IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 8, 2019, accepted April 24, 2019, date of publication May 16, 2019, date of current version May 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917398

Indicator Proposal for Measuring Regional Political Support for the Electoral Process on Twitter: The Case of Spain's 2015 and 2016 General Elections

JOSE N. FRANCO-RIQUELME[1], **ANTONIO BELLO-GARCIA**[2], **AND JOAQUÍN ORDIERES-MERÉ**[1]

[1]Department of Industrial Engineering, Business Administration and Statistics, Universidad Politécnica de Madrid, 28006 Madrid, Spain
[2]Department of Construction and Manufacturing Engineering, Universidad de Oviedo, 33203 Gijon, Spain

Corresponding author: Jose N. Franco-Riquelme (j.franco.riquelme@upm.es)

This work was supported by the European Commision (Erasmus Mundus Sustain-T Project).

ABSTRACT Research on electoral events in conjunction with social media provides opportunities to describe an interesting phenomenon that can be analyzed using sentiment analysis techniques. The goal of the study is to analyze the support of political parties during electoral periods from Twitter comments, including 250 000 tweets regarding the Spanish general elections of 2015 and 2016, respectively. Text mining and natural language processing techniques enable information analysis, and the methodology emphasizes good practices for large-scale data collection retrieved from Twitter through a quantitative analysis of text collection written in the Spanish language. After information extraction obtained in three Spanish regions defined by geolocation, as well as feature selection based on keywords of the main four political parties, we conducted an in-depth examination of Twitter users' support during the course of the election. By weighting the tendency of tweets, we were able to obtain a proposed indicator of support: the positiveness ratio (PR). The results suggest that PR is a feasible barometer to demonstrate the measurable patterns of support tendency regarding political parties and users' behavioral activity to track their affinity on Twitter. The findings indicate consistent support behavior by users toward traditional parties and optimistic users' behavior regarding emerging political parties.

INDEX TERMS Geolocation, positiveness ratio, sentiment analysis, natural language processing, Spanish general election, text mining, Twitter.

I. INTRODUCTION

In the last decade, social researchers have witnessed significant political changes [1], [2]. Social media has transformed communication in politics: candidates, political parties, and society have created a major platform on which all users can share their opinions and comments on political causes. With the advent of Web 2.0, a plethora of discussions on online network channels has provided researchers with a primary source for analysis on the role of social media in politics.

Events such as elections generate a vast amount of data that can be processed using sentiment analysis techniques, which are also known as "opinion mining." This can be defined as the computational treatment of opinions, sentiments, and subjectivity in the text, and it has numerous applications in political science [3], [4]. Through social media, researchers can analyze trends, evaluate public opinion, gauge reactions and appraise voters.

Based on this analysis, online forums offer valuable opportunities for political debate to occur because they transcend geographical boundaries and provide easy access to low-cost discussion platforms. Consequently, new media has become increasingly important during election campaigns [5], [6]. Additionally, users tend to have a predisposition for shorter text to demonstrate or relate their opinion, as opposed to writing longer excerpts or documents [7], and this is one of the main reasons to choose Twitter as a research platform.

Several studies have addressed the relationship between politics and social media platforms [8]–[11] and seek to describe this relationship through communication and a focus

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

on the Twittersphere. This involves users of the social media application Twitter being considered collectively as analysis subjects.

A limited number of researchers have determined the indicators of voters' behavior regarding political parties—for example, philias and phobias, and other emotions with respect to voter preferences. A question emerges in this situation: Can the degree of support for political parties be efficiently measured, for instance, during electoral periods? To answer this question, we examined the work of previous researchers who developed indicators that were implemented for elections, such as the relative support (RS) parameter. This indicator yielded evidence of a connection between Twitter users' activity and election outcomes [12], [13].

In this research, we undertook an in-depth examination of Twitter users' support by weighting the tendency of tweets based on our proposed indicator: the positiveness ratio (PR). Using geolocation, we were able to obtain a basis upon which to compare different sentiments by region in Spain; in this case, it was crucial to compare the dissimilarities of political region preferences. Similarly, we considered it essential to have an estimate of Twitter users' preferences based on positive sentiments (philia) about political parties in the context of an election process.

The aim of this paper was to develop an indicator based on positive mentions on Twitter (per period and region) to address the correspondence of support with a view to describing whether variations exist based on election outcomes, using text mining and geolocation. In addition, we contribute to the best practices in natural language processing techniques of the Spanish language.

Our study focused on the Spanish general elections of 2015 and 2016 and analyzed user activity during this period. We had the unique opportunity to study the same election (because a re-election of the first election was required) on two different dates: December 20, 2015, and June 26, 2016. Analyzing these events enabled us to obtain remarkable insights by comparing our data results through the aid of our proposed indicator.

There are reasons to justify the selection of Spain as a case study for estimating the representation of data retrieved from social media analysis within the political sciences field. In fact, choosing Spain is based primarily on guidelines regarding accessibility to traditional media and political party presence in the media, in addition to the high penetration of the Internet in the population. Moreover, according to previous studies [14], [15], the main political parties have received more coverage than newer parties based on previous electoral results. In addition, there is a lack of state regulations for Twitter campaigning, and the emergence of Spanish popular movements such as 15-M—the Indignados—and other new political parties [16], which have strategically implemented digital channels as a new way of spreading information, represented an ideal open scenario for online social media analysis.

Once we understood the current sentiment analysis, we were able to observe that existing English literature discussing such analysis is extensive and has been thoroughly discussed. We realized that the non-English-language research (in this case, Spanish) in this field is quite scarce, however, and does not provide the same quantity of available information as English language literature. Consequently, we emphasized new methods employed in Spanish-language research in the natural language processing (NLP) field and improved practices using text mining techniques. To facilitate our analysis, we used LinguaKit [17], which is a multilingual open-source toolkit that performs NLP tasks to accomplish our sentiment analysis proficiently. Our research illustrated the role of social media in politics and new ways to measure events related to social media. We performed an enhanced analysis based on Twitter data, which may be applicable to other fields, such as marketing studies. Hence, trends related to products and services, behavioral patterns regarding publicity campaigns, and their effects on customers according to the knowledge generated from the social platforms can be measured.

The remainder of the study is organized as follows: the background is described in Section II in two main topics, social media and politics, and the studies related to Twitter, that measure emotions in elections. In Section III, the data preparation section describes our Twitter database retrieval, our research approach to text mining and NLP, the tweet geolocation, and the feature selection. Section IV comprises the methodology for developing our indicator based on our cleaned data collection. In Section V, we present the results related to the Spanish general elections of 2015 and 2016. The discussion of our results based on the indicator proposed is analyzed in Section VI. Conclusions and future research directions are summarized in Section VII.

II. BACKGROUND

A. SOCIAL MEDIA AND POLITICS

The interactions between politics and social media have generated significant attention among researchers since their joint activities emerged in the first decade of the XXI century. These activities influence society and the population of a country or region. For decades, politicians and political parties have been accustomed to spreading their messages via newspapers, radio, and television. In this sense, modern online sociotechnological systems are catalyzing profound changes in the manners in which we communicate [8], and various social activities are now organized via social media platforms.

This research is focused on Twitter, which is an extended microblogging platform whose users share short messages called "tweets,[1]" as well as links to other websites, videos,

[1] On September 26, 2017, Twitter increased the text limit of a tweet from the original 140 characters to 280: http://www.telegraph.co.uk/technology/2017/09/26/twitter-tests-doubling-length-tweets-280-characters/

and pictures [18], [19] and generally express their likes and dislikes. Moreover, users look for information, such as breaking news, data about celebrities and influencers, and comments on different topics, including politics, business, and economics, which they also discuss. As previously mentioned, Twitter has been considered a valuable tool and an important source for political analysis [20], providing a ''convenient source of data on users' opinions, interactions, and reported behaviors'' [21]. In fact, it has become an important tool for political actors in political campaigns [22].

It is crucial to note the limitations of social media analysis, which cannot replace traditional polls [23] because a relatively nonrepresentative portion of the population uses this new communication tool. Indeed, the difficulties related to language interpretation—for instance, the pitfalls associated with sarcasm and ambiguity, which are inherent in political debates—must be considered.

Precedents in the literature have explained behavioral patterns among Twitter users regarding elections in different countries. During the British and Dutch general elections in 2010 [24], Dutch politicians were more active than their British counterparts. It was observed by Ahmed *et al.* [10] that prior to the 2014 Indian elections, Twitter was used primarily to push timely, on-demand information to followers regarding campaign updates and political party promotions.

In the 2014 European Parliament (EP), the content of social media communications [25] have been examined from the perspective of the use and adoption of Twitter and based on evaluations of the tone and variations of campaigns, using text analysis tools such as linguistic inquiry and word count (LIWC), to analyze the emotional tone and polarity of tweets.

1) THE OPPOSING APPROACHES ON POLITICAL DATA ANALYSIS

There are two opinion trends on social media and political data analysis, and they can be classified into two main categories with opposing approaches to this issue. The first comprises studies that claim to have predictive capabilities; it is argued that social networks are even more accurate forecasters of election outcomes than the traditional media [26]–[31]. The authors have achieved a certain degree of success using different machine-learning and lexicon-based prediction views on Twitter. They performed techniques such as ordinary least squares regression (OLS) models, root mean square error (RMSE) and correlation, the previously mentioned lexicon approach, and the rule-based sentiment analysis, and polarity extraction.

Caldarelli *et al.* [13] also addressed this research proposition, performing a multiscale analysis of the Twitter time evolution regarding the 2013 Italian elections. They monitored the behavior of Italian Twitter users, claiming that the numerical indicator—the ratio support (RS) parameter— [12] implemented in their data seemed rather accurate regarding predicted votes. Nevertheless, they could find no conclusive evidence of its predictive capability regarding the trends of the final electoral results.

Furthermore, leading up to the 2015 general elections in the United Kingdom (UK), Burnap *et al.* [32] studied Twitter data in an effort to forecast the outcome of the election results; they searched for users' feelings and associated the polarity scores of tweets regarding each political party's supporters. For this purpose, automated sentiment analysis was incorporated using software developed by Thelwall *et al.* [33], who organized the text into scales of positive and negative tweets regarding nine political parties in the UK. Their results indicated that the Labor party gained the majority of seats, which proved inaccurate regarding the final electoral outcomes, although they correctly predicted the order of the top three parties regarding vote share: Conservative, Labor and UKIP.

In contrast, the authors who were critical of Twitter's predictive capabilities with regard to political data analysis found no correlation between the analysis performed on this social media platform and the election outcomes, nor did they determine it to be an accurate predictor of electoral success [34]–[38]; rather, they argued that this was a simplistic analysis. Therefore, critics argue that it is imperative for social network users' representativeness to be considered when considering the age groups of the majority of these users, most of whom are between 16 and 29 years old [21], [39], [40].

Correspondingly, in a study performed in the context of the 2012 United States Republican presidential primary elections—in which an approach based on lexicon, using the lexicoder sentiment dictionary (LSD), was adopted—it was acknowledged that ''tweets are reactive rather than predictive'' in regard to the political context [41]. After conducting an exhaustive literature review, Gayo-Avello *et al.* [39] affirmed that electoral predictions based on Twitter data cannot replace traditional polls. They enumerated the main weaknesses of research predictions based on Twitter and made recommendations for further research in this area.

Moreover, a study carried out by Jungherr *et al.* [38] caused controversy due to its strong refutation of the possibility that Twitter could be used for predicting the results of the 2009 German elections based on its users' activities. The authors stated that ''the number of party mentions in the Twittersphere is thus not a valid indicator of offline political sentiment or even of future election outcomes''; this statement was in response to the study conducted by Tumasjan *et al.* [26], in which their predictions approach was assumed to be correct regarding the 2009 elections. Responding to the question posed by Jungherr *et al.*, they restated their position, arguing that their conclusions were ''well supported by both data and analysis'' [42].

The prediction approach based on Twitter data might be better addressed with the help of polling information, such as that obtained from traditionally conducted public opinion surveys, to improve forecasts [27] if there is a significant statistical relationship between social networks and voting results.

The challenges that arose were due to the difficulty of conducting analysis in the Spanish language, as well as the cleaning and processing steps. During the analysis, retweets were not excluded, nor were tweets mentioning multiple parties, among other factors. Altogether, this experience expressed the importance of carefully processing data while implementing the methodology selected for achieving the study goals.

In brief, considering both sides, efforts should be made to seek a third method, to establish trends and obtain useful information using appropriate techniques to improve the measurement methods used on social media regarding the political context. According to Gayo-Avello [36], not everyone uses Twitter, and only a minority of users tweet about politics. In addition, researchers in social media analysis must be aware of fake news and data and should avoid simplistic analyses of opinion mining.

2) THE SPANISH POLITICAL SCENARIO AND TWITTER ANALYSIS

Since the emergence of online political activism represented by actions such as 15-M and organizations such as Podemos in Spain, major innovations have been introduced to the field of political communications, which has a preponderant role in a modern democracy [16]. Hence, it is worth mentioning some research on the interrelationship between social and political media in electoral periods in the context of Spain.

Borondo *et al.* [12] suggested that there was "a strong correlation between the activity taking place in Twitter and election results" regarding the 2011 Spanish elections. Developing this issue, they proposed an indicator—the RS parameter—according to the cumulative volume of mentions between two parties; the resulting quantity might be a forecaster of the success of the most frequently mentioned party. They stated, however, that the approach could not ensure the predictability of a political campaign outcome based on Twitter.

In particular, voting behavior has been analyzed using mathematical modeling with sociological, psychological, demographic, and economic variables based on a finite time-discrete compartmental dynamic [43] over the course of the Spanish election process between 2011 and 2015. It predicted the performance of the election winner, the PP, as well as the entrance of the new political players, Ciudadanos and Podemos, which resulted in a four-party political scenario [44].

Singh *et al.* [45] used a sentiment analysis approach focused on the 2016 Spanish election with the help of the open-source web framework ASP.Net. They calculated the positive sentiment score of each party to predict and compare the electoral outcomes. The lack of accuracy was evident, however, as the results of only two political parties were predicted.

In addition, through a sample of 9,042 tweets of quantitative content analysis, and digging into the communication strategies of political actors, Lopez-Meri *et al.* [46] found hybridization between new and old media usage of Twitter by politicians, focusing them on innovative digital platforms for establishing links with the mainstream media. In the same way, Alonso-Muñoz and Casero-Ripollés [22] contributed to understanding the political interactions of the 2016 election in terms of the self-communication of parties and leaders, and the impact of citizens' activities on Twitter. They noticed different forms of political programs, a low degree of thematic fragmentation, and a degree of dissonance in the agenda between politicians and the interest of Twitter users.

B. MEASURING EMOTIONS ON TWITTER

Research carried out in the 1940s and 1950s by Alan Turing, among others, established a new field in the sciences—artificial intelligence (AI)—and advances in computer calculation [47]. Consequently, this research was acknowledged for scholars in this area as fundamental to the development of disciplines such as the one that is relevant to this investigation—NLP—which is also called computational linguistics [48]. In fact, NLP is recognized as a subfield of computer science that comprises learning, understanding, and human language production content, and it has helped with the scrutiny of expansive collections of texts from numerous sources.

Among the aspirational goals of research in NLP is to examine and reveal the content of a vast amount of data to find underlying information in order to understand trends and sentiments of people, in this case, Twitter users. One should consider the fact that NLP is not exempt from difficulties. Seeking accuracy and efficient measurements is essential to resolve problems related to social media text mining. A great deal of literature exists on this topic [49], since the ability to rely on measuring techniques is critical. Moreover, an NLP system must be accurate to avoid problems regarding the disambiguation of words, sentences, syntactic structures, and semantic scope [50]. Sarcasm can also be a pitfall in analysis that is used specifically to correctly categorize opinions about products, services, and political candidates, among other areas [51]. Considerable progress in computational linguistics has been made in the last decade due to increased computing power, however, which allows highly mathematically complex models and available large text corpora to be processed.

The possibility of exploring and tracking the preferences of people on Twitter was an opportunity for us to use this information in a reliable manner. Hence, there are research approaches that are based on different methods of revealing nonexplicit information about predilections in the field of politics. Consequently, Ceron *et al.* [52] endeavored to measure political preferences in Italy and France using supervised sentiment analysis on blogs and social networks, according to the method outlined by Hopkins and King [53], which provided useful information for enriching traditional offline polls.

Analyzing dynamics regarding Twitter message volume in relation to political parties [9], a statistical pattern was identified using geometric Brownian motion. Hence, a positive autocorrelation over a short time (i.e., daily tweet volume), and a broad distribution such as log-normal in political parties was found. In addition, the optimal averaging tendency of tweet volume was identified, suggesting a limited capacity of prediction, despite the strong fluctuation of Twitter activity.

This microblogging platform was also used to classify supporters and antagonists of organizations, such as the Islamic State of Iraq and Syria (ISIS). This was viewed as research into online polarization [54] and as a way to measure engagement patterns to illustrate underlying phenomena in organizations.

It is worth mentioning the FreeLing tool [55], which is an open-source language NLP implementation that is used extensively in some languages (including English, Spanish, Portuguese, and Italian) to conduct text analysis. It has been implemented in several studies ....([56]–[58] in Spanish language. This tool has been used to perform NLP to support primary analytical tasks and data processing, with certain success. For the objectives of this analysis, however, we did not utilize the FreeLing tool to obtain Twitter commentary polarities.

We found that once examined, of all the resources that were necessary for working with our Spanish language database and were considered in the implementation of this research, LinguaKit[2] was the most useful computational tool. The system is programmed in Perl, which was developed in a Spanish University [59] and is defined as a suite for analysis, extraction, annotation, and grammatical correction. It also enables the performance of different tasks, including opinion mining, which provides language detection in four languages: Spanish, Galician, Portuguese, and English [17]. For the first three mentioned languages, the suitability of this software stands out for the inclusion of a verbal conjugator as an independent module, significantly improving the analysis in these languages.

To summarize, LinguaKit proved to be beneficial not only as an NLP application used in the early stage of the text process but also for obtaining reliable results in the Spanish language using sentiment analysis [60]. The accurate detection of entities (i.e., organizations, names, and references) and a practical method for finding text polarities are among the most important means to achieve the proposed objectives using the NLP process, which will be described in the following section.

III. DATA PREPARATION

From our point of view, it is necessary to measure the degrees of positiveness on Twitter to compare these with the results of the elections. We conducted our quantitative analysis in the context of the 2015 and 2016 Spanish general elections, using the Twitter database, and set the geolocation to obtain tweets according to region.

We adopted text mining and NLP techniques to retrieve and preprocess the data, and we carried out sentiment analysis with the help of the LinguaKit computational tool to perform feature selection and polarity of the tweets. In addition, we developed an indicator to compare the election outcomes as a proxy of the tendency in the electoral events. Fig. 1 displays the framework developed in this study.

A. RETRIEVAL PHASE

For our data collection, we focused on tweets posted about the Spanish general elections held on December 20, 2015 and June 26, 2016. The dataset is provided via Zenodo (https://zenodo.org/record/2662543#.XNAoqugzZD9) and can be processed using R and the libraries provided in the scripts.

For this research, it was fundamental to identify the evidence in a vast amount of data concerning the elections. In the case of social media platforms, hashtags (#) provided the keywords for classifying the messages related to events and topics. Accordingly, we followed the method by which the media and its users broadly identified the conversation regarding these elections—that is, the hashtags **#20D** and **#26J**, taken together with the keywords assigned to each event (see Fig. 2). According to the trending topic in Spain, much of the conversation about the election revolved around these hashtags.

Data were collected from a sample of 250,000 public tweets in accordance with Twitter policy (section F.2.b.i), as the motivation is research with noncommercial interest, in accordance with its licensing.[3] The process of how the database was selected and recorded is depicted in Fig. 3.

We retrieved posts by searching on the Twitter page and saving the searches using the Python computational language. The library Twitter Scraper[4] was used to extract the messages in JSON[5]-based (raw) format and based upon the command lines containing references as the mentioned hashtags and the election dates, we built our tweets database.

After the data retrieval stage, we built our Twitter database centering on eight attributes: tweet creation date, the text of the tweets, favorites (likes on other social networks), retweets (which is the same as reposted tweets), the path of each tweet (i.e., the URL address), tweet ID (which is a unique number used to identify a tweet), user ID (in this case, the unique number used for user identification), and username.

All attributes gave us precise information about each tweet, addressing the base for the analysis. The approach that we used to undertake this study, however, focused on five attributes: 1) tweet creation date, which we used to divide our database into four separate periods; 2) text, which provides the primary base for dealing with polarity and entity recognition (e.g., PSOE, Rivera, Podemos); 3) favorites and retweets, which provide the method for creating the weighting to be

[2] See https://linguakit.com/en/about

[3] See https://developer.twitter.com/en/developer-terms/policy.html

[4] See https://github.com/taspinar/twitterscraper/

[5] JavaScript Object Notation (JSON)

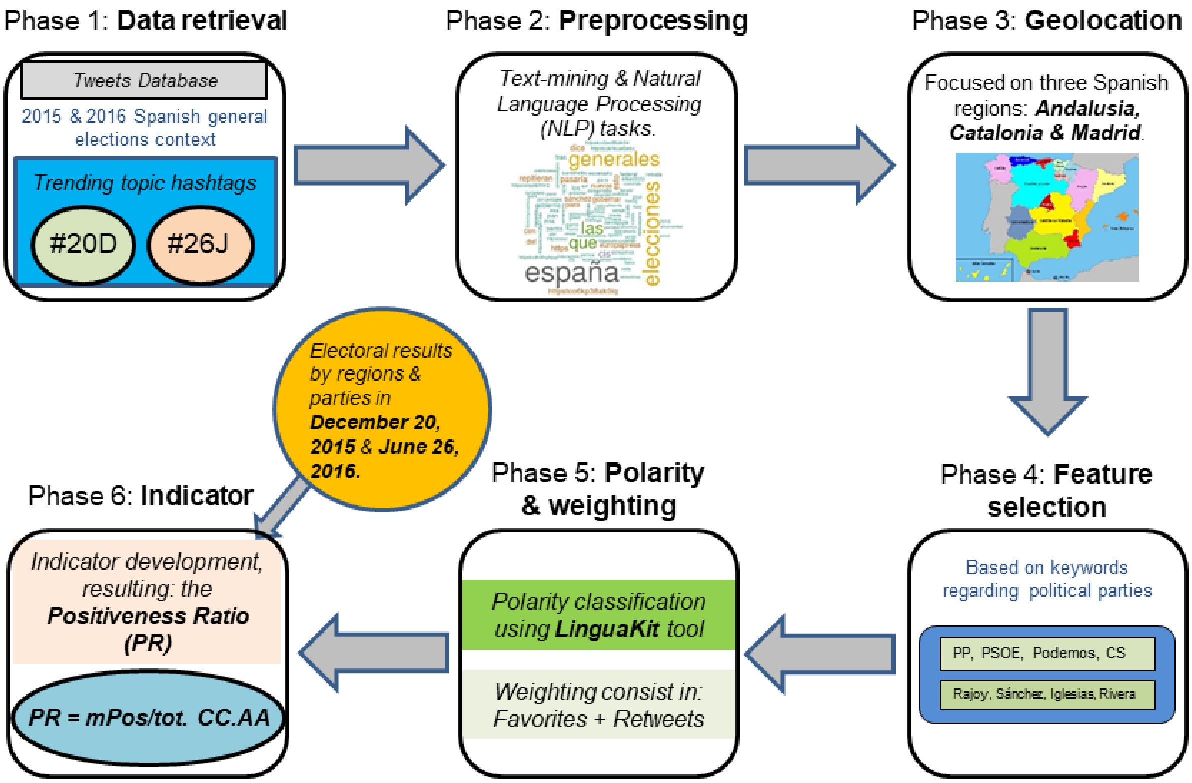
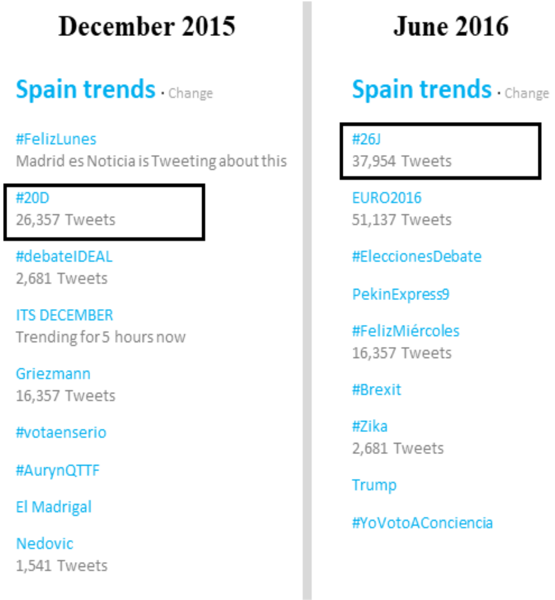

Phase 1: Data retrieval
Tweets Database
2015 & 2016 Spanish general elections context
Trending topic hashtags
#20D
#26J
Phase 2: Preprocessing
Text-mining & Natural Language Processing (NLP) tasks.
generales
las
que
españa
elecciones
Phase 3: Geolocation
Focused on three Spanish regions: Andalusia, Catalonia & Madrid.
Phase 4: Feature selection
Based on keywords regarding political parties
PP, PSOE, Podemos, CS
Rajoy, Sánchez, Iglesias, Rivera
Phase 5: Polarity & weighting
Polarity classification using LinguaKit tool
Weighting consist in: Favorites + Retweets
Electoral results by regions & parties in December 20, 2015 & June 26, 2016.
Phase 6: Indicator
Indicator development, resulting: the Positiveness Ratio (PR)
PR = mPos/tot. CC.AA

FIGURE 1. The framework of the approach proposed in this research.



December 2015

Spain trends · Change

#FelizLunes
Madrid es Noticia is Tweeting about this

#20D
26,357 Tweets

#debateIDEAL
2,681 Tweets

ITS DECEMBER
Trending for 5 hours now

Griezmann
16,357 Tweets

#votaenserio

#AurynQTTF

El Madrigal

Nedovic
1,541 Tweets

June 2016

Spain trends · Change

#26J
37,954 Tweets

EURO2016
51,137 Tweets

#EleccionesDebate

PekinExpress9

#FelizMiércoles
16,357 Tweets

#Brexit

#Zika
2,681 Tweets

Trump

#YoVotoAConciencia

FIGURE 2. Trending Twitter topics in Spain during the months of December 2015 and June 2016.

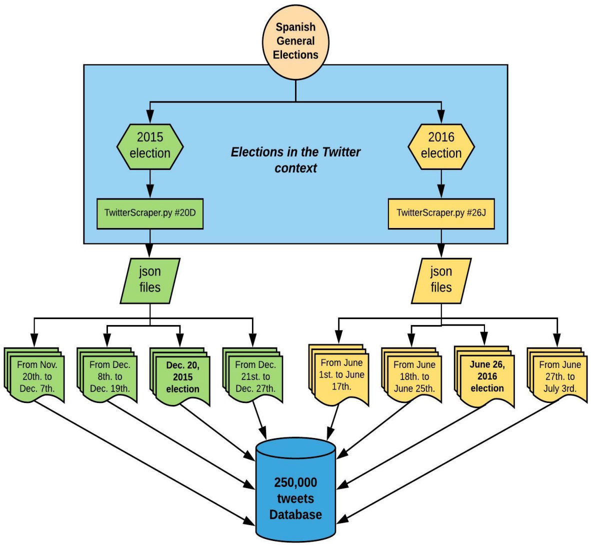

Spanish General Elections
2015 election
Elections in the Twitter context
2016 election
TwitterScraper.py #20D
TwitterScraper.py #26J
json files
json files
From Nov. 20th. to Dec. 7th.
From Dec. 8th. to Dec. 19th.
Dec. 20, 2015 election
From Dec. 21st. to Dec. 27th.
From June 1st. to June 17th.
From June 18th. to June 25th.
June 26, 2016 election
From June 27th. to July 3rd.
250,000 tweets Database

FIGURE 3. Tweet data acquisition and processing depiction.

used in our indicator proposed in this work; and 4) username, which is crucial for obtaining the geolocations of the Twitter users.

B. THE PREPROCESSING PHASE

To perform text mining, we first chose tweet dates to define the periods under study. Second, we focused on the text, extracting all the information for our analysis based on the tweets. Third, tweet metadata, such as favorites and retweets, have been used to develop our weighting indicator, which will be examined in detail later.

We divided the dataset into two stages based on two electoral events: the first election date—December 20, 2015—and the second election date—June 26, 2016. Regarding the first electoral event, we covered the pre-electoral period between November 20 and December 19, 2015. This period was divided into two sections, which were 30 and 12 days before election day—that is, the day of the general elections on December 20, 2015—and the postelectoral period between the day after elections (December 21) and December 27, 2015.

Likewise, for the second electoral event, we established the pre-electoral period as June 1 to June 25, 2016, which we divided into two sections: 30 days and 12 days before election day—that is, the day of the repeated general elections (June 26, 2016)—and the post electoral period between June 27 and July 3, 2016.

Once we defined the periods, we carried out text cleaning from the database of tweets to establish the corpora, which can be defined as a collection of texts that is representative of a particular natural language or language variety [61]. The set of tweets plays an essential role in providing the material basis and a test bed for building NLP systems.

We structured the preprocessing steps based on the work of Dubiau and Ale [62] and Jurafsky and Martin [49], where we performed information retrieval such as regular expressions (RE) to identify patterns, as well as normalization, tokenization, filtering, and other NLP tasks.

C. GEOLOCATION

Considering the aim of obtaining tweets by region, an essential task was to obtain the geolocation of Twitter users. One of our aims in this study is to achieve certain levels of granularity regarding each municipality, province and autonomous community in Spain. This perspective enriched the information, giving us the view of where it was generated and thus providing an automated analytical view of where the event of interest occurred. Most of the tweets, however, did not indicate their location. Less than 1% of tweets are geolocated; therefore, it is difficult to obtain information from users' profiles [63].

Given this problem, we developed a technique for determining the user's location, and in performing this task, one of the variables helped us to identify the location of the user: the user ID. We used the R library twitteR[6] to process data extraction through the Twitter API, merging the extracted tweets with the data retrieved from each user profile. The location of the user provides us with the geolocation through the information obtained from his or her profile.

The database was reduced because few users include their location in the personal information section of their profiles. Nevertheless, we obtained 40% of the geolocated tweets from the entire dataset. Having extracted the users with locations in their profiles and compared them, we merged the tweets and the locations. We obtained the locations from the National Institute of Statistics of Spain, which provided precise and reliable information regarding political divisions based on more than 8,000 municipalities, 50 provinces, and 17 autonomous communities.

In this work, we are based on Spain's population representativeness, which are the three autonomous communities with the largest population and therefore, the greatest tweet generators (as can be seen in the web page: https://es.statista.com/estadisticas/472413/poblacion-de-espana-por-comunidad-autonoma/we). For this reason, we focused on the three autonomous Spanish communities—Andalusia, Catalonia, and Madrid. Thus, based on this premise, our database corresponds to the population of the selected autonomous communities (henceforward CC.AA), from which the majority of the tweets were retrieved.

D. FEATURE SELECTION

The process of taking unstructured information embedded in a collection of text after preprocessing and converting it into structured data is called information extraction (IE). Furthermore, feature selection, which entails classifying a collection of documents, is considered a critical stage of our analysis [49]. In our case, we analyzed a collection of tweets, having built our database on two key terms associated with the elections—#20D and #26J. Therefore, period definition, text cleaning, geolocation, and tweet classification were the beginning of our analytical construction.

During this phase, with the help of LinguaKit, we undertook named entity recognition (NER), which is also known as entity chunking or entity extraction. This was applied based on a set of rules for proper names used for entity identification, and elements were labeled and placed into categories such as persons, organizations, and locations [64], [65]. As a result, we classified our data based on keywords related to the four political parties on which we focused, as shown in Table 1.

It is worth mentioning that there are names referring to each political party that was researched for this study; consequently, the parties themselves might be referenced by naming their leaders: "Rajoy," "Soraya," "Sanchez," "Iglesias," and "Rivera"; regionalisms: "Populars" or "Ciutadans"; places: "Genova"; and references, such as "Naranja."

IV. INDICATOR APPROACH

This study established a criterion that helps to explain the evolution of the political parties involved in this electoral process. To analyze the relationship between the results of two electoral events, we compared the official results with those obtained from Twitter. We aim to identify the sentiment behind the tweets located in different regions of Spain.

Among the analyzed parties we find two of long political tradition in Spain (Popular Party, or PP, and the Socialist Workers' Party, or PSOE) against two other emerging parties

[6] See https://cran.r-project.org/web/packages/twitteR/twitteR.pdf

TABLE 1. Implementation of named entity recognition for a dataset of keywords related to the four political parties selected for this research.

Political Party	Keywords samples
PP	“Partido Popular”, “PP”, “Rajoy”, “Soraya”, “Mariano”, “Ppopular”, “Genova”, “Cifuentes”, “Albiol”, “Populars”
PSOE	“PSOE”, “Pedro_Sanchez”, “Susana_Diaz”, “JoveSocialistesYa”, “Carme_Chacon”, “PSC”, “Zapatero”, “Josep_Borrell”
Podemos	“Partido_Podemos”, “Pablo_Iglesias”, “Errejon”, “Podemos”, “AdaColau”, “EnComuPodem”,“JMKichi”, “XavierDomenech”
CS	“AlbertRivera”, “Cs”,“Ines_Arrimadas”, “Ciudadanos”, “Ciutadans”, “Rivera”, “Naranja”, “Riveranosotros”

of recent creation (Citizens, or CS, and We Can, or Podemos) participating for the first time in elections at the national level. Subsequently, NER was defined based on the number of tweets mentioning these political parties, leaders, and the number of references made about them (see Table 1).

Taking the dataset into consideration, we accomplished polarity recognition using LinguaKit to establish the three categories for this phase: positive, neutral, and negative tweets. For instance, we recognized each tweet based on polarity, and a numerical value was assigned to each tweet as a percentage indicator of its positive (1), negative (-1), or neutral value (0).

Thereafter, we identified and linked entities and polarities before proceeding to sum the occurrences of positive and negative tweets in the same way that each tweet was weighted using Twitter metadata, such as favorites and retweets. Altogether, our collected data comprised the summation of positive and negative tweets, as well as their weighting, divided by CC.AA and time periods. In this study, we focused on collecting only the positive results of occurrences and weighting its values.

A. THE POSITIVENESS RATIO (PR) PROPOSAL

The ratio is a mathematical expression that is used to compare quantities; it indicates the relationship between two numbers, as well as how many times the first number contains the second, and the result is expressed in the form of a decimal fraction. Using this definition, previous researchers have sought to determine the relationship between Twitter users' activity and election outcomes [12], [13], [66]. Borondo *et al.* developed an indicator—the relative support (RS) parameter—for studying political sentiment, and it was “used to indicate and quantify which candidate and in which proportion is getting more benefits from events occurring offline.” This approach was applied to the 2011 Spanish general elections and, likewise, the 2013 Italian elections, for which evidence of correlation was found between Twitter users' activity and electoral outcomes.

This measure consists of the ratio between the aggregate allusions to the two political parties as an indicator of their activities, performed within the RS parameter, which is defined as RS A/B, according to the following expression:

$$\mathrm{RS\ A/B} = \mathrm{mA/mB} \tag{1}$$

where mA and mB are the slopes for the accumulated mentions of the A and B political parties, obtaining a decimal number that is considered the ratio of the A party.

It is important to note, however, that this indicator might have some limitations, considering the tools that we utilized, along with the literature review and the methods presented in this section. Hence, omitting the fact that there are positive and negative mentions in tweets, differentiating users who were supporters from those who were not was extremely challenging. The metadata weighting was not taken into account, nor was the possibility of determining users' geolocations to establish their behaviors by region.

Inspired by the work of Borondo *et al.* [12], [66] and Caldarelli *et al.* [13], we established our primary goal, which was to find an effective way to estimate the support for political parties. Thus, we focused on the positive tweets and their weighting. In this sense, we proposed a new indicator to measure the support for each political party based on the weighting of the positive amounts divided by the sum of the positive accumulation of tweets (mPos) per autonomous community as the denominator (tot. CC.AA). By doing so, we achieved an accurate method to measure a parameter focusing on positiveness (i.e., estimated support) on Twitter. Thus, we introduced our positiveness ratio (PR) approach, which is expressed as follows:

$$\mathrm{PR} = \mathrm{mPos/tot.\ CC.AA} \tag{2}$$

given as a decimal number that represents the support of a political party. In light of the preceding, which is an indicator of the tendency of an electoral event performed, we compared the support for a selected political party during the elections.

Our point of view in this analysis is not to conduct a predictability approach, but rather to demonstrate an understanding of how Twitter users behave in an electoral context by identifying patterns regarding political parties in certain periods and regions.

In addition to the use of the indicator proposed here, to measure the degree of support in the context of political parties on Twitter, we examined the distribution of the data flow of the combined electoral events. Similarly, measures of central tendency help us to identify the users' data behavior by estimating positivity using the PR indicator but summarized by time in this instance, the four periods, which were divided into the elections of December 20, 2015 and June 26, 2016. In this sense, identifying the sum defined in a certain period could give us some evidence about the overall performance concerning the positivity track. Thus, this represented most of the data coming from the Twitter dataset that defined users' average behavior and their interquartile range.

V. RESULTS

In this section, we present our findings based on our analysis of Twitter and the results of the 2015 and 2016 Spanish general elections, as well as the implementation of the proposed indicator: the positiveness ratio (PR).

Our analysis shows the sequencing of each event divided into four periods: "30 days before the elections," "12 days before the elections," "election day," and "8 days after the elections." Therefore, we selected three autonomous communities—Andalusia, Catalonia, and Madrid—and the four main national political parties—"PP," "PSOE," "Podemos, " and "CS"—for examination in this study.

We explain our approach using scatter plot visualizations for the sake of clarity and to enable a better understanding of the electoral events. The figures illustrate the results of our indicator approach compared with the electoral outcomes, that track the support for each political party.

In brief, the advantages of this research include measuring our findings by contrasting the election outcomes against the data collected from Twitter. Consequently, because the official data were expressed as decimals (see Table 2 and Table 3) after the elections occurred, we believe that the results can be compared by normalizing the scales from these sources. For this reason, we divided our results into two dates according to the Spanish electoral calendar.

TABLE 2. The December 20, 2015 (20D) election results of the four analyzed political parties, shared by the party's percentage expressed in decimals (Source: Ministry of the Interior, Spain, 2015).

CC.AA	Political Party	Results (20D)
Andalusia	PP	0.29
Andalusia	PSOE	0.32
Andalusia	Podemos	0.17
Andalusia	CS	0.14
Catalonia	PP	0.11
Catalonia	PSOE / PSC	0.16
Catalonia	Podemos / En Comú	0.25
Catalonia	CS	0.13
Madrid	PP	0.33
Madrid	PSOE	0.18
Madrid	Podemos	0.21
Madrid	CS	0.19

TABLE 3. The June 26, 2016 (26J) election results of the four analyzed political parties, shared by the party's percentage expressed in decimals (Source: Ministry of the Interior, Spain, 2016).

CC.AA	Political Party	Results (26J)
Andalusia	PP	0.34
Andalusia	PSOE	0.31
Andalusia	Podemos	0.19
Andalusia	CS	0.14
Catalonia	PP	0.13
Catalonia	PSOE / PSC	0.16
Catalonia	Podemos / ECP	0.25
Catalonia	CS	0.11
Madrid	PP	0.38
Madrid	PSOE	0.20
Madrid	Podemos	0.21
Madrid	CS	0.18

A. ELECTORAL RESULTS PER REGION

Through the source provided by the Ministry of the Interior of Spain, the outcomes per party in decimals divided by regions were established, as shown in Table 2 and Table 3. Our quantitative approach involved examining the results of the December 20, 2015 and June 26, 2016 elections according to region. Therefore, we selected three Spanish regions—Andalusia, Catalonia, and Madrid—and related these election outcomes, expressed as decimals, to standardize the data based on the indicator that we proposed in this study.

In the case of the December 20, 2015 (20D) election, the results are detailed by autonomous communities and the three national level political parties analyzed in this study, as shown in Table 2.

Similarly, we focused on the June 26, 2016 (26J) election, and the results of this event are shown in Table 3.

These data provided an analytical basis upon which to compare our proposed indicator, contrasting this with the Twitter outcomes and the results of the electoral events.

Table 3 shows that in Catalonia, a 35% participation share was not captured. This effect is because the prevalence of the regional parties is very high compared to the other analyzed Spanish regions, and this study is focused on the major national parties only.

It was decided to retain the data without applying the mathematical transformation (i.e., a normalization task that can be compared with the data coming from the other two regions), so that they can keep the effect in their original context. It is recognized, however, that such a decision will make an impact in the indicator values of the Catalonia results (see Fig. 4 and Fig. 5).

B. THE DECEMBER 20, 2015 ELECTION

For the December 20, 2015 election, the outcomes and the support were analyzed based on the four periods represented

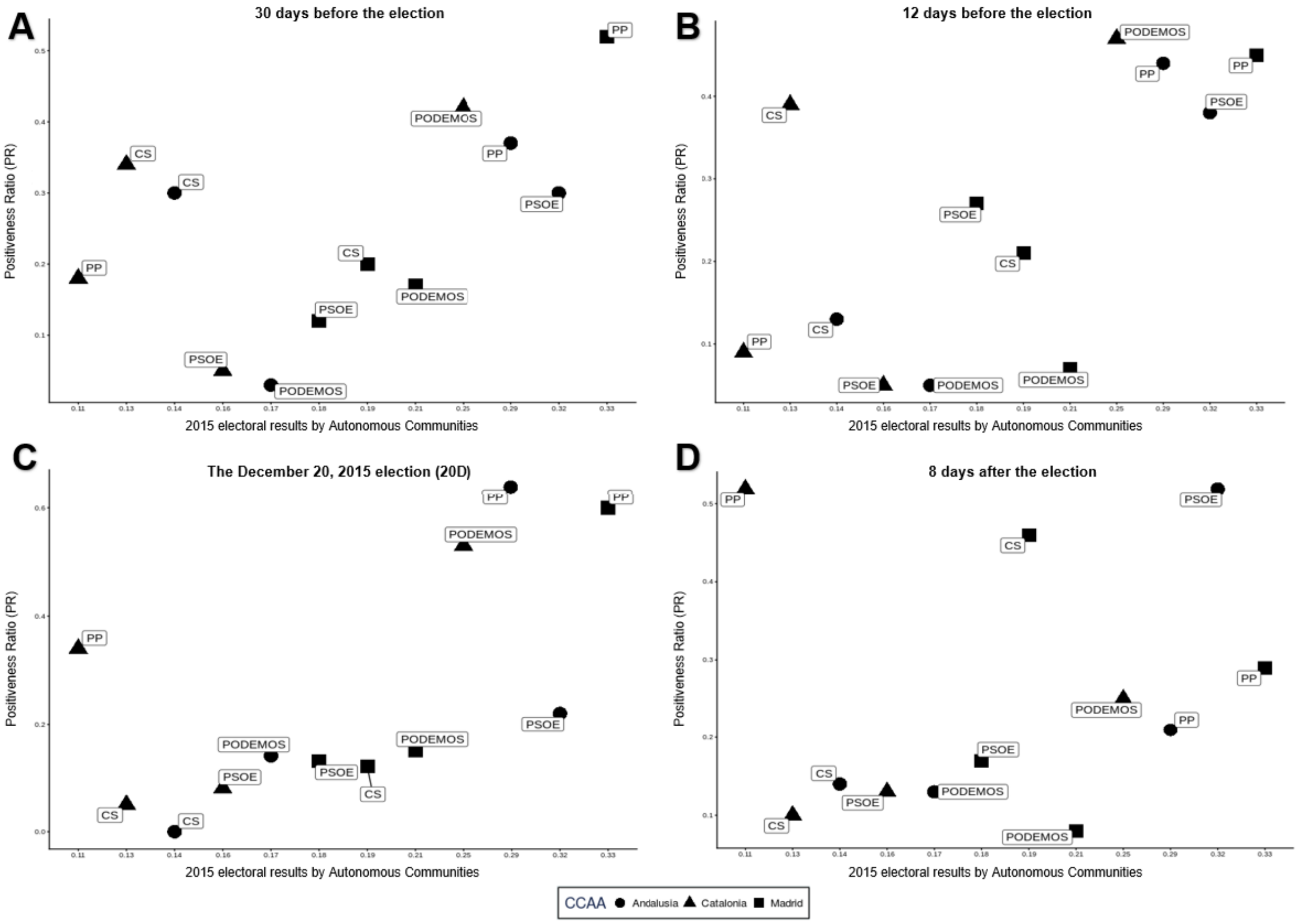


FIGURE 4. **Visualization of the December 20, 2015 (20D) election results and the positiveness ratio (PR) in the four periods analyzed.**

by scatter plots, as depicted in Fig. 4 (A, B, C, D). First, Fig. 4 (A) displays the analysis of "30 days before the elections," beginning with the PP, which had the highest performance in Madrid in correspondence with the PR; in Andalusia had a medium PR but high results; and the lowest support in Catalonia, in accordance with low electoral outcomes. Regarding PSOE, the highest results and PR correspond to Andalusia; thus, in Madrid, there were average results and PR, and these decreased in Catalonia, which had the lowest electoral outcomes and support among the three regions. Regarding the emerging parties, Podemos yielded its highest PR in Catalonia—corresponding with its electoral result— followed by Madrid (with average support), and the lowest support and results were evident in Andalusia. CS had low electoral outcomes in Catalonia, but high PR. Andalusia had low electoral results and medium support reflected in tweets, but there were better vote results in Madrid, contrasting with the lower PR of their supporters received across the regions.

Second, concerning Fig. 4 (B), for the "12 days before the elections," minor changes in this period were reported with respect to the previous interval. Starting this analysis with PP, its highest electoral performance was in Madrid, reflecting the same level of support. In Andalusia, there were high electoral results in combination with the support on Twitter, whereas in Catalonia, there were both low electoral outcomes and low PR level. In the case of PSOE and the behavior of users regarding their support, it reached a peak in Andalusia, but it had a medium vote percentage and PR in Madrid and had the lowest performance in Catalonia. At this point, PP and PSOE displayed similar behavior, both of which were also shown in the previous period. Podemos achieved the highest PR in Catalonia in the same way of high results; however, it had low support in Madrid and Andalusia—having average results. CS, despite the low results in Catalonia, retained high support. Its PR dropped in Andalusia, but in Madrid, it increased its PR.

Third, in Fig. 4 (C), more changes in the behavior of Twitter users regarding the "December 20, 2015" election, became noticeable. PP retained its high PR level and voting outcomes in Madrid, as in Andalusia, and dropped in Catalonia corresponding with its low results. PSOE showed low support, considering the high electoral outcomes in Andalusia. In Madrid and Catalonia, there was also low PR

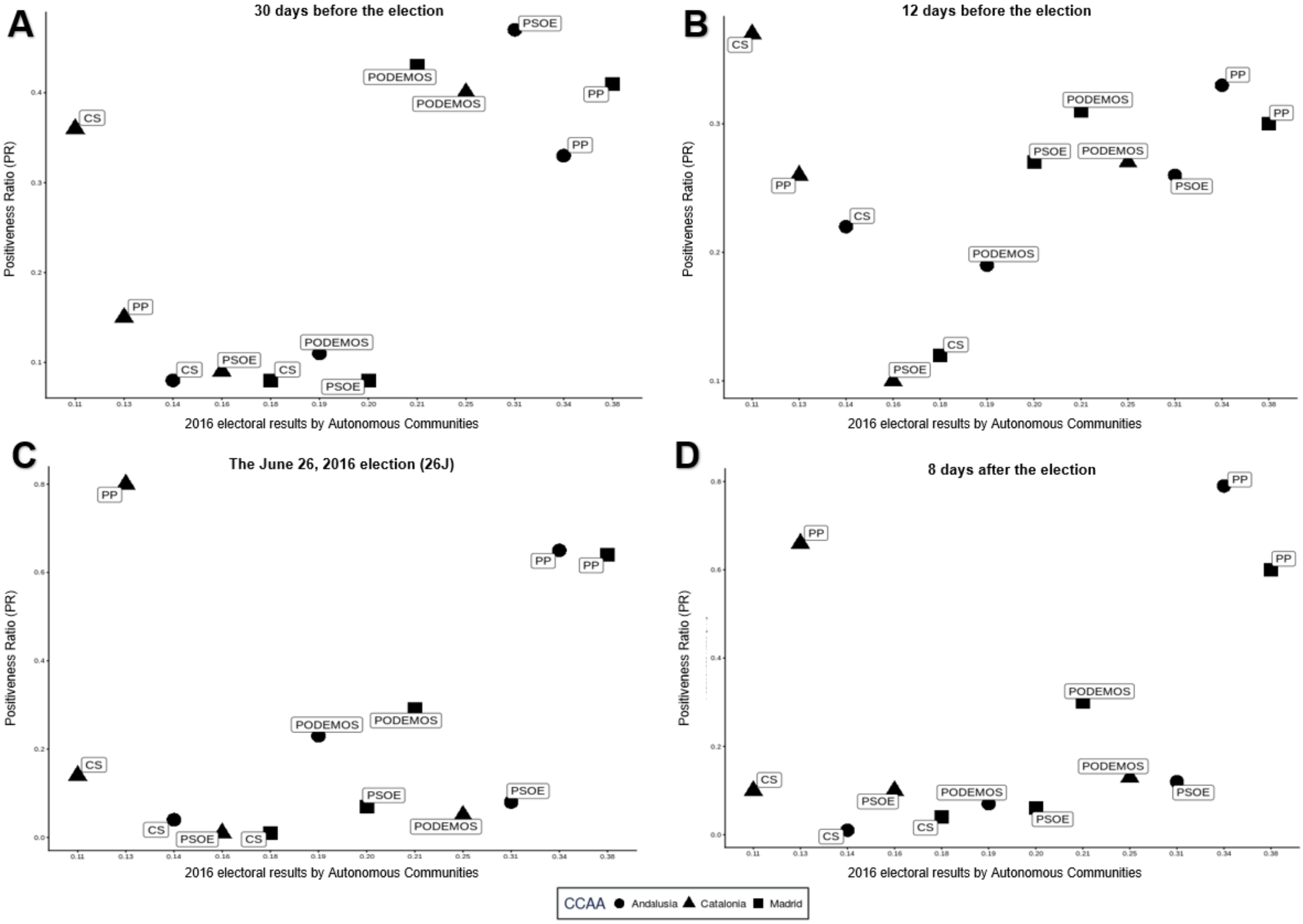


FIGURE 5. Visualization of the June 26, 2016 (26J) election results and the positiveness ratio (PR) in the four periods analyzed.

concerning electoral results in both regions. For Podemos, high support was observed, demonstrating its correspondence with the electoral outcomes in Catalonia and the low support in Madrid and Andalusia. Differing from the previous periods, CS yielded its support in the same way as its low outcomes in Catalonia. Additionally, its PR decreased in Madrid and Andalusia.

Finally, as Fig. 4 (D) shows, in the period "8 days after the elections," changes in the behavior of Twitter users became evident. Regarding PP, support in Madrid surprisingly decreased, as in Andalusia, while it reached a considerably high support level in Catalonia, contrasting with the low outcomes. PSOE reached a high PR in Andalusia but maintained low levels of PR in Madrid and Catalonia as depicted in the previous intervals. After having high support in the previous stages, Podemos' PR decreased in Catalonia, maintaining low support levels in Madrid and Andalusia. In this period, CS substantially increased its support in Madrid (in accordance with its best performance in this region). In Catalonia and Andalusia, as on election day, the low PR continued with its electoral outcomes.

C. THE JUNE 26, 2016 ELECTION

In the election held on June 26, 2016, we followed the same outline performed for the previous election, as shown in Fig. 5 (A, B, C, D). First, in Fig. 5 (A), related to the "30 days before elections," PP support in Madrid corresponded to its electoral results, while in Andalusia, there was high PR and good results. In Catalonia, a coherent low level of support and results were observed. PSOE reached its highest PR in Andalusia, corresponding with the electoral outcomes. In Madrid and Catalonia, however, the support on Twitter was very low, in comparison to its electoral performance, with medium and low vote percentages, respectively. Podemos displayed a high PR in Madrid and Catalonia in accordance with its high vote results but had low support in Andalusia, which did not coincide with the electoral outcomes. Rather, CS experienced a peak in its PR in Catalonia with its low electoral results and similarly low PR and results in Madrid and Andalusia.

Second, in Fig. 5 (B), during the "12 days before the elections," high support for PP in Andalusia and Madrid is apparent, according to the electoral outcomes. In Catalonia, despite its low voting percentage, a considerably high PR was observed during this period. PSOE support in Andalusia slightly decreased, but the PR in Madrid soared according to its preceding period, and in Catalonia, the PR and electoral outcomes remained relatively low. Regarding the new parties, Podemos maintained relatively high support in both Madrid and Catalonia, corresponding with the results of the elections, and in Andalusia, it increased its PR compared with the previous period. CS retained its peak of support in Catalonia, in opposition to its low performance in terms of electoral outcomes, which was the same as in Madrid, where a low PR was maintained, while support in Andalusia had a medium increment.

Third, in Fig. 5 (C), on the "June 26, 2016" election, PP had high support was observed in all regions, which was consistent with its positive electoral outcomes (excepting Catalonia). In the case of PSOE, supporters' behavior on Twitter was consistent: its PR was low, especially in Catalonia and Madrid, considering the average electoral outcomes, but its PR was surprisingly low in Andalusia considering its high voting performance and the performance shown in the previous periods. Podemos ostensibly decreased its support in Catalonia, considering the previous time intervals compared with its results. In Madrid, there was a regressive trend of its PR, showing lower performance on the election day. Andalusia displayed a similar trend regarding average support and electoral outcomes. Despite CS having high PR in Catalonia during the periods before the elections, on the election day, however, this fell dramatically, in line with its low electoral outcomes, and similarly, decreased its support in Andalusia. Except for Madrid, there was continuous low support and electoral outcomes in the previous periods.

Finally, in Fig. 5 (D), which displays the period "8 days after the elections," a constant high level of support for PP can be observed in all regions, continuing the trend of all periods before. PSOE maintained a low PR equal to that of election day in all regions, especially in Catalonia, where its low support was noticed during all periods. Similar to election day, Podemos retained the low support trend in Madrid and Catalonia (where it reached its best result), and even decreasing it in Andalusia. Despite having high support during the previous campaign but poor electoral results, especially in Catalonia, the support for CS dramatically decreased on election and the days after, continuing a regressive trend, as seen in the low support and electoral voting results in Madrid and Andalusia.

D. SUPPORT FLOW DURING THE ELECTIONS

We will now explore our data distribution, based on the positive results of accumulated values that converge on the indicator proposal. Without party differentiation, we estimate the levels of measures of central tendency delineated into the four periods established in the 2015 and 2016 Spanish elections. Therefore, identifying the trends during these events helps us to distinguish between users' activities to confirm whether there have been changes in each event or to make comparisons, considering representativeness in the set of data collected from Twitter.

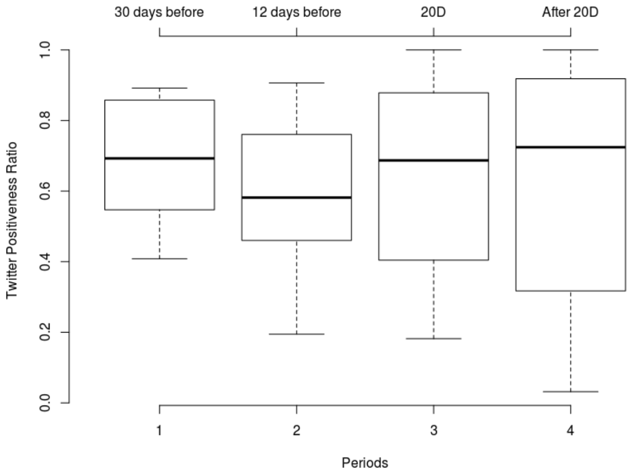


FIGURE 6. Support flow during the 2015 Spanish general election (20D) divided into four periods.

Fig. 6 illustrates the support flow of the 2015 election through measures of central tendency with the help of our indicator, which was used to evaluate all parties divided into four periods. Changes were found in each period, starting with the "30 days before," as seen in the high median of its PR. Next, the decreasing support during the "12 days before" period reached a medium level of near 50% of the median on election day (20D). In contrast, the median of the support progressed, expanding the area of positive values, and finally, the period "after 20D" witnessed a small upward median and a dispersion of the support range, where the majority of positive comments are in the Q1 area.

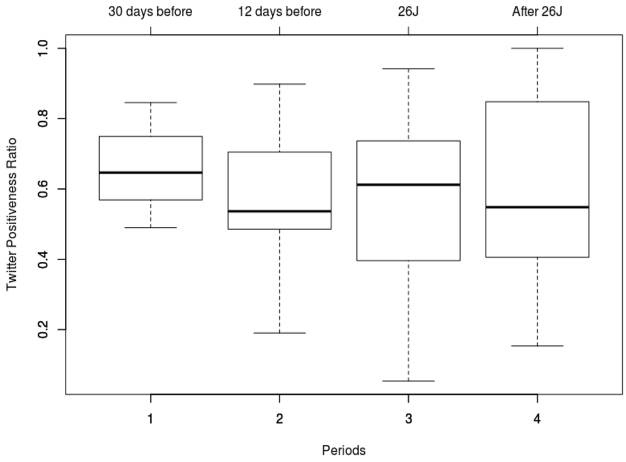


FIGURE 7. Support flow during the 2016 Spanish general election (26J) divided into four periods.

Fig. 7 presents some differences regarding the flow of the 2016 election using the same approach for the measures of central tendency and the indicator results, whereby the dispersion of the positivity is lower compared with the 2015 election in the ''30 days before'' period. Very similar behavior is also shown concerning the median, with a high PR. Then, the ''12 days before'' period displays a downward trend regarding positivity, which is similar behavior to that seen during the previous election event; however, it describes higher occurrences regarding the median and the Q3. In this case, on election day (26J)—similar to the context of the previous elections but less dispersed with regard to the median—the higher appearances are located in the Q1 zone. Finally, in the ''after 26J'' period, differently from the 2015 election, it was noticed that the median dropped compared to the first election period, with a regressive trend of positivity in the Q1 area.

VI. DISCUSSION

Measuring preferences is always a concern with events such as elections, and for this reason, we developed a feasible set of techniques for this task. According to previous recommendations [39], [67], we emphasize the good practices of handling data retrieved from social media platforms. Ultimately, text mining, NLP, and the LinguaKit computational tool were chosen as adequate for addressing this problem, which involved a careful selection of methods. Thus, unlike previous studies [12], [13], in which measures of political events have been characterized based on cumulative mentions regarding two political parties, we separated tweets into three categories: positive, neutral, and negative. Our approach was to focus on users' support based on the positive posts.

In this study, we highlight the behavioral trend evolution on Twitter and elections with a view to proposing new alternatives to measure political events such as the electoral process. We confirmed evidence of correspondence after the comparison of our proposed indicator and the election outcomes, finding a more coherent behavior of supporters in traditional parties, rather than the emerging parties. Consequently, preferences emerged regarding what was expected based on election outcomes. Optimistic users' behavior regarding the emerging political parties can also be observed, especially during the periods before either election day. Nonetheless, considering the fluctuation of users' behavior situations, contradictory results have also been reported.

In the 2015 election (see Fig. 4 [A, B, C, D]), correspondence between the parties' behavior can be seen during the period before and on election day. For instance, the high support for the PP and PSOE parties was reported in the regions in which they have more voters: Madrid and Andalusia, respectively. Surprisingly, the PP showed high support in Catalonia in contrast to the low percentage of votes gained after election day. Conversely, after election day, the support for PP decreased in Madrid and Andalusia, but PSOE support remained high in the last region. It is worth observing that Andalusia and Madrid—as opposed to Catalonia, having regional majority parties' votes—are considered regions where traditional parties such as the PSOE and PP, respectively, predominate [68].

Different behavior concerning emerging parties was reported, however, and their trends fluctuated depending on the region. Podemos retained its high support levels in Catalonia (in accordance with its electoral outcomes), but that support dropped after election day. Its PR stayed relatively low in the other regions. In the case of CS, after showing high support before the election in Catalonia, support dramatically fell on election day and in the succeeding days, in accordance with its low electoral performance. In Madrid and Andalusia, support remained relatively low during all periods, but unexpectedly increased greatly after the election in Madrid.

Concerning the 2016 election (see Fig. 5 [A, B, C, D]), high support can be seen in the case of PP, which maintained high PR during all periods in accordance with their good electoral results (excepting Catalonia). The support of PSOE was constantly high, particularly in Andalusia, before election day; however, it decreased, especially on election day and thereafter, despite the good electoral outcomes in this region.

Regarding the emerging parties—Podemos and CS—support reached high levels for the former in Madrid and Catalonia, in the previous periods before the election day; in contrast, on the ensuing days, support evidently decreased. The latter party, excepting Catalonia, had shown a low level of support, suggesting that having experienced poor results in the 2015 elections, its followers might have demonstrated a lack of support during the 2016 election. In sum, in the 2016 election, the support activities in most of the regions and parties—excluding the winning party, PP—decreased, thereby suggesting that the supporters stopped their activity on social networks after the electoral date.

It can be demonstrated that due to their structure, traditional parties supporters' are more active with their behavior, and it can be more prevalent, as shown in Fig. 4 (A, B, C, D) and Fig. 5 (A, B, C, D), as opposed to less prevalent within emerging parties. Although a high degree of fluctuation among Twitter users in the emerging parties was noticed, particularly in the case of Podemos, a mixed variation along periods and regions was displayed. In addition, we compared these traditional parties with their supporters—in the influenced regions—who demonstrated greater discipline in terms of their behavior, presenting a consistent level of PR during the electoral contest. What the support showed in both instances—that is, Twitter and the election results—was remarkable, as the PP won more seats in the 2016 election. The PP effectively replicated, and even improved, its performance in the 2015 election.

Visualizations of measures of central tendency (see Fig. 6 and Fig. 7) represent the accumulative support during the election periods to enable an understanding the behavior of political party followers. Thus, using these measures of support flow during elections, we confirm the same trends in our indicator, carried out in the 2015 and 2016 elections, with peaks in support 30 days before the elections, on election day, and in the days afterwards. Hence, we realized that the support decreased largely in the period just prior to election day. This situation might underpin the influence of negative comments on political parties and consequently cause a decrease in positive tweets. Curiously, in the period after the 2016 election, the support decreased slightly in a different way from that during the previous 2015 election, which might corroborate that the victorious party (PP) replicated its triumph of the first election date and even gained more votes.

In light of what was published in the media,[7] it is known that the PP party hired the services of The Messina Group (TMG[8]) in 2016. This consulting company specializes in big data and electoral strategies. The information published in the Spanish press suggested that by using social media platforms, the company pushed the party to gain tactical votes during the June 26, 2016 election. The party carried out its political campaign in a strategic way to deliver its message to potential voters. In fact, we argued that this might have influenced Twitter users' behavior in this election; for example, in the days following the election, a positive downward trend is noticeable (see Fig. 7). This result suggests that the company contract had ended, based on the decline in positive activity overall, which might be reflected in the analyzed period.

The present study has some limitations, and their identification should facilitate our efforts to conduct new research. First, finding a method to improve generalizability—that is, including all age groups—to manage our database is relevant, as we could mention the age group of the sample, which is undoubtedly represented by the majority who were young users to the detriment of older users according to a study carried out in the UK [40], referring to those who are active in social media platforms. Second, it was necessary to rely on the indicator presented in this work but to do so in real time, thereby giving us a stream of results, and at the same time, having the progression of a determined electoral event. Finally, it would be necessary to complement the elections with traditional polls or surveys to enable us to compare whether our method could help to track future trends.

Despite these limitations, our results were consistent, suggesting that the proposed methodology may be an effective method of researching events based on Twitter user activity. While it is true that a specific misalignment occurs in some instances, which could be explained by regional factors or campaign strategies that altered users' behavior on social platforms, the research findings revealed measurable positive tendencies during electoral periods. Indeed, due to the preeminence of regional parties in Catalonia, there could be some dissimilarities considered about the indicator proposed. In summary, depending on factors such as the context, culture, regions, and adding more variables, the indicator might be more accurate and experimenting with it is a challenge for further research.

VII. CONCLUSION

In this paper, we have provided evidence that the support for political parties can be measured and tracked through social media platforms, such as Twitter, to obtain intelligence information before the election. With the intention of contributing to a state-of-the-art of Twitter analysis in politics in Spanish language, we proposed the development of the PR indicator as a tool to measure positivity during political events. To our knowledge, the electoral process is a rich source from which emotions can be mined [69], and regarding analytical outputs, the classification of polarity is a good proxy from which to explore polling procedures and other fields.

Nevertheless, the evolution of social media analysis regarding politics is still considered to be in its infancy [13], despite the importance of the topic for political science. Moreover, we presented an advantageous measuring process that can be used to associate Twitter user behavior using a method of quantitative orientation.

The research findings of this study have provided evidence related to support indicators (and their variations) in the field of human activities based on platforms, such as Twitter, which are provided for analysis. Thus, it would be worthwhile to investigate how different event types can be measured regarding decision support systems for products or services and marketing campaigns, among other fields.

An aim of researchers in this field is to provide practical resources for the Spanish language [57], [58], [60], [70], [71]. With the help of open-source tools and establishing new experimentation in keeping with a lexicon-based approach and machine-learning methods, it is worth combining the best practices with a view to achieving steady improvement over time. Despite the limited availability of resources in non-English languages, the interest in and importance of literature in other languages has increased in recent years, and it is because of our focus that we are able to share techniques and learn from others. It would also be interesting to experiment with models such as unsupervised learning to categorize data and to search for differences and similarities among political groups.

Forthcoming research may focus on the "negativeness" approach and how political opponents—termed as "haters"—could affect the performance of parties on political campaigns. In fact, it might be necessary to expand research to other social media platforms to widen the scope with new types of users so that different kinds of indicators can be developed, or reshape its functioning, for example, by adding new variables. In the same way, as a future research line, it would be interesting to apply this measurement process to associate the behavior of the Twitter users in other areas beyond the politics, such as brands and corporations. Moreover, it may be desirable to complement new information sources—that is, interviews or surveys—to supplement the input data with a view to providing details that were not explained because they were outside of the scope of this work.

[7] See "The San Francisco guru who won the elections for Mariano Rajoy"http://www.elmundo.es/cronica/2016/07/03/57779fc0ca4741301d8b4609.html

[8] As they published in their website: https://themessinagroup.com/industries/politics/

ACKNOWLEDGMENT

This work has benefited from discussions with M. Guadalupe Sánchez-Escribano, Luis Ballesteros, T. Bulut Durmaz, John Kline, Xiaochen Zheng, and Ermal Hetemi.

REFERENCES

[1] T. Persson and G. Tabellini, "Democratic capital: The nexus of political and economic change," *Amer. Econ. J. Macroecon.*, vol. 1, no. 2, pp. 88–126, Jun. 2009.

[2] R. Inglehart and P. Norris, "Trump, Brexit, and the rise of populism: Economic have-nots and cultural backlash," *SSRN Electron. J.*, Jul. 2016.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, vol. 10, 2002, pp. 79–86.

[4] B. Pang and L. Lee, *Opinion Mining and Sentiment Analysis*, vol. 2, nos. 1–2. Boston, MA, USA: Now, 2008.

[5] S. Kruikemeier, "How political candidates use Twitter and the impact on votes," *Comput. Human Behav.*, vol. 34, pp. 131–139, May 2014.

[6] S. Quinlan, M. Shephard, and L. Paterson, "Online discussion and the 2014 Scottish independence referendum: Flaming keyboards or forums for deliberation?" *Elect. Stud.*, vol. 38, pp. 192–205, Jun. 2015.

[7] D. Jiang, X. Luo, J. Xuan, and Z. Xu, "Sentiment computing for the news event based on the social media big data," *IEEE Access*, vol. 5, pp. 2373–2382, 2017.

[8] J. Borge-Holthoefer *et al.*, "Structural and dynamical patterns on online social networks: The Spanish May 15th movement as a case study," *PLoS ONE*, vol. 6, no. 8, Aug. 2011, Art. no. e23883.

[9] Y.-H. Eom, M. Puliga, J. Smailović, I. Mozetič, and G. Caldarelli, "Twitter-based analysis of the dynamics of collective attention to political parties," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0131184.

[10] S. Ahmed, K. Jaidka, and J. Cho, "The 2014 Indian elections on Twitter: A comparison of campaign strategies of political parties," *Telematics Inform.*, vol. 33, no. 4, pp. 1071–1087, 2016.

[11] S. Hong and S. H. Kim, "Political polarization on Twitter: Implications for the use of social media in digital governments," *Government Inf. Quart.*, vol. 33, no. 4, pp. 777–782, 2016.

[12] J. Borondo, A. J. Morales, J. C. Losada, and R. M. Benito, "Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish Presidential election as a case study," *Chaos Interdiscip. J. Nonlinear Sci.*, vol. 22, no. 2, 2012, Art. no. 023138.

[13] G. Caldarelli *et al.*, "A multi-level geographical study of italian political elections from Twitter data," *PLoS ONE*, vol. 9, no. 5, May 2014, Art. no. e95809.

[14] P. Aragón, K. E. Kappler, A. Kaltenbrunner, D. Laniado, and Y. Volkovich, "Communication dynamics in Twitter during political campaigns: The case of the 2011 Spanish national election," *Policy Internet*, vol. 5, no. 2, pp. 183–206, Jun. 2013.

[15] P. Barberá and G. Rivero, "Understanding the political representativeness of Twitter users," *Soc. Sci. Comput. Rev.*, vol. 33, no. 6, pp. 712–729, Dec. 2015.

[16] R. A. Feenstra, S. Tormey, A. Casero-Ripollés, and J. Keane, *Refiguring Democracy: The Spanish political laboratory*, 1st. London, U.K.: Routledge Focus, 2017.

[17] P. Gamallo and M. Garcia, "LinguaKit: A multilingual tool for linguistic analysis and information extraction," *Linguamática*, vol. 9, no. 1, pp. 19–28, Jun. 2017.

[18] E. Martinez-Camara, M. T. Martín-Valdivia, L. A. Ureña-Lopez, and A. Montejo-Raez, "Sentiment analysis in Twitter," *Nat. Lang. Eng.*, vol. 20, no. 01, pp. 1–28, Jan. 2012.

[19] J. S. Cesteros, A. Almeida, and D. L. De Ipiña, "Sentiment analysis and polarity classification in Spanish Tweets," in *Proc. TASS*, 2015, pp. 23–28.

[20] D. Vilares and M. A. Alonso, "A review on political analysis and social media," *Procesamiento Lenguaje Natural*, vol. 56, pp. 13–23, Mar. 2016.

[21] T. H. McCormick, H. Lee, N. Cesare, A. Shojaie, and E. S. Spiro, "Using Twitter for demographic and social science research: Tools for data collection and processing," *Sociol. Methods Res.*, vol. 46, no. 3, pp. 390–421, Aug. 2017.

[22] L. Alonso-Muñoz and A. Casero-Ripolles, "Political agenda on Twitter during the 2016 Spanish elections: Issues, strategies, and users' responses," *Commun. Soc.*, vol. 31, no. 3, pp. 7–25, 2018.

[23] F. Bravo-Marquez, D. Gayo-Avello, M. Mendoza, and B. Poblete, "Opinion dynamics of elections in Twitter," in *Proc. 8th Latin Amer. Web Congr.*, Oct. 2012, pp. 32–39.

[24] T. Graham, D. Jackson, and M. Broersma, "New platform, old habits? CandidatesŠ use of Twitter during the 2010 British and Dutch general election campaigns," *New Media Soc.*, vol. 18, no. 5, pp. 765–783, May 2016.

[25] P. Nulty, Y. Theocharis, S. A. Popa, O. Parnet, and K. Benoit, "Social media and political communication in the 2014 elections to the European Parliament," *Elect. Stud.*, vol. 44, pp. 429–444, Dec. 2016.

[26] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Election forecasts with Twitter: How 140 characters reflect the political landscape," *Soc. Sci. Comput. Rev.*, vol. 29, no. 4, pp. 402–418, Nov. 2011.

[27] M. P. Cameron, P. Barrett, and B. Stewardson, "Can social media predict election results? Evidence from New Zealand," *J. Political Marketing*, vol. 15, no. 4, pp. 416–432, Oct. 2016.

[28] H. G. Yoon, H. Kim, C. O. Kim, and M. Song, "Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling," *J. Informetrics*, vol. 10, no. 2, pp. 634–644, 2016.

[29] J. DiGrazia, K. McKelvey, J. Bollen, F. Rojas, and C. Danforth, "More Tweets, more votes: Social media as a quantitative indicator of political behavior," *PLoS ONE*, vol. 8, no. 11, Nov. 2013, Art. no. e79449.

[30] M. Gaurav, A. Kumar, A. Srivastava, and S. Miller, "Leveraging candidate popularity on Twitter to predict election outcome," in *Proc. 7th Workshop Social Netw. Mining Anal.*, 2013, Art. no. 7.

[31] K. Singhal, B. Agrawal, and N. Mittal, "Modeling Indian general elections: Sentiment analysis of political twitter data," in *Information Systems Design and Intelligent Applications*. New Delhi, India: Springer, 2015, pp. 469–477.

[32] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, "140 characters to victory?: Using Twitter to predict the UK 2015 general election," *Elect. Stud.*, vol. 41, pp. 230–233, Mar. 2016.

[33] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Assoc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.

[34] D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj, "Limits of Electoral Predictions using Twitter," in *Proc. 5th Int. AAAI Conf. Weblogs Social*, 2011, pp. 1–4.

[35] D. Gayo-Avello, "No, you cannot predict elections with Twitter," *IEEE Internet Comput.*, vol. 16, no. 6, pp. 91–94, Nov. 2012.

[36] D. Gayo-Avello. (Apr. 2012). "'I wanted to predict elections with Twitter and all I got was this Lousy Paper'—A balanced survey on election prediction using Twitter data." [Online]. Available: https://arxiv.org/abs/1204.6441

[37] P. T. Metaxas, E. Mustafaraj, and D. Gayo-Avello, "How (not) to predict elections," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust, IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 165–171.

[38] A. Jungherr, P. Jurgens, and H. Schoen, "Why the pirate party won the german election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. 'Predicting elections with twitter: what 140 characters reveal about political sentiment,'" *Soc. Sci. Comput. Rev.*, vol. 30, no. 2, pp. 229–234, May 2012.

[39] D. Gayo-Avello, "A meta-analysis of state-of-the-art electoral prediction from Twitter data," *Soc. Sci. Comput. Rev.*, vol. 31, no. 6, pp. 649–679, Dec. 2013.

[40] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who Tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0115545.

[41] D. Murthy, "Twitter and elections: Are Tweets, predictive, reactive, or a form of buzz?" *Inf., Commun. Soc.*, vol. 18, no. 7, pp. 816–831, Jul. 2015.

[42] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Where there is a sea there are pirates: Response to Jungherr, Jürgens, and Schoen," *Soc. Sci. Comput. Rev.*, vol. 30, no. 2, pp. 235–239, May 2012.

[43] W. M. Haddad, V. S. Chellaboina, and Q. Hui, *Nonnegative and Compartmental Dynamical Systems*. Princeton, NJ, USA: Princeton Univ. Press, 2010.

[44] E. De La Poza, L. Jódar, and A. Pricop, "Modelling and analysing voting behaviour: The case of the Spanish general elections," *Appl. Econ.*, vol. 49, no. 13, pp. 1287–1297, Mar. 2017.

[45] P. Singh, R. S. Sawhney, and K. S. Kahlon, "Predicting the outcome of Spanish general elections 2016 using Twitter as a tool," in *Advanced Informatics for Computing Research* (Communications in Computer and Information Science), vol. 712, D. Singh, B. Raman, A. Luhach, and P. Lingras, Eds. Singapore: Springer, 2017, pp. 73–83.

[46] A. Lopez-Meri, S. Marcos-Garcia, and A. Casero-Ripolles, "What do politicians do on Twitter? Functions and communication strategies in the Spanish electoral campaign of 2016," *El Profesional Información*, vol. 26, no. 5, pp. 795–804, Sep. 2017.

[47] S. Rusell and P. Norvig, *Artificial Intelligence. A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2010.

[48] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015.

[49] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.

[50] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[51] S. Mukherjee and P. K. Bala, "Detecting sarcasm in customer tweets: An NLP based approach," *Ind. Manage. Data Syst.*, vol. 117, no. 6, pp. 1109–1126, Jul. 2017.

[52] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizensŠ political preferences with an application to Italy and France," *New Media Soc.*, vol. 16, no. 2, pp. 340–358, Mar. 2014.

[53] D. J. Hopkins and G. King, "A method of automated nonparametric content analysis for social science," *Amer. J. Political Sci.*, vol. 54, no. 1, pp. 229–247, Jan. 2010.

[54] W. Magdy, K. Darwish, and I. Weber. (Mar. 2015). "#FailedRevolutions: Using Twitter to study the antecedents of ISIS support." [Online]. Available: https://arxiv.org/abs/1503.02401

[55] L. Padró, "Analizadores Multilingües en FreeLing," *Linguamática*, vol. 3, no. 1, pp. 13–20, 2011.

[56] G. Sidorov *et al.*, "Empirical study of machine learning based approach for opinion mining in Tweets," in *Advances in Artificial Intelligence*, I. Batyrshin and M. G. Mendoza, Eds. Berlin, Germany: Springer, 2013, pp. 1–14.

[57] F. Pla and L.-F. Hurtado, "Sentiment analysis in Twitter for Spanish," in *Proc. Int. Conf. Appl. Natural Lang. Data Bases/Inf. Syst.*, 2014, pp. 208–213.

[58] M. S. Deas, O. Biran, K. Mckeown, and S. Rosenthal, "Spanish Twitter messages polarized through the lens of an English system," in *Proc. TASS*, 2015, pp. 81–86.

[59] P. Gamallo, M. Garcia, C. Pineiro, R. Martinez-Castaño, and J. C. Pichel, "LinguaKit: A big data-based multilingual tool for linguistic analysis and information extraction," in *Proc. 5th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, 2018, pp. 239–244.

[60] D. Vilares, M. García, M. A. Alonso, and C. Gómez-Rodríguez, (Aug. 2017). "Towards syntactic Iberian polarity classification." [Online]. Available: https://arxiv.org/abs/1708.05269

[61] R. Xiao, "Corpus creation," in *Handbook of Natural Language Processing*, N. Indurkhya and F. Damerau, Eds., 2nd ed. London, U.K.: CRC Press, 2010, pp. 147–165.

[62] L. Dubiau and J. M. Ale, "Análisis de Sentimientos sobre un Corpus en Español: Experimentación con un Caso de Estudio," in *Proc. ASAI*, 2013, pp. 36–47.

[63] J. Mahmud, J. Nichols, and C. Drews, "Where is this Tweet from? Inferring home locations of Twitter users," in *Proc. ICWSM*, 2012, pp. 511–514.

[64] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.

[65] P. Gamallo, J. C. Pichel, M. García, J. M. Abuín, and T. F. Pena, "Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data," *Procesamiento Lenguaje Natural*, vol. 53, pp. 17–24, Sep. 2014.

[66] J. Borondo, A. J. Morales, J. C. Losada, and R. M. Benito, "Analyzing the usage of social media during Spanish presidential electoral campaigns," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 785–792.

[67] Z. Wang, V. Joo, C. Tong, and D. Chan, "Issues of social data analytics with a new method for sentiment analysis of social media data," in *Proc. IEEE 6th Int. Conf. Cloud Comput. Technol. Sci.*, Dec. 2014, pp. 899–904.

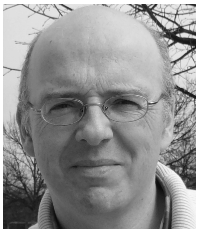
[68] S. Balfour, "The 2015 Spanish general election: A final look at the parties and the polls," in *Proc. LSE Eur. Politics Policy (EUROPP)*, 2015. [Online]. Available: https://blogs.lse.ac.uk/europpblog/2015/12/17/the-2015-spanish-general-election-a-final-look-at-the-parties-and-the-polls/

[69] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, "Sentiment, emotion, purpose, and style in electoral tweets," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 480–499, 2015.

[70] A. M. F. Montraveta, "La construcción del WordNet 3.0 en español," in *La lexicografía en su dimensión teórica*. 2010, pp. 201–220.

[71] D. Vilares, M. Thelwall, and M. A. Alonso, "The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets," *J. Inf. Sci.*, vol. 41, no. 6, pp. 799–813, 2015.

JOSE N. FRANCO-RIQUELME received the M.Sc. degree in innovation management from the National University of Asuncion (UNA), Paraguay, in 2015. He is currently pursuing the Ph.D. degree in economics and innovation management with the Universidad Politécnica de Madrid (UPM), Spain. He is also a Research Fellow of the Center for Technology Innovation, Universidad Politécnica de Madrid (*Centro de Apoyo a la Innovación Tecnológica*, CAIT-UPM). His main research interests include business analytics, machine learning, text mining, and sentiment analysis in social media. In addition, his investigation includes open innovation models, knowledge innovation business services (KIBS), and data driven innovation.



ANTONIO BELLO-GARCIA has been a Full Professor of computer graphics in engineering with the University of Oviedo, Spain, since 2008. His research is focused on computer-aided design, computer graphics and data mining, and visualization. He teaches courses on CAD/BIM, optimization techniques in engineering, and digital image processing. He has participated in over 30 national and international projects, funded by the European Commission and the Spanish Ministry for Science and Innovation. He has also published over 30 papers in national and international scientific journals.

JOAQUÍN ORDIERES-MERÉ received the Ph.D. degree in industrial engineering from the Universidad Nacional de Educación a Distancia (UNED), in 1987.

He was a Full Professor in industrial management with the Universidad de la Rioja, in 1997, and at the Universidad Politécnica de Madrid, since 2009. His research interests are connected to managerial dimension in industry, including project management and business analytics. In particular, his focuses are process data modelling to improve the knowledge and optimize those processes. He was involved in over 50 research projects, most of them international and competitive. He has published over 100 research papers, with accumulated cites over 3700 and h-factor of 20. He participates into ISO TC groups and serves regularly as a Reviewer for different journals, including the Editor Member Board in some journals, such as the *International Journal of Data Mining, Modelling, and Management*. He also represents his country as a member of some European Union established committees such as RFCS TSG8.