

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.

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}→ {donald-trump}* are transformed to *{anger}→ {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.

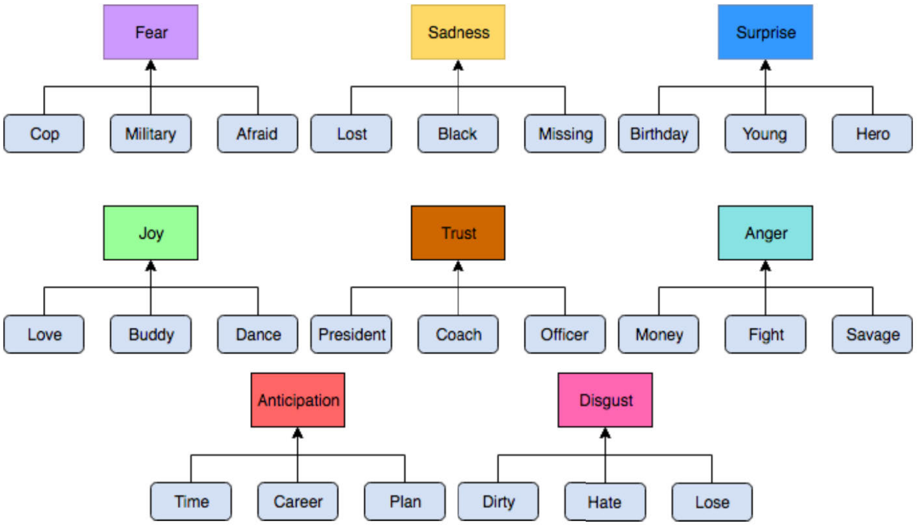

FIGURE 6. Generalization of words based on emotions.

TABLE 2. Machine specifications.

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

TABLE 3. Cluster specifications.

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

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 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 and Zhang [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.

TABLE 4. Results for the Apriori algorithm.

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

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.

TABLE 5. Results for the Eclat 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 6. Results for the FP-Growth algorithm.

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

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 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.

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 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

[3] NoSQL database used in Big Data problems.

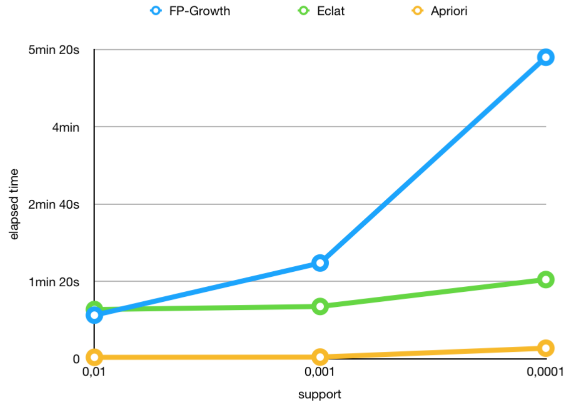

FIGURE 7. Algorithm comparison regarding the execution time.

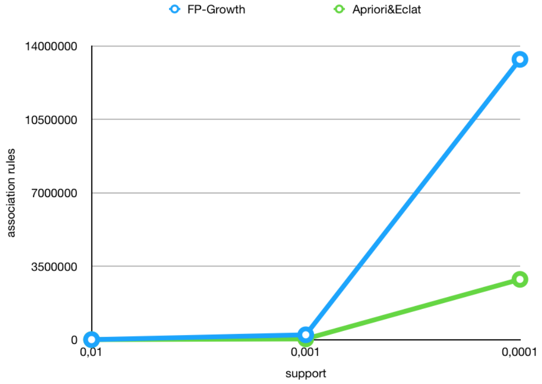

FIGURE 8. Comparison of the amount of rules discovered by the algorithms.

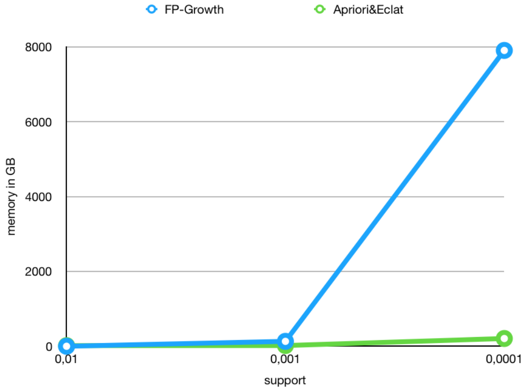

FIGURE 9. Comparison of the memory used.

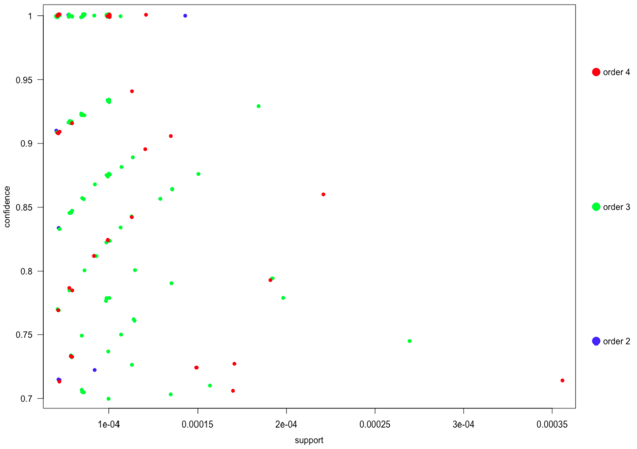

FIGURE 10. Rules distribution for Donald Trump. Support in x-axis and confidence in y-axis.

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.

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.

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

TABLE 7. Interesting rules about Donald Trump.

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

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.

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.

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.

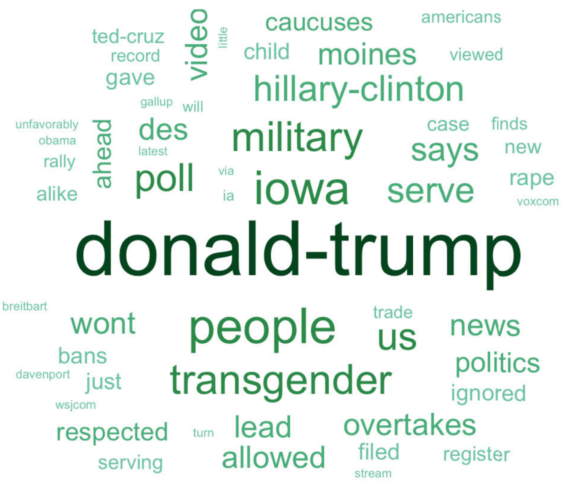

FIGURE 11. Word Cloud for Donald Trump.

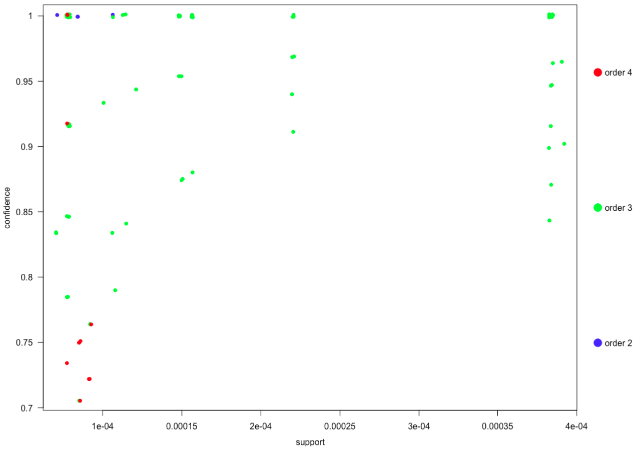

FIGURE 12. Distribution of rules for Hillary Clinton. Support in x-axis and confidence in y-axis.

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

TABLE 8. Interesting rules about Hillary Clinton.

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

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}* → *{hillary-clinton}*, *{bernie-sanders, better}* → *{hillary-clinton}* and *{bernie-sanders, race}* → *{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}* → *{hillary-clinton}*, *{attack, republicans}* → *{hillary-clinton}* and *{attack, emails}* → *{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.

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.

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.

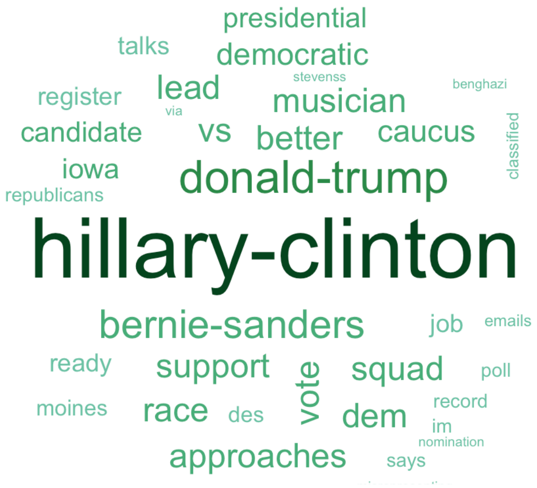

FIGURE 13. Word Cloud for Hillary Clinton.

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 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.

TABLE 9. Rules based on sentiments about Donald Trump.

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 10. Rules based on sentiments about Hillary Clinton.

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

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. CONCLUSION 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.

REFERENCES

[1] S. A. Moosavi and M. Jalali, "Community detection in online social networks using actions of users," in *Proc. Iranian Conf. Intell. Syst. (ICIS)*, Feb. 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. T. Ngai, Y. Hu, Y. H. 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," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, Feb. 2011.

[4] K. Kwon, Y. Jeon, C. Cho, J. Seo, I.-J. Chung, and H. Park, "Sentiment trend analysis in social Web environments," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 261–268.

[5] M. D. P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. Á. Rodríguez-García, and R. Valencia-García, "Sentiment analysis on tweets about diabetes: An aspect-level approach," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–9, Feb. 2017.

[6] B. Krawczyk, B. T. McInnes, and A. Cano, "Sentiment classification from multi-class imbalanced Twitter data using binarization," in *Proc. Int. Conf. Hybrid Artif. Intell. Syst.* Cham, Switzerland: 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, 2015, Art. no. e0124993.

[8] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "SenticNet: A publicly available semantic resource for opinion mining," in *Proc. AAAI Fall Symp., Commonsense Knowl.*, vol. 10, 2010, pp. 1–5.

[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," *Amer. Sci.*, 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," *Appl. Soft Comput.*, vol. 49, pp. 212–223, Dec. 2016.

[11] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993.

[12] R. Agrawal and R. Srikant, "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," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 1–12, Jun. 2000.

[14] Z. P. Ogihara, M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in *Proc. 3rd Int. Conf. Knowl. Discovery Data Mining*, 1997, pp. 283–286.

[15] R. Srikant and R. Agrawal, "Mining generalized association rules," *Future Gener. Comput. Syst.*, vol. 13, nos. 2–3, pp. 161–180, Nov. 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 *Proc. Sci. Inf. Conf.* Cham, Switzerland: 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 *Proc. 1st Int. Conf. Sustain. Technol. Comput. Intell.* Singapore: Springer, 2020, pp. 323–338.

[18] F. Colace, M. De Santo, and L. Greco, "A probabilistic approach to Tweets' sentiment classification," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 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*. Cham, Switzerland: 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 Gener. Comput. Syst.*, vol. 106, pp. 92–104, May 2020.

[21] J. Tao and X. Fang, "Toward multi-label sentiment analysis: A transfer learning based approach," *J. Big Data*, vol. 7, no. 1, pp. 1–26, Dec. 2020.

[22] A. Mohammed and R. Kora, "Deep learning approaches for arabic sentiment analysis," *Social Netw. Anal. Mining*, vol. 9, no. 1, p. 52, Dec. 2019.

[23] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 8, no. 4, p. e1253, Jul. 2018.

[24] A. Da'u, N. Salim, I. Rabiu, and A. Osman, "Recommendation system exploiting aspect-based opinion mining with deep learning method," *Inf. Sci.*, vol. 512, pp. 1279–1292, Feb. 2020.

[25] J. Chen, M. S. Hossain, and H. Zhang, "Analyzing the sentiment correlation between regular tweets and retweets," *Social Netw. Anal. Mining*, vol. 10, no. 1, p. 13, Dec. 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, Sep. 2019.

[27] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable techniques for mining causal structures," *Data Mining Knowl. Discovery*, vol. 4, nos. 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 *Proc. 1st Workshop Social Media Anal. (SOMA)*, 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 *Proc. LREc*, vol. 10, 2010, pp. 1320–1326.

[31] L. A. Daher, I. Elkabani, and R. Zantout, "Identifying influential users on Twitter: A case study from paris attacks," *Appl. Math. Inf. Sci.*, vol. 12, no. 5, pp. 1021–1032, Sep. 2018.

[32] H. T. Phan, N. T. Nguyen, and D. Hwang, "A tweet summarization method based on maximal association rules," in *Proc. Int. Conf. Comput. Collective Intell.* Cham, Switzerland: 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. 10569–10583, 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," *Int. J. Eng. Technol.*, 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 *Proc. 37th Australas. Comput. Sci. Conf.*, vol. 147. Darlinghurst, NSW, Australia: Australian Computer Society, 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 Netw. Anal. Mining*, vol. 4, no. 1, p. 212, Dec. 2014.

[37] R. C. Mosley, Jr., "Social media analytics: Data mining applied to insurance Twitter posts," in *Casualty Actuarial Society E-Forum*, vol. 2. 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 Syst. Appl.*, vol. 55, pp. 351–360, Aug. 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," *Knowl.-Based Syst.*, vol. 163, pp. 666–674, Jan. 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*. Singapore: Springer, 2019, pp. 323–334.

[41] X. Zhou, X. Tao, J. Yong, and Z. Yang, "Sentiment analysis on tweets for social events," in *Proc. IEEE 17th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, Jun. 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 Res.*, vol. 43, no. 1, pp. 51–59, Apr. 2019.

[43] Z. Hai, K. Chang, and J.-J. Kim, "Implicit feature identification via co-occurrence association rule mining," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Berlin, Germany: Springer, 2011, pp. 393–404.

[44] M. Yuan, Y. Ouyang, Z. Xiong, and H. Sheng, "Sentiment classification of Web review using association rules," in *Proc. Int. Conf. Online Communities Social Comput.* Berlin, Germany: Springer, 2013, pp. 442–450.

[45] R. Dehkharghani, H. Mercan, A. Javeed, and Y. Saygin, "Sentimental causal rule discovery from Twitter," *Expert Syst. Appl.*, vol. 41, no. 10, pp. 4950–4958, Aug. 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 *Proc. 8th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Dec. 2016, pp. 341–345.

[47] L. Bing, K. C. C. Chan, and C. Ou, "Public sentiment analysis in Twitter data for prediction of a company's stock price movements," in *Proc. IEEE 11th Int. Conf. e-Business Eng.*, Nov. 2014, pp. 232–239.

[48] M. Hahsler and R. Karpienko, "Visualizing association rules in hierarchical groups," *J. Bus. Econ.*, vol. 87, no. 3, pp. 317–335, Apr. 2017.

[49] L. Cagliero and A. Fiori, "Analyzing Twitter user behaviors and topic trends by exploiting dynamic rules," in *Behavior Computing*. London, U.K.: Springer, 2012, pp. 267–287.

[50] L. Cagliero and A. Fiori, "Discovering generalized association rules from Twitter," *Intell. Data Anal.*, vol. 17, no. 4, pp. 627–648, Jun. 2013.

[51] J. A. Diaz-Garcia, M. D. Ruiz, and M. J. Martin-Bautista, "Generalized association rules for sentiment analysis in Twitter," in *Proc. Int. Conf. Flexible Query Answering Syst.* Cham, Switzerland: 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 Evol. Hist.*, 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, Feb. 1995.

[54] X. Wang, A. McCallum, and X. Wei, "Topical N-grams: Phrase and topic discovery, with an application to information retrieval," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 697–702.

[55] K. Hornik, C. Buchta, and A. Zeileis, "Open-source machine learning: R meets Weka," *Comput. Statist.*, vol. 24, no. 2, pp. 225–232, May 2009.

[56] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2005, pp. 363–370.

[57] C. Borgelt, "Efficient implementations of Apriori and eclat," in *Proc. IEEE ICDM Workshop Frequent Itemset Mining Implement. (FIMI)*, Nov. 2003, pp. 1–9.

[58] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60.

[59] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*. Boston, MA, USA: 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, and M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, and I. Stoica, "Apache spark: A unified engine for big data processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, 2016.

JOSE ANGEL DIAZ-GARCIA received the degree in computer engineering from the University of Granada, in 2016, the master's degree in computer engineering, in 2017, and the master's degree in data science, in 2019. He is currently pursuing the Ph.D. degree in data mining. He is a Predoctoral Fellow of the Department of Computer Science and Artificial Intelligence, University of Granada. He works with the research group of Databases and Intelligent Information Systems, in collaboration with projects, such as COPKIT, in the topics of text mining, big data, and social media mining.

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

She held non-permanent teaching positions at the Universities of Jaen, Granada, and Cadiz. She is currently with the Department of Statistics and OR, University of Granada. She has organized several special sessions about data mining in international conferences and was a part of the organization committees of the FQAS 2013 and SUM 2017 conferences. She belongs to the Approximate Reasoning and AI Research Group and the Cybersecurity Lab, 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. She has participated in more than ten 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.

MARIA J. MARTIN-BAUTISTA (Member, IEEE) is a Full Professor with the Department of Computer Science and Artificial Intelligence, University of Granada, Spain. She is a member of the Intelligent Data Bases and Information Systems IDBIS 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. theses and published more than 100 articles in high-impact international journals and conferences. She has participated in more than 20 Research and Development projects, including several European projects, and has supervised several research technology transfers with companies. She has served as a program committee member of several international conferences.