World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:9, 2018

Real Time Classification of Political Tendency of Twitter Spanish Users based on Sentiment Analysis

Marc Solé, Francesc Giné, Magda Valls, Nina Bijedic

***Abstract*—**What people say on social media has turned into a rich source of information to understand social behavior. Specifically, the growing use of Twitter social media for political communication has arisen high opportunities to know the opinion of large numbers of politically active individuals in real time and predict the global political tendencies of a specific country. It has led to an increasing body of research on this topic. The majority of these studies have been focused on polarized political contexts characterized by only two alternatives. Unlike them, this paper tackles the challenge of forecasting Spanish political trends, characterized by multiple political parties, by means of analyzing the Twitters Users political tendency. According to this, a new strategy, named Tweets Analysis Strategy (TAS), is proposed. This is based on analyzing the users tweets by means of discovering its sentiment (positive, negative or neutral) and classifying them according to the political party they support. From this individual political tendency, the global political prediction for each political party is calculated. In order to do this, two different strategies for analyzing the sentiment analysis are proposed: one is based on Positive and Negative words Matching (PNM) and the second one is based on a Neural Networks Strategy (NNS). The complete TAS strategy has been performed in a Big-Data environment. The experimental results presented in this paper reveal that NNS strategy performs much better than PNM strategy to analyze the tweet sentiment. In addition, this research analyzes the viability of the TAS strategy to obtain the global trend in a political context make up by multiple parties with an error lower than 23%.

***Keywords*—**Political tendency, prediction, sentiment analysis, Twitter.

I. Introduction

WITH the growth of social network usage and its published content, a great deal of information can be found about its users and opinions. This occurs because most of the content of social networks reflects opinions and interests of its users [3], [4]. Thus, social networks have turned into an excellent platform for analysis of social trends in different fields, such as sport, fashion, politics, etc... [5].

Nowadays, there are three most used social networks: Facebook, Instagram and Twitter. Each of these social networks has a different predominant usage type. For example, people tend to use Facebook to share their live events or feelings with their friends. However, the access to the users' information is limited, which difficults the extraction of knowledge. Another social network with a huge number of users is Instagram; which is based on sharing image contents and it is again more related to lifestyle. In like manner, it does not contain many opinions of its users. In contrast to the previous two social networks, Twitter users typically express their opinion related to real time social events in a limited small number of characters. Likewise, the majority of them are usually public [6]. Those reasons have caused that in the last years, twitter users accounts have been increasing year after year and, as a consequence, Twitter is turning into a powerful tool to know real time user opinions related to social events in an easy and direct way [4].

According to this, the classification of the Twitter users in relation to its public opinion and how to recognize them as representatives of the society opinion has turned into a big challenge for the research community. Thus, in the last years, researchers have proposed different algorithms to classify users in different ways: according to the user's profile and their usage of the application (personal, professional, business, spam and viral users) [7], according to their entity (individual user or organization) [8], according to their interests (entertainment, culture, business finance, politics, technology, sports and others) [9], according to different attributes (gender, age, regional origin and political orientation) [10] or even according to their political affiliation or ethnicity [10]. The amount of published papers proves that Twitter is a relevant social network to gather data to classify the users in different types and relate them to parts of the society.

In this way, this research is focused into analyzing Twitter users opinion in order to forecast their political trends inside the Spanish political context. The motivation is given by two different facts. First of all, the traditional methodology for obtaining the political tendency of the citizens, such as telephone polls, has a high cost [14] and it does not result in satisfactory indicators [12]. As a representative example, the case of the Spanish elections in 2016 [13], where the results of the polls showed a significant difference from the real results. In some cases, the differences between the real results and the polls were higher than 20 points, which gives an idea about the difficulty to analyze the political tendency [14]. Secondly, in the last years, the politicians have been using their Twitter accounts to send messages to potential voters and citizens in general, as an easy and cheap way of frontal interactions. In such a manner, politicians interact with the citizens, and this interaction is related to the intention to motivate voters to vote for them [11]. Also the political TV programs and news use the politicians tweets to relate the politicians' opinions on some event.

M. Solé is with the Department of Computer Science at University of Lleida, Spain (e-mail: marcsolefarre90@gmail.com).
F. Giné is with the Department of Computer Science at University of Lleida, Spain.
M. Valls is with the Department of Mathematics at University of Lleida, Spain.
N. Bijedic is with the Department of Mathematics at Dzemal Bijedic University (Bosnia & Herzegovina).

This paper tackles the challenge of forecasting Spanish political trends by means of analyzing the Twitters Users political tendency. According to this aim, a novel strategy, named tweets Analysis Strategy (TAS), is proposed. This is based on analyzing the users tweets by means of discovering its sentiment and classifying them according to the political party that they support. In order to do this, TAS implementation is divided into three main sections:

A. *Data Initialization* contextualizes the environment and retrieves the input data from Twitter. Contextualization is related to the Spanish political context and it is based on creating data sets with typical Spanish words, which are used to collect the spanish political tweets.
B. *Data Processing* classifies the user tweets to the political parties and obtains the tweets' sentiment that can be positive, negative or neutral. In order to do this, two different methods for sentiment analysis are discussed: the first one is based on Positive and Negative words Matching (PNM) and the second one is based on a Neural Networks Strategy (NNS). The results reveal that neural networks performs much better than matching keywords strategy.
C. *Data Post-Processing* generates the output of the strategy; it means the political tendency for each user and the global political tendency of Spanish users by averaging users' tendencies.

It is worth pointing out that although this proposal is focused on the Spanish context, the proposed methodology can be adapted for any other context characterized by multiple political parties.

The effectiveness of the both proposed sentiment analysis strategies is evaluated by a data corpus with 312,369 positive and 343,011 negative tweets. The NNS strategy performs better than PNM with a best assert of 71%. Additionally, we have evaluated the global trend given by the TAS strategy by means of comparing a specific set of 184 users, previously analyzed manually, with the results given by the TAS strategy. In this comparison, the TAS strategy has achieved a 60% of assert, 22% of error and 18% of neutral results. Likewise, it is shown that the system is able to scale properly when a huge number of tweet are retrieved and analyzed in streaming process, because the strategy are developed and deployed in a Apache Storm Cluster [28] and that provides an architecture to process all data in parallel processes.

The remainder of this paper is organized as follows. Section II compares the main contributions of the literature about forecasting political trends from Twitter in relation to the proposal. Section III describes in detail the three sections making up the TAS strategy proposed in this paper. Section IV evaluates the performance of both sentiment analysis strategies together with the global political prediction given by the TAS strategy. Finally, Section V outlines the main conclusions and future work.

II. Twitter Political Tendency

The analysis of the political tendency using the opinion of people through social networks is a challenge for the research community. Therefore, researches have proposed different strategies to predict the users' political tendency.

There are a significant set of works, which are focused on the US elections system, which is limited to a binary prediction between liberal/democrat and conservative/republican [15], [17]-[19]. In contrast to them, the proposal is oriented to the Spanish context, which is characterized by multiple dimensions. For instance, in the elections on 2016, nine political parties obtained representation in the parliament, which are stratified from the right to the left wing, and also according to their general/regional affiliation. So, the magnitude of the problem faced up by the TAS strategy is higher than in the previous reported works.

Among the previous reported works, special interest has the strategy, presented in [15], which uses data set of 3,938 users with political ideology labels self-reported through survey and classifies them in two groups, Conservative and Liberal. Authors characterize the political groups of users through the language used on Twitter. From that, they build a fine-grained model to predict political ideology of unseen users. In this way, they are able to identify politically moderate and neutral users, such as the TAS strategy presented in this paper does.

Another interesting strategy is presented in [19]. This is also focused on classifying the political users by the right or left orientation, labeling political orientation of 1000 twitter users. Authors propose two types of classification, content analysis and communication networks. The content analysis is based on analyzing tweet's text, word frequency, hashtag frequency for each user, and Latent semantic Analysis of hashtags (LSA), a technique to discover a set of topics. The communication networks approach ignores tweet content and instead focused on relationships between users. Authors claim that the analysis of political communication networks provide highest accuracy, although the information-rich hashtag features are almost as effective and have the benefit of generalizing without the need to re-cluster the network to accommodate new users. This last conclusion has been taken into account to develop the TAS strategy.

One work focused on the German election which is related to a multi-dimensional problem, such as this paper does, is presented in [16]. Authors predict by means of a sentiment analysis tool the political tendency of the twitter users analyzing 100,000 Twitter messages mentioning political parties or politicians. In one hand, the authors conclude that the simple count of the number of tweets is related to the voter tendencies and they are similar to the traditional election sampling. In the other hand, the sentiment of Twitter messages could be related to the political programs, candidate profiles, and evidence from the content of the election campaign. In contrast to that, this research presents a complete system able to tackle the political prediction problem from the data input retrieval, the data processing to extract the sentiment of the tweets and the computation of the global political tendency.

III. Tweets Analysis Strategy (TAS)

The TAS strategy presented in this paper is based on retrieving political tweets, analyzing their sentiment and

classifying each tweet according to the political party that it is related to. Hence, each time a tweet is analyzed, in return it was obtained both its sentiment and the political parties related to it. Once all the tweets for one user have been analyzed, the overall polarity was computed for each political Party relevant for the user, and hence the political tendency of a given user. The last step is based on computing all the tendencies of all users to obtain the global political tendency among the analyzed users in the society. It is important to denote that all of the strategy to extract and analyze tweets and users is performed with a streaming platform. This means that the processes of tweet retrieval and analysis, as well as tendency update, are all performed in real time. In that sense, tendency updates always reflect actual political pulse of Spanish Twitter users.

The TAS strategy is designed in the following three main blocks (see Fig. 1):

A. *Data Initialization*: The first block is focused on contextualization and input data retrieval from Twitter. This block is divided in three parts:
   1) *Context Definition*: In order to focus on a particular context, several sets of keywords will be defined prior to TAS strategy implementation. Context definition has two input purposes, the first one is to extract tweets within a concrete context (Spanish politics, in this case) and the second is to classify tweets according to the political parties they refer to.
   2) *Twitter Application programming Interface (API)*: Retrieves input information to TAS about users and their tweets. There are several types of Twitter API functions. In this case, the streaming API functionality has been used.
   3) *Data Extraction*: This block extracts tweets from the Twitter API, filtering those tweets launched by Spanish users and which are related to political context.

B. *Data Processing*: The second block is devoted to obtain the sentiment of each tweet, and classify them according to its political affinity. This block is divided in two parts:
   1) *Sentiment Analysis*: In order to obtain the sentiment of a tweet text two methods were discussed, *Positive and Negative Matching (PNM)* and *Neural Networks Strategy (NNS)* [2].
   2) *Collection Classifier*: Using the keywords sets defined in the context definition block, each tweet was classified according to its affinity to Spanish political parties.

C. *Data Post-Processing*: This block provides two outputs based on the process of tendency update for each user:
   1) *User Tendency Update*: When a analyzed tweet enters this process, the strategy recalculates both the User political tendency and the global political tendency. One of the outputs of this precess is the political tendency related to each political party for one user.
   2) *Global Political Tendencies Update*: The other output is the global political tendency related to each political party for all of the users, there are calculated in this process.

A. Data Initialization

1) Context Definition: In order to analyze social data, it is necessary to focus on a context target. In particular, this paper is focused on the Spanish political context. However, the overall strategy can be applied to different contexts, such as a different countries' political elections, a particular sport competition, fashion tendencies or the most popular product items.

Context initialization is given by appropriately selecting two keyword sets. These sets are defined before performing the Data extraction phase and are relevant words extracted from official web pages, twitter accounts or campaign hashtags from the different political parties.

The first data set is the Keywords set, K, which contains the main keywords related the context. Hence, those tweets containing any of the words in K will be captured when retrieving tweets.

Once the target tweets are retrieved, the next step will be classify each of them with regard to the political parties they refer to. So, let X be the set of political parties (in the case of Spain, $|X|$ = Partido Popular (PP), Partido Socialista Obrero (PSOE), Ciudadanos (CS), Esquerra Republicana de Catalunya (ERC), Partit Democrata Catala (PDECAT) and Podemos). Then, for each party $x \in X$ the Collections Keywords set C_x is defined, containing the relevant Keywords that will be used to classify tweets which are related to x. There are three different types of keywords in K and C_x: user entities starting with @, hashtags starting with # and political words (like "Partido Popular", for instance). Notice that the keywords of the C_x Collection Keywords can be similar to K Keywords Set but not necessary equal. This is because some words that appear in C_x could produce a confusion during the extracting process of the tweets, because of their ambiguity. It is discussed in more detail in the Data Extraction block.

2) Twitter API: The tool for tweets extraction has been the Twitter API [20], and more specifically the StreamingAPI, because it provides real-time information on tweets and users. To obtain the political tweets, the track function was used of the Twitter API. It allows retrieving tweets whose content matches some of the words in the Keywords set K.

3) Data Extraction: The main purpose of the Data extraction block is the refinement of the political tweets extracted with the Twitter Streaming API. Therefore, the input of this process is potential political tweets and the output is tweets that have political context. The need for refinement arises from the fact that the used keywords sometimes can retrieve tweets with non-political context. In particular, with this strategy, there were some problems with ambiguous Keywords named "soft keywords" that can be a political text in K, but can not be neither a Hashtag nor a user. The reason for this is that those words could have different meanings, so the procedure could return non-political tweets. An example of a soft keyword could be "Podemos", since it can be the Spanish

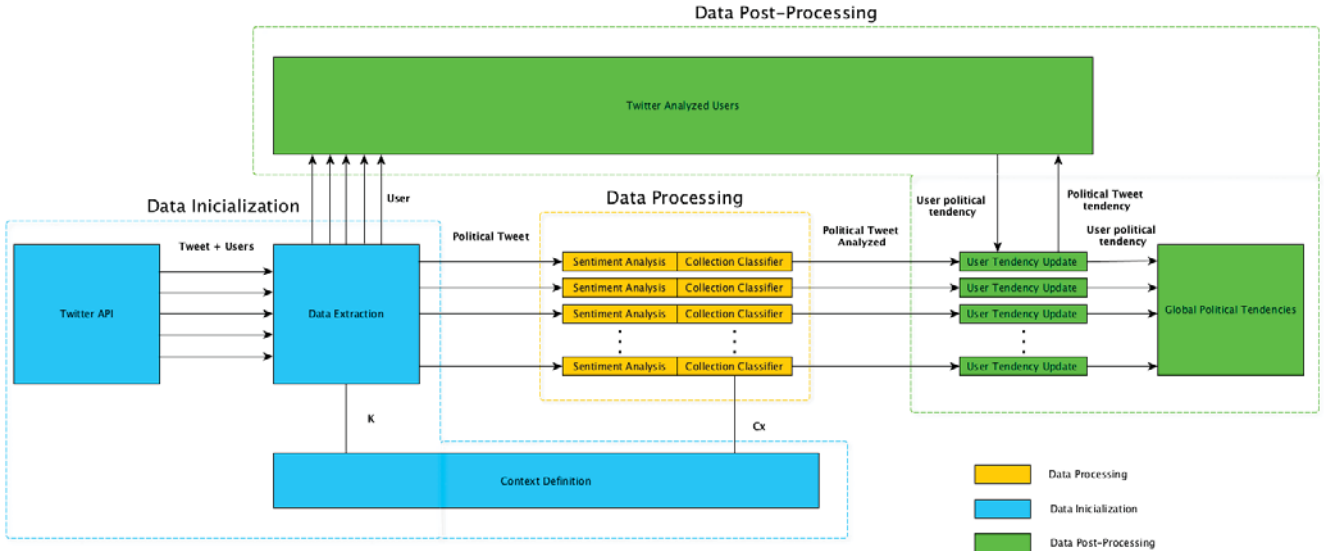

Fig. 1 TAS Diagram

political party or the conjugation of the verb "Poder". To solve this problem, the soft keywords are identified and when they appear in a tweet, model searches for another matching with another non-ambiguous word. If it finds a soft keyword with another non-ambiguous word, the tweet is analyzed. On the other hand, if the soft keyword is the only $k \in K$ in the tweet, then the tweet is discarded.

As an output, this process sends political tweets to the Data Processing block and also collects those twitter users, who have launched political-related tweets in a given context.

B. Data Processing

1) Sentiment Analysis: There are numerous research papers and studies that focus on sentiment classification for Twitter. In the literature appear different techniques and methodologies to detect and identify twitter data sentiment, such as:

- The Machine Learning Approach [22], which uses linguistic features.
- The Lexicon-based Approach [21] that relies on a sentiment lexicon, a collection of known and precompiled sentiment terms.
- The last is divided into dictionary-based approach and corpus-based approach which uses statistical or semantic methods to find sentiment polarity. The hybrid Approach [23] combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.

These studies describe methodologies to detecting and identifying twitter data sentiment.

The proposal of this paper relies partially on the methodology described in [23], which uses emoticons for tweets feature extraction and includes the usage of unigrams and/or bigrams (couple of words that appear together), as well as parts of speech tags, to design an algorithm to classify the sentiment of the tweets.

The objective of the Sentiment Analysis block is to obtain a positive, negative, or neutral value per tweet text. For such a purpose, two possible and different strategies have been developed:

- *Positive Negative Match (PNM)*, which is focused on matching and counting positive and negative words/emoticons inside a tweet.
- *Neural Network Sentiment (NNS)*, which is based on the usage of Linear Classifiers, more concretely neural networks.

To perform these both strategies, there had to exist previous databases with positive and negative tweets, which will be built in the initialization step of the sentiment analysis, as follows.

Sentiment Analysis Inicialization: In order to build these databases, 343,011 positive and 312,369 negative tweets were extracted (which are not necessarily in the political context, and which contain at least one emoticon). As in [24], a tweet that has positive emoticons is saved into the positive database and the tweet that contains negative emotions is saved into the negative database. Figs. 2 and 3 show the positive and negative emoticons used, respectively:

Fig. 2 Positive Emoticons

Fig. 3 Negative Emoticons

Previously to classify the tweets into positive or negative, they are processed in two senses: on the one hand, for the sake of enriching the information of the tweet text, the content of the URL appearing in the tweet text has been added, when it is applicable. When the content is a news headline, the title is

retrieved. When the content is a tweet, the complete tweet text is retrieved. This is done because many of the tweets had some URLs in the text and information concerning the meaning of the text would be lost if the URL content was excluded.

On the second hand, for the sake of uniformization, the tweets collected in both positive and negative databases, have to be treated according to a cleaning process based on the following steps:

1) Escape HTML characters: tweet text can contain html characters like "😄" and it is necessary to convert them to unicode, like "\uD83D\uDE04"
2) Decode Data: With the aim to make ready the tweet text to manage it, it is required to standardize and encode the text to unicode.
3) Remove users that contains simbol "@": Remove the users of the tweet because they are not relevant to the text sentiment analysis.
4) Remove hashtags: The hashtags are often related to a event context and not to a sentiment, for these reason they are not relevant to analyze the sentiment.
5) Slangs lookup: Since every language has slangs, in this part of the cleaning, slangs are mapped into words related to them.
6) Split joined sentences: To reduce the number of characters, Twitter users join the sentences using the capital letters in the words, an example: "ThisIsVeryDisgusting". In the cleaning process, they are split up.
7) Standardize words: To make "cool" tweets, some users write words emphasizing some letters, for example: "I am veeeeeery happy". This part of the process consists of standardizing these words. So in this case, the result would be "I am very happy".
8) Apply spell corrector Hunspell [26]: tweets are texts which have often been written in a rush and users do not take much care on the spelling. Hence, many misspelled words may be found. A spell corrector has been used to write these words correctly.
9) Remove the words with only one character: In relation with to the two strategies presented, it was considered that one-character words are not relevant and they could even introduce noise to the results.
10) Remove the double spaces: As a consequence of the cleaning process, some of the previous steps could have generated double spaces, which are removed.

Hence, after these processes, two collections of positive and negative tweets will be identified. These databases are used in both sentiment analysis strategies developed in this implementation. These strategies will take as inputs the target tweets, and its sentiment polarity will be obtained as an output.

Positive Negative Match (PNM): This strategy is based on counting positive and negative words in a tweet. Hence, it requires the previous creation of two databases with positive and negative words. To build up this words databases, the positive and negative tweets database defined earlier will be used.

The positive words database (negative, respectively) includes all the words that appear in the tweets belonging to the positive tweets database. In order to increase significance, the bigrams (couples of consecutive words) are also collected. The words/bigrams are collected along with their frequency of apperance. However, notice that many words will appear in both the positive and negative words database. In this case, they are removed from the database with less appearances, and its frequency is the difference between its frequencies in both sets. Following this procedure, two collections of 11,027 positive and 10,484 negative words/bigrams have been obtained.

Then, for each input tweet, after applying the previously described cleaning process, its polarity is obtained by counting the emoticons and the appearances of positive and negative words. The final result is the difference between the positive and negative words/emoticons. This difference is the polarity of the tweet, which can be positive ($difference > 1$), negative ($difference < 0$) or neutral ($difference == 0$).

As it will be discussed in the experimental results section, this procedure has been run for a test set of tweets, whose polarity is known beforehand. In particular, different experiments have been launched, using only some percentage of the most significant words (with maximum frequency) in the words databases. It turns out that the maximum assert has been reached when using the whole set of words.

Neural Network Sentiment (NNS): This strategy is based on the usage of a convolutional neural network which is trained using the positive and negative tweets database. Once the neural network is trained, it is ready to return the sentiment of a new tweet. In fact, the evaluation of one tweet returns the percentage of positive and negative.

In particular, the implementation is based on the project [25], which originally uses Tensorflow [27] to detect positive and negative English tweets. It is worth pointing out that this had to be modified to be adapted to the Spanish language. Notice that, since users may use irony when launching a tweet, the assert of the sentiment of a tweet based on its emoticons may be not be 100% correct. In order to deal with this, when the tweets are introduced to train the neural network, a weight percentage of positivity or negativity is assigned to each one. Note that, to analyze a tweet, the neuronal network needs to recognize all the word in the tweet text. So, when a specific word is not known, this word is removed from the tweet text and the updated tweet is again analyzed. If all the words were discarded, then the tweet would be classified as neutral.

2) Collection Classifier: This process receives political tweets along with its sentiment. As an output, it returns the list of political parties to which each tweet is related to, along with its sentiment. Hence, for each political party $x \in X$, the process looks for matchings between the keywords in C_x and the tweet text. Notice that one tweet can be classified as affine to one or more political parties, because matching keywords belonging to different political parties can be found.

As an upgrade, the overall system benefits from this step, by updating the C_x sets. Indeed, when some of the words appearing in a tweet, which is related to party $x \in C_x$, are detected as significant for this political party, they are assimilated to the set C_x. In this way, the increment of C_x

will improve the Data Classification process.

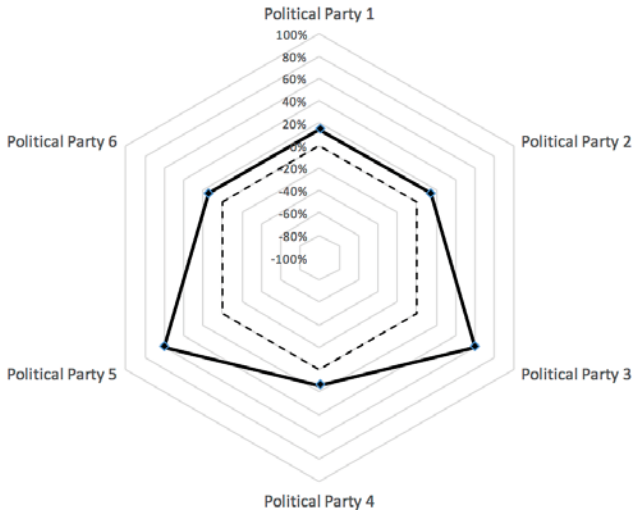

Fig. 4 Example User Political Tendency

C. Data Post-Processing

1) User Tendency Update: The final objective of the TAS strategy is to obtain the global political tendency. For such a purpose it is needed to analyze the tendency of each user. Hence, it is needed to get all the analyzed tweets for each user and identify the political party occurrences and its sentiment.

Let n be the number of political parties, and let u be an analyzed twitter user who has launched m analyzed tweets. Then, the data is arranged by means of an $m \times n$ matrix $M_u = (c_{ix})$, where c_{ix} is the coefficient referring to the tweet t_i launched by user u and related to the political party $x \in X$.

The content of each cell c_{ix} of the matrix can take four possible values:

$$c_{ix} = \begin{cases} 1, \text{ if } t_i \text{ has positive sentiment and mentions } x, \\ -1, \text{ if } t_i \text{ has negative sentiment and mentions } x, \\ 0, \text{ if } t_i \text{ has neutral sentiment and mentions } x, \\ \text{NULL, if } t_i \text{ does not mention } x, \end{cases}$$

Once, the data is collected in M_u, w_x is computed as the global weight sentiment of user u related to each political party x, as follows:

$$w_x = \sum_{i=1}^{m} c_{ix}, \forall c_{ix} \neq NULL. \tag{1}$$

On the other hand, let r_x be the number of citations of political party x from all the tweets that u has launched, that is

$$r_x = \#\{c_{ix} : \ 1 \leq i \leq m, \ c_{ix} \neq NULL\}. \tag{2}$$

The political affinity s_x for a specific user u and for each political party x, is calculated as the mean of sentiments, that is, as the ratio of tendency of political party x, w_x, to the number of appearances r_x, as follows

$$\mu_x = w_x \div r_x. \tag{3}$$

For each user and each political party the standard deviation is also calculated,

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (c_{ix} - \mu_x)^2}, \tag{4}$$

where n stands for r_x, or if there are less than 30 tweets, n stands for $r_x - 1$. Coefficient of variation is then calculated as the ratio of standard deviation and mean, where mean is not equal to zero,

$$\rho_x = \sigma_x \div |\mu_x|. \tag{5}$$

Coefficient of variation is used to explain how much a user's tweets vary in sentiment with regard to one political party; the less they vary, the more one is certain in the obtained average sentiment. Nevertheless, since mean can be negative, only the absolute value is taken into account. Furthermore, since mean is between -1 and 1, and it can be considerably smaller than standard deviation, in further text the interpretation of variability is adjusted with respect to the values that coefficient of variation can take, considering the mean value of coefficient of variation as the limit of small variability.

In order to clarify this process, a simple example is presented in Table I and Fig. 4, where it is assumed that there are 6 political parties ($n = 6$) and 10 tweets ($m = 10$) for a given user u. The last five rows of the table show the values w_x, r_x, μ_x, σ_x and ρ_x for each political party. Notice that, for this user, one can trust most to the tendency for the political parties 3 and 5.

Twitter Analyzed Users is one of the outputs of the TAS is a collection of analyzed users, which contains, for each user, all the information retrieved from twitter plus the data analyzed by TAS strategy, that are the sentiment for each political party, the number of tweets related to each political party and the result of the TAS strategy.

When the user political tendency has been calculated it is refreshed for each user and sent to the Global political Tendencies Update.

2) Global Political Tendencies Update: The global political tendencies update process computes the global political tendency of the Spanish twitter users. This process is executed each time a user is analyzed and the output is the global political tendency that contains the sentiment for each political party, the number of tweets related to each political party and the result of the TAS strategy. The process to calculate the global political tendency is calculated as the mean of users' tendencies.

IV. Experimental Results

The aim of this section is to illustrate the results obtained from the implementation of TAS, that is the sentiment analysis, user tendency and global tendency. The sentiment analysis results are composed of the two strategies presented in the TAS section: NNS strategy block presents results of the neural

TABLE I
CALCULATION EXAMPLE OF A WEIGHT MATRIX M_u FOR A GIVEN TWITTER USER u

Test User	Political Party 1	Political Party 2	Political Party 3	Political Party 4	Political Party 5	Political Party 6
t1	1	1	1	1	1	1
t2	1	NULL	1	NULL	1	NULL
t3	NULL	NULL	NULL	NULL	-1	NULL
t4	1	1	1	1	NULL	1
t5	-1	-1	NULL	-1	NULL	-1
t6	1	1	NULL	1	1	1
t7	-1	NULL	-1	NULL	NULL	NULL
t8	-1	-1	NULL	-1	NULL	-1
t9	NULL	-1	NULL	-1	NULL	-1
t10	NULL	1	1	1	1	1
w_x	1	1	3	1	3	1
r_x	7	7	5	7	5	7
_x	0,1429	0,1429	0,6000	0,1429	0,6000	0,1429
_x	1,0690	1,0690	0,8944	1,0690	0,8944	1,0690
_x	7,48	7,48	1,49	7,48	1,49	7,48

network using training, validation, and test datasets. PNM strategy block presents results of the PNM using training and the test data sets. The user tendency block results presents the experimentation and results of the user tendency calculation.

The results have been obtained using Storm Cluster [28] in OpenNebula [29]. The cluster is composed of 4 virtual computers with 4 processors and 4Gb of RAM each. Using this environment the ratio of tweets retrieved and analyzed are 5 tweets per second.

A. Sentiment Analysis

This section describes the results and discussion of NNS and PNM. The datasets used in the experiment is a data set with 312369 positive and 343011 negative tweets.

1) NNS: In order to obtain relevant results using neural network, the data set has been divided into training(70%), validation(20%) and test(10%) data. Training dataset is used to adjust the weights of neural network, validation dataset serves to further explore if additional data is improving the accuracy (or at least not decreasing it), and the test dataset is used to present sample results of the trained neural network. The output of the neural network is the percentage of positive and negative sentiment of a tweet. During the process of analyzing a tweet, if the neural network cannot recognize a word, the word is removed and the cleansed text is resent to the neural network. This process is repeated as long as there are unknown words in the input tweets, and if all the words are removed, the tweet is considered neutral. Results of the neural network training and validation has been the following:

Train set size = 458766
Test set size = 65538
Epoch 1, validation accuracy: 0.667114
Epoch 2, validation accuracy: 0.69687
Epoch 3, validation accuracy: **0.719812**

The evolution of the epoch confirms that the training data set provides a great training assert to the neural network. The best assert achieved is 71.98%. This can be considered as a good result compared to the results obtained in [1], which are also related to the analysis of Spanish tweets sentiment and they achieve an assert of 70%, over a significantly smaller dataset.

2) PNM: In order to obtain relevant results using PNM strategy, the data set has been divided into training (70%) and test (30%) data. Training dataset is used to extract the positive words (11027) and the negative words (10484) databases, while the test dataset is used to present sample results of the matching of the positive and negative words. The output of the PNM is the number of matches for each positive and negative word and the positive and negative emoticons in an analyzed tweet.

In the experimentation PNM has been tested with different sets of positive and negative words, in order to analyze its impact on the final result and obtain maximum assert. More concretely, we have run experimentation starting with the 10% of the most frequent words, and then increasing this percentage until using the whole database.

TABLE II
PNM RESULTS

Database %	Succes %	Error %	Neutral %
10	42	19	38
20	45	20	34
30	47	20	31
40	49	20	29
50	50	20	29
60	50	20	28
70	51	20	27
80	51	20	27
90	52	21	26
100	53	21	25

Table II shows the results of the different percentages used to check PNM, and the best result is achieved when the 100% of the words database is used. However, the best result obtained is 53% assert.

Taking into account the experimental results obtained, the NNS has been the selected strategy to analyze the users tweets sentiment.

B. TAS User Tendency

This subsection, presents a comparison between the political tendency obtained with the TAS Strategy in front of the real tendency of a sample sets of users, which has been used as a control sample and whose real tendency has been obtained manually. This sample sets of users is referred to as Tendency

TABLE III
TAS USER TENDENCY RESULTS

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Minimal number of tweets per user	25	20	10	20	20	20
Minimal number of tweets per user per political party	7	4	4	4	4	4
Maximal coefficient of variation	1,20	1,20	1,20	1,80	2,00	0,47
Assert	55,88%	60,00%	58,57%	59,65%	57,86%	71,43%
Error	23,53%	21,54%	21,43%	22,81%	22,86%	0,00%
Neutral	20,59%	18,46%	20,00%	17,54%	19,29%	28,57%
Users Number	15	28	30	43	51	3

of the Users Analyzed set (TEN). Manually analyzing the tendency of this set of users has been a challenge because this task is not straightforward, given that ordinary users post few political tweets and even some of them may contain irony, images or neutral text. The way for analyzing the profiles has been based on identifying tweets supporting a political party or tweets against a political party. Also, when an user has retweeted a tweet launched by a politician, it has been counted as a positive tendency to this political party. Following this procedure, a set of 184 random users has been considered, but it was reduced to 112 after discarding those accounts related to non Spanish users, prank accounts or TV programs. Next step is to compare the users political tendency obtained with the TAS Strategy, in relation to the results of TEN.

Table III shows the results of 6 different tests performed over the set of 112 TEN users. From each test, a particular subset of users has been selected in order to evaluate the success of the TAS strategy over significant users. These subsets have been chosen according to three different sieve parameters: the minimal number of tweets per user, the minimal number of citations per political party per user r_x and the maximal value of correlation coefficient ρ_x. The goal of the experimentation is to adequately tune these parameters in order to achieve maximal assert, minimal error and maximal number of users.

The results presented in Table III show that the largest assert is achieved with the smallest coefficient of variation (column Test 6) because the expressed opinion tendency varies the least in all analyzed tweets. Nevertheless this restriction results in removal of huge number of users, so in this case it was able to analyze only three users.

In order to perform less restrictive experiments, Tests 1, 2 and 3 have been performed with the maximal coefficient of variation of 1,20, which is the average of the coefficient of variation of all the political parties of all the users. Among these three experiments, the best assert is achieved in Test 2, with minimum of 20 tweets per user, at least 4 tweet for political party. The amount of users analyzed in this case has increased until 28, while the assert has been 60%. So, the following tests have been designed with the aim to maintain this assert while increasing the amount of users. This is shown in Tests 4 and 5, where the maximal coefficient of variation has been relaxed. Results show that Test 4 achieves the best results given that it obtains an assert close to 60% and analyses the maximum number of users (43).

Table IV presents a sample of an analyzed user applying the criteria of the Test 4, who has posted 20 tweets. In this case it can be seen that the political parties significant for this user are ERC and Podemos, since the coefficient of variation for these parties satisfies $\rho_x \leq 1.8$ and the number of citations of each political party are $r_x \geq 4$. Likewise the political party PSOE could be another candidate, because the $\rho_x \leq 1.8$, but the number of citations of political party is $r_x < 4$, and for this reason this political significance for this user is discarded. Hence, in this sample one can conclude that this user has a negative tendency to the political party ERC and a positive tendency to the political party Podemos.

TABLE IV
SAMPLE USER TENDENCY ANALYZED

User	PP	PSOE	CS	ERC	Podemos	PDECAT
t1	1	NULL	1	NULL	1	1
t2	1	NULL	1	NULL	1	NULL
t3	NULL	0	0	NULL	NULL	NULL
t4	-1	NULL	NULL	-1	NULL	-1
t5	NULL	NULL	1	NULL	1	1
t6	NULL	NULL	NULL	NULL	NULL	NULL
t7	NULL	NULL	-1	-1	NULL	-1
t8	1	NULL	1	NULL	NULL	1
t9	NULL	NULL	-1	-1	NULL	-1
t10	1	NULL	1	NULL	NULL	1
t11	-1	-1	-1	-1	NULL	-1
t12	NULL	-1	NULL	-1	NULL	-1
t13	-1	NULL	-1	-1	NULL	-1
t14	1	NULL	1	NULL	NULL	1
t15	-1	NULL	NULL	-1	NULL	-1
t16	1	NULL	NULL	1	1	1
t17	NULL	NULL	1	1	1	1
t18	-1	NULL	-1	-1	-1	-1
t19	1	NULL	1	NULL	NULL	1
t20	1	NULL	NULL	NULL	NULL	1
w_x	3	-2	3	-6	4	1
r_x	13	3	14	10	6	17
_x	0,2308	-0,6667	0,2143	-0,6000	0,6667	0,0588
_x	1,0127	0,5774	0,9750	0,8433	0,8165	1,0290
_x	4,39	0,87	4,55	1,41	1,22	17,49

Fig. 5 presents the global political tendency for the 43 users analyzed with respect to the criteria used in Test 4. It turns out that for these users all political parties have average negative global tendencies, so the best tendency can be considered the one that has the highest mean. In this case, political party PP has the best global tendency. Furthermore, the results of Test 4 shows that political party PDECAT is not significant for the tested users. So, this political party has not been reflected in the Global Tendency Analyzed.

V. CONCLUSION

This paper proposes a new strategy, named Tweets Analysis Strategy (TAS), to analyze the Spanish political users tweets by means of discovering its sentiment (positive, negative or neutral) and classifying them according to the political party

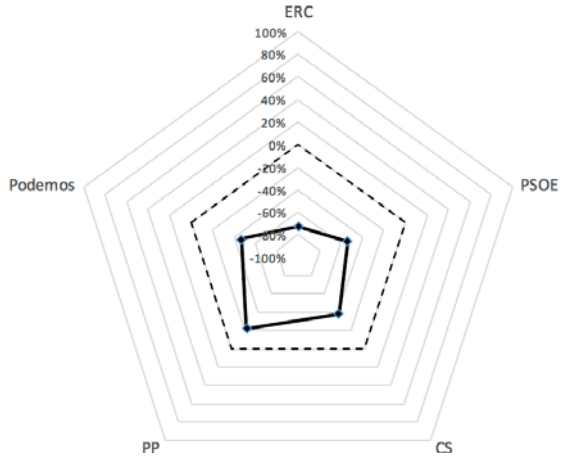

Fig. 5 Sample Global Tendency Analyzed

they support. From this individual political tendency, the global political prediction for each political party is calculated.

In order to do this, two different strategies for analyzing the sentiment analysis are proposed: one is based on Positive and Negative words Matching (PNM) and the second one is based on a Neural Networks Strategy (NNS). Both strategies have been evaluated with a set of 655380 tweets, previously processed by a cleaning process, obtaining an accuracy of 53% and 71.98% for the PNM and NNS strategy, respectively. It is worth pointing out that NNS provides the best results because it takes into account the relationships between the words, while the PNM only matches the positive and negative words.

With the aim of determining the political tendency for each user, the TAS strategy needs to define the threshold of some different parameters, which were identified by means of experimentation. The most relevant thresholds identified for each political party and user are: the number of tweets greater than 20, the number of citations $r_x \geq 4$ for each political party and the coefficient of variation $\rho_x \leq 1.8$. In this case, the error obtained has been 22.81%. It turns out that coefficient of variation can be an useful indicator when applied to one user's political tendency expressed in tweets, in the sense that if user's tendency towards political party does not vary much throughout tweets, one can assume that this is their reliable tendency. The strategy based on the available number of tweets per user, per political party, as well as the number of analyzed users would allow to decide if it is feasible to exclude tweets with medium and large variation in sentiment or to exclude only those with large variation.

The future work is directed towards extending our analysis about discovering and modeling the relationships between users. In this way, the integration of the content analysis, developed in this paper, with the communication networks of the users, can improve significantly the accuracy of the prediction for each user. Likewise, we plan to apply the TAS strategy method to discover the tweets sent by bots, in order to determine the level of intrusion in the political campaigns.

Acknowledgments

This work has been supported by the MEI under contract TIN2014-53234-C2-2-R, TIN2017-84553-C2-2-R and MTM2017-83271-R.

References

[1] M. Cmara, G. Cumbreras, V. Romn, and G. Morera, "TASS 2015 The Evolution of the Spanish Opinion Mining Systems" *Procesamiento del Lenguaje Natural*, vol.56, 2016.

[2] J. Schmidhuber. 2015. Deep learning in neural networks. Neural Netw. 61, C (January 2015), 85-117. DOI=http://dx.doi.org/10.1016/j.neunet.2014.09.003

[3] B. Wellman, A. Quan Haase, J. Witte and K. Hampton. "Does the Internet Increase, Decrease, or Supplement Social Capital?: Social Networks, Participation, and Community Commitment", *American Behavioral Scientist*, vol.45, n3, 2014, pp.436–455.

[4] Statista. "Number of monthly Active Twitter Users Worldwide from 1st quarter 2010 to 3rd quarter 2017 (in millions)", https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/, Date accessed: 01/12/2017.

[5] K. Lee, D. Palsetia, R. Narayanan, Md. Mostofa Ali Patwary, A. Agrawal and A. Choudhary. "Twitter Trending Topic Classification". *In Proceedings of the IEEE 11th International Conference on Data Mining Workshops (ICDMW '11)*. IEEE Computer Society, 2011 pp.251–258.

[6] Beevolve. "An Exhaustive Study of Twitter Users Across the World", http://temp.beevolve.com/twitter-statistics/, 2012, Date accessed: 01/12/2017.

[7] M.M. Uddin, M. Imran and H. Sajjad. "Understanding Types of Users on Twitter", *ArXiv e-prints*, abs/1406.1335, 2014.

[8] L. De Silva and E. Riloff. "User Type Classification of tweets with Implications for Event Recognition". *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 2014, pp.98–108.

[9] N. Mangal, R. Niyogi and A. Milani. "Analysis of Users' Interest Based on tweets", *16th International Conference on Computational Science and Its Applications (ICCSA)*, 2016.

[10] M. Pennacchiotti and A. Popescu. "A Machine Learning Approach to Twitter User Classification", *A Machine Learning Approach to Twitter User Classification*, 2011, pp.281–288.

[11] EJ. Lee, SY. Shin. "Are They Talking to Me? Cognitive and Affective Effects of Interactivity in Politicians' Twitter Communication", *Cyberpsychology, behavior and social networking*, vol.15, n.10, 2012, pp.515–520.

[12] DS. Hillygus. The Evolution of Election Polling in the United States, *The Public Opinion Quarterly*, vol.75, n.5, 2011, pp. 962-981.

[13] La Vanguardia. "Bad Results in the Forecast of the Elections", http://www.lavanguardia.com/politica/elecciones/20160627/402794314538/sondeos-elecciones-generales-26j-fallo.html , June 27, 2016, Date accessed: 01/12/2017 (Spanish version).

[14] W. Wanga, D. Rothschil, S. Goel, A. Gelman. "Forecasting Elections with Non-representative Polls", *International Journal of Forecasting*, Vol.31, Issue 3, 2015, pp. 980-991.

[15] D. Preotiuc-Pietro, Y. Liu, D. Hopkins, L. Ungar. "Beyond Binary Labels: political Ideology Prediction of Twitter Users", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp.729–740.

[16] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. "Predicting Elections with Twitter: What 140 Characters Reveal about political Sentiment", *International AAAI Conference on Weblog and Social Media*, 2010, pp.178–185.

[17] Preotiuc-Pietro, D.; Volkova, S.; Lampos, V.; Bachrach, Y. and Aletras, N. Studying User Income through Language, Behaviour and Affect in Social Media. PLoS ONE,10(9), 2015.

[18] K. Sylwester and M. Purver. "Twitter Language Use Reflects Psychological Differences between Democrats and Republicans", *PLoS ONE*, vol.10, n.9, 2015.

[19] MD. Conover, B. Gonalves, J. Ratkiewicz, A. Flammini and F. Menczer. "Predicting the political Alignment of Twitter Users", *IEEE Third International Conference on Social Computing (SocialCom)*, 2011, pp:192–199.

[20] Twitter Developers. "Twitter API", https://dev.twitter.com/, Date accessed: 01/12/2017.

[21] L. Jiang, M. Yu, M. Zhou, X. Liu and T. Zhao. "Target-dependent Twitter sentiment classification", *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp.151–160.

[22] Y. Wang, Y. Rao, X. Zhan, H. Chen, M. Luo and J. Yin. "Sentiment and Emotion Classification over Noisy Labels", *Know.-Based Syst.*, 2016, pp.207–216.

[23] W. Medhat, A.Hassan and H. Korashy. "Sentiment Analysis Algorithms and Applications: A survey", *In Ain Shams Engineering Journal*, Vol.5, Issue 4, 2014, pp.1093–1113.

[24] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau. "Sentiment Analysis of Twitter Data", *In Proceedings of the Workshop on Languages in Social Media (LSM '11)*, 2011, pp.30–38.

[25] D. Grattarola. "Twitter Sentiment Classification", https://github.com/danielegrattarola/twitter-sentiment-cnn, Date accessed: 1/12/2017.

[26] Hunspell. Hunspell is the spell checker, http://hunspell.github.io/, 2017, Date accessed: 14/12/2017.

[27] TensorFlow. TensorFlow is an open source software library for numerical computation using data flow graphs, https://www.tensorflow.org/, 2017, Date accessed: 14/12/2017.

[28] Apache Storm. Apache Storm is a free and open source distributed realtime computation system, http://storm.apache.org/, 2017, Date accessed: 14/12/2017.

[29] OpenNebula. OpenNebula is a cloud computing platform for managing heterogeneous distributed data center infrastructures, https://opennebula.org/, 2017, Date accessed: 14/12/2017.