2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)

A Sentiment Analysis System of Spanish Tweets and Its Application in Colombia 2014 Presidential Election

Jhon Adrián Cerón-Guzmán and Elizabeth León-Guzmán
Departamento de Ingeniería de Sistemas e Industrial
Universidad Nacional de Colombia
Bogotá D.C., Colombia
{jacerong,eleonguz}@unal.edu.co

***Abstract*—What people say on social media has turned into a rich source of information to understand social behavior. Sentiment analysis of Twitter data has been widely used to capture trends in public opinion regarding important events such as political elections. However, current research in social media analysis in political domains faces two major problems, namely: sentiment analysis methods implemented are often too simple, and most of the works assume that all users and their tweets are trustworthy. This research is aimed at dealing with these problems to achieve more reliable public opinion measurements. First, a dataset of 513K tweets referring to Colombia 2014 presidential election was collected. To distinguish spammer accounts from non-spammer ones, a supervised learning technique was implemented on a labeled collection of users. Next, a sentiment analysis system was developed by following a supervised classification approach. Lastly, the system was applied in the Colombian election to investigate the potential of social media for voting intention inference. Experimental results show that inference methods based on Twitter data are not consistent, despite obtaining the lowest mean absolute error and correctly ranking the highest-polling candidates in the first round election with the proposed inference method.**

I. Introduction

In an increasingly connected world taking advantage of what people say about factual or subjective issues might bring gains not only in the economic and political arena, but also in the social one. However, finding and monitoring such information is a formidable task due to the large amount of user-generated content that is spread on the web [1]. And, not least, language diversity in the web [2] becomes a major issue to be considered.

Social media platforms such as Facebook and Twitter have led to deep changes in the paradigm of information generation and consumption. For example, real-time Twitter content on natural disasters has been exploited to support disaster management teams [3]. Twitter is nowadays a popular microblogging site where users receive and exchange information instantaneously with a global audience; this is, users are not limited to their friendship networks, as it happens in Facebook. 'Tweeting', therefore, has become an activity *par excellence* to say what one thinks or feels, because of brevity of tweets and the widespread use of mobile devices [4].

What people say on social media about issues of their everyday life, the society, and the world in general has turned into a rich source of information to understand social behavior [5]. This large amount of user-generated content has brought new research opportunities to explore the human subjectivity at large scale, which was not feasible using traditional methods. However, analyzing this content also presents several challenges, including: distinguishing noisy, useless, and irrelevant information from valuable data; and developing text analysis approaches based on Natural Language Processing (NLP) techniques, which properly adapt to the informal genre and the free writing style of these platforms. Addressing these challenges would lead to more reliable results, because new forms of spam have been spread to manipulate social media discourse [6] and the performance of traditional NLP tools degrades on social media data [7], [8].

An appealing application of social media analysis is to determine the opinion orientation expressed in text streams. Sentiment analysis or opinion mining, as this application is known, deals with the task of rating the opinion orientation as either positive, negative, or neutral [1]. This computational approach has been implemented in a diversity of domains ranging from marketing to politics [5]. The latter has caught the attention of researchers, whom have investigated the feasibility of supplementing or substituting traditional electoral polling with sentiment classification of text streams [9], [10], [11]. Despite the relative success reported in the literature, the following problems have been identified [12], [10]: most of the works implemented the simplest of sentiment analysis methods, whose performance is only slightly better than that of a random classifier; and they have assumed that all users and their tweets are trustworthy. These problems, therefore, need to be tackled in order to achieve more reliable public opinion measurements from social media data.

This research deals with the challenges and problems described above. Specifically, spammer detection and a sentiment analysis system of Spanish political tweets, which implements state-of-the-art approaches and adapts to the informal genre and the free writing style of Twitter, are the main contributions of this work in the search for more reliable public opinion

978-1-5090-3936-4/16 $31.00 © 2016 IEEE
DOI 10.1109/BDCloud-SocialCom-SustainCom.2016.47

measurements from Twitter data regarding a political election. Furthermore, the potential of social media to infer voting intention is investigated. Colombia 2014 presidential election was proposed as case study. The above challenges are not, however, the only open ones [10]. Figurative language and demographic characterization of Twitter users, among others, are major issues to be addressed in further research.

This paper is organized as follows. Section II describes previous work on sentiment analysis of Twitter data and discusses research focused specifically on political domain. Next, a brief background on the Colombian election is presented in Section III. In Section IV the datasets used are described. Then, the sentiment analysis system, as well as the voting intention inference, are discussed in Sections V and VI, respectively. Finally, Section VII concludes the paper.

II. Related Work

A. Sentiment Analysis of Twitter Data

Sentiment classification is not a recent task. To the best of our knowledge, the seminal works on sentiment analysis were carried out by Pang et al. [13] and Turney [14]. They proposed the approaches typically used in sentiment analysis, whatever the source of the textual data: classification based on unsupervised learning [14] and classification based on supervised learning [13]. The former uses a lexicon to filter words of known polarity in a document in order to assign it a label class. Although this approach is appealing due to its simplicity and the ease of implementing it, it is not able to understand subtle expressions (e.g., sarcasm) and the different meanings that a same word may acquire in nonidentical domains [1]. Instead, state-of-the-art in sentiment analysis of tweets follows the supervised classification approach [15], [16], [17]. Below, a literature review on sentiment analysis of Twitter data is presented.

Mohammad et al. [15] used a Support Vector Machine (SVM) with a large number of features, which are grouped into: word ngrams, character ngrams, all-caps, part-of-speech (POS), hashtags, lexicons, punctuation, emoticons, elongated words, clusters, and negation. To classify a tweet, they first normalized it by replacing URLs and user mentions by placeholders, and then tokenized and POS tagged it. The vectorization [18] was based on the bag-of-words (BOW) representation.

Miura et al. [19] developed a sentiment analyzer based on supervised text classification. They used the Logistic Regression algorithm to predict the label class of a tweet; however, because the class distribution was unbalanced, they introduced an weighting factor w_l to adjust a probability output $Pr(l)$ of class l, and thus the class with the highest updated probability was chosen. The groups of features was inspired by Mohammad et al. [15], namely: word ngrams, character ngrams, lexicons, clusters, and word senses. They also used a spelling corrector and a word sense disambiguator, which were applied in the text preprocessing.

Amir et al. [20] proposed the following three groups of features: word-based, lexicon, and syntactic. In order to compute the word-based features, they used, in addition to the BOW representation, the word2vec method [21]. Under this method, neural networks are trained to learn vector representation of words from a (commonly) large dataset; this vector representation is characterized by full density and low dimensionality. To assign a class label to a tweet, they implemented the Logistic Regression algorithm. Class weights set to be inversely proportional to the class distribution were introduced.

Hagen et al. [16] reproduced four state-of-the-art approaches to sentiment analysis in Twitter and combined them in an ensemble. The ensemble combination was not based on the final decision of each approach (reimplemented as a classifier), but rather it requested the classifiers' probabilities for each class. Thus, the class with the highest average probability was chosen.

Saralegi and Vicente [22] developed a supervised system using three groups of features to support the classification of Spanish tweets. In order to transform the tweet text into a feature vector, they used the BOW representation filtered by the words of a polarity lexicon, in addition to the frequency of each POS tag and the frequency of each emoticon and interjection type (positive and negative). Because most of the lexicons exist for English, they created a polarity lexicon for Spanish by semi-automatically translating an English polarity lexicon and automatically extracting the words most associated with a certain polarity from a training corpus.

Finally, Díaz-Galiano and Montejo-Ráez [23] used the word2vec and doc2vec methods for vector representation [21], [24]. Doc2vec, unlike word2vec, induces a vector representation for each paragraph. In order to represent a tweet as a feature vector, they concatenated the vector obtained by the doc2vec and the vector as the average of the word2vec vectors. In this way, a 500-dimensional vector fed a SVM to assign a class label to a tweet.

B. Predicting Voting Intention from Twitter Data

Predicting real-world events from social media data has turned into an appealing line of research from social sciences to computer science [5]. What people say about an electoral race or its contestants has been exploited to predict or forecast election outcomes, given the large amount of user-generated content by the ever-growing virtualization of human behavior. In this way, a new alternative to gauge public opinion has been developed, which also benefits from the increasing cost and difficulty of the traditional opinion polls [25]. However, the numerous researches that claim to have successfully forecasted, face reproducibility problems [10]. More importantly, most of the works dealing with election outcome prediction were only post hoc analysis [12].

Tumasjan et al. [26] claimed that the proportion of tweets mentioning an electoral option can be considered as a plausible reflection of its voting share. They reported a very low error in forecasting the 2009 German Bundestag elections on the assumption that the larger number of tweets, the larger the vote. However, despite being aware of the low representativeness, such that a small number of users generated most of the tweets, they claimed that Twitter is a predictor of election outcomes. This proves that Twitter's user base is not a representative sample of the population [10]. Jungherr et al. [27] reproduced the research, finding that the apparent success was due to data manipulation on the part of researchers.

O'Connor et al. [9] correlated sentiment scores with opinion polls in order to determine if sentiment classification would respond faster to changes in the consumer confidence or the presidential job approval, compared with the traditional opinion polling. They defined the sentiment score to be the ratio between the number of positive tweets and the number of negative tweets; tweets were labeled by a lexicon-based classifier. Based on the obtained results, they claimed that sentiment analysis is a substitute and supplement for the traditional polling. However, Metaxas et al. [12] concluded that a lexicon-based classifier wrongly interprets the subtleties of propaganda and disinformation, and that its performance is only slightly better than that of a random classifier.

In the same way that social media provides a rich source of information, it could, however, contain noisy, useless, and irrelevant information. In this regard, Metaxas et al. [12] warned: "spammers and propagandists write programs that create lots of fake accounts and use them to tweet intensively, amplifying their message, and polluting the data for any observer." Those problems, therefore, need to be tackled in order to achieve reliable public opinion measurements.

Gayo-Avello [10] presented a comprehensive literature revision on electoral prediction from Twitter data. He concluded with recommendations for future research from which the following are highlighted: the state-of-the-art approaches to sentiment analysis in Twitter should be implemented, and spam and disinformation should be removed from the study data. These recommendations have greatly inspired this research.

Shi et al. [28] and Tsakalidis et al. [11] developed prediction models that did not strictly rely on sentiment analysis or Twitter volume. Shi et al. [28] correlated a set of 19 features with opinion polls in order to predict the Republican Party presidential primaries in 2012. They also claimed that the traditional electoral polling can be supplemented or supplanted with analysis of Twitter data. Tsakalidis et al. [11] developed regression models to predict elections for multiple countries. Their results in most of the cases were better than those of the poll-based prediction, although they used a lexicon-based classifier.

III. Background on the Colombian Election

In race for the presidency in 2014 five candidates competed for the most important Colombian political office, including the incumbent President Juan Manuel Santos. Óscar Iván Zuluaga, Marta Lucía Ramírez, Clara López, and Enrique Peñolosa were the other candidates.

The presidential election was held under a two-round voting system. In the first round, held on May 25, 2014, no candidate received an absolute voting majority, and for that a run-off took place 21 days later between Zuluaga and Santos, whom were the highest-polling candidates with 29.28% and 25.72% support from voters, respectively. In the run-off election, Santos was re-elected President with 50.98% support.

Table I
Summary of the collected Twitter data on the electoral process

Candidate	Collection Period	Tweets	Users
Santos	Apr 30–Jun 24, 2014	332,575	117,783
Zuluaga	Apr 30–Jun 24, 2014	202,405	81,979
Ramírez	Apr 30–May 29, 2014	9,273	6,198
López	Apr 30–May 29, 2014	13,711	9,457
Peñalosa	Apr 30–May 29, 2014	12,072	7,391
Blank vote	Apr 30–Jun 24, 2014	39,203	27,148

IV. Datasets

The datasets are characterized by the sources from which they were collected, namely: Twitter and opinion polls. In order to enable reproducibility, the datasets are made publicly available at http://dx.doi.org/10.5281/zenodo.58435.

A. Twitter Data

During the course of the presidential election, in a two-month period from April 30, 2014 to June 24, 2014, a dataset ("COpres14") of 513,324 tweets contributed by 149,831 different users was collected from Twitter Search API. To conduct the research relying on tweets referring to the aforementioned political context, a set of criteria was defined to filter them. Thus, only tweets containing at least one keyword or hashtag related to the presidential election (i.e., *elecciones* (`election`), *presidenciales* (`presidential`), *#Elecciones2014* (`#2014Election`), *#ColombiaElige* (`#ColombiaChooses`), *#EleccionesColombia* (`#ColombianElection`), and *#ColombiaDecide* (`#ColombiaDecides`)) and full name or user mention that identifies a given candidate, were collected. Table I shows the amount of collected data in terms of users and tweets per candidate. Note that a larger amount of data was collected for the candidates Santos and Zuluaga, as well as for the *blank vote* option, because they were the contestants in the run-off election.

1) Spammer Detection: Because not enough tweets were collected per user, up to the 40 most recent tweets were crawled from each user timeline using Twitter User Timeline API. In this way, a dataset of 134,625 users and 1,765,225 tweets was collected. In order to classify a collection of users into spammer and non-spammer, three methods commonly used in the literature to create a ground truth for spammer detection were applied: harmful links in tweets were detected by automatically checking them against five URL blacklists, and thus the users who spread them were thoroughly analyzed to identify spammers and not misclassify non-spammers; suspended accounts by Twitter in the dataset were used as spammer instances; and a random sample of 1,245 users was manually labeled by considering the set of criteria proposed by Chu et al. [29]. As a result, a labeled collection of 3,455 users was created, including 2,660 spammers and 795 non-spammers.

On the other hand, 30 features were proposed to support the discriminative power in spammer detection, such as user mention ratio, age of the user account, number of tweets from manual and automated devices, retweet rate, and number of followers; these are computed from tweet text and metadata. Then, a Twitter user classification system was developed by implementing the Random Forest algorithm on the ground truth, using the features to support the classification. The system was trained on a training set (66% of samples in the ground truth) via cross validation, and its evaluation was performed on a test set (34% of samples in the ground truth). Experimental results [30] show that the system achieved an overall accuracy of 92.62%, with a true positive rate (spammers classified as spammers) of 93.01% and a false positive rate (non-spammers misclassified as spammers) of 7.78%.

As a result of applying the Twitter user classification system on the COpres14 set, 22.01% of users were classified as spammers, whom generated 15.67% of tweets. These findings prove that spammers could significantly affect measurements based on Twitter data, and thus, it can be argued that the mere number of tweets is not a reliable source of voting preferences. Accordingly, spammers and their tweets were removed from the dataset. Extensive research on spammer detection can be read in [30].

2) Sentiment Labeled Dataset: A sentiment analysis system is highly sensitive to the domain from which the data used to train it were extracted [1]. For this reason, a system may obtain poor results when it is applied on a dataset whose domain differs from the one learned [31]. Although an important and large resource exists to build sentiment analysis systems of Spanish tweets [32], it was decided to create a dataset by labeling a random sample of tweets drawn from the COpres14 set, because the context of extraction of the cited resource has a Spain-focused bias and its domain deals with topics of general interest from politics to celebrities; instead, a dataset whose domain exclusively deals with the topic of interest of this research, results more appropriate because the sentiment analysis system will be applied on the COpres14 set to infer voting intention in the Colombian election.

A random sample of 1,170 tweets was drawn from the COpres14 set. In order to label a tweet as either positive, negative, or neutral, two volunteers assigned it a label according to the sentiment they understood was conveyed.[1] If there was no agreement among volunteers regarding the polarity label of a tweet, a third independent volunteer was heard. Volunteers agreed in 40.38% of tweets, thus supporting the statement with respect to humans often disagree on the sentiment of a text [32]. In total, 1,030 tweets were labeled by 234 different volunteers, each of whom labeled 10 tweets. The class distribution is as follows: positive, 22.43%; negative, 41.10%; and neutral, 36.47%.

Table II
OPINION POLLS TO GAUGE VOTING INTENTION IN THE FIRST ROUND ELECTION

Pollster	Survey period	
	Start date	End date
Infométrika [33]	May 03, 2014	May 06, 2014
Centro Nacional de Consultoría [34]	May 06, 2014	May 10, 2014
Cifras y Conceptos [35]	May 09, 2014	May 12, 2014
Datexco [36]	May 10, 2014	May 13, 2014
Gallup [37]	May 10, 2014	May 13, 2014
Ipsos [38]	May 13, 2014	May 15, 2014

Table III
OPINION POLLS TO GAUGE VOTING INTENTION IN THE RUN-OFF ELECTION

Pollster	Survey period	
	Start date	End date
Cifras y Conceptos [35]	May 26, 2014	May 27, 2014
Centro Nacional de Consultoría [39]	May 26, 2014	May 30, 2014
Datexco [40]	May 31, 2014	June 04, 2014
Gallup [41]	May 31, 2014	June 03, 2014
Cifras y Conceptos [35]	May 31, 2014	June 03, 2014
Ipsos [42]	June 02, 2014	June 04, 2014

B. Opinion Polls

Opinion polls were collected and then aggregated by hand. These were filtered by their survey period in order that they corresponded with the collection period of Twitter data. Thus, polls conducted between May 03, 2014 and May 15, 2014 were collected to infer voting in the first round election (as seen in Table II); likewise, those whose survey period was in the range from May 26, 2014 to June 04, 2014 were collected to infer voting in the run-off election (as seen in Table III).

In order to aggregate the data from the opinion polls, the approach proposed by Tsakalidis et al. [11] was applied as follows: because a poll is usually conducted in a two- or three-day period, the voting share each candidate would receive is treated as if the election would have taken place on any of these days. If two or more polls were conducted on a same day, the voting share of each candidate is considered as the weighted average value, using the sample size of every poll as the weight. Finally, the votes of undecided voters were proportionally distributed to all contenders.

V. The Sentiment Analysis System

A. The System Architecture

The tweet text is passed through the pipeline of the sentiment analysis system in order to label it as either positive, negative, or neutral. The pipeline, which goes from text preprocessing to machine learning classification, is shown in Fig. 1 and described below.

[1] A website was developed to help volunteers to manually label tweets.

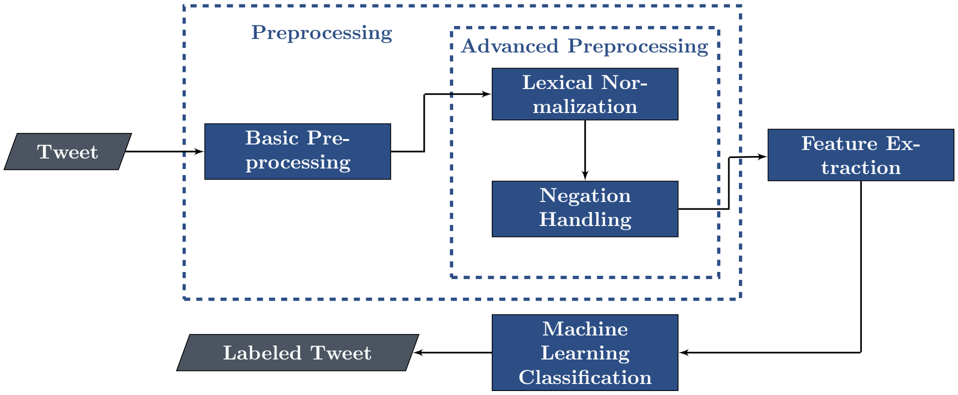

Figure 1. The system architecture

1) Preprocessing: The process of text cleaning and normalization is performed in two phases: basic preprocessing and advanced preprocessing.

Basic Preprocessing:

- Removing URLs and emails.
- HTML entities are mapped to textual representations (e.g., "&lt;" → "<").
- Specific Twitter terms such as mentions (@user) and hashtags (#topic) are replaced by placeholders.
- Unknown characters are mapped to their closest ASCII variant, using the Python *Unidecode* module for the mapping.
- Consecutive repetitions of a same character are reduced to one occurrence.
- Emoticons are recognized by simple regular expressions and classified into positive and negative, according to the sentiment they convey (e.g., ":)" → "EMO_POS", ":(" → "EMO_NEG").
- Unification of punctuation marks [43].

Advanced Preprocessing. Once the set of simple rules has been applied, the tweet text is tokenized and morphologically analyzed by FreeLing [44]. In this way, for each resulting token, its lemma and Part-of-Speech (POS) tag are assigned. Taking these data as input, the following advanced preprocessing is applied.

- **Lexical normalization.** Each token is passed through a set of basic modules of FreeLing (e.g., dictionary lookup, suffixes check, detection of numbers and dates, and named entity recognition) for identifying standard word forms and other valid constructions. If a token is not recognized by any of the modules, it is marked as out-of-vocabulary (OOV) word. Then, a confusion set is formed by normalization candidates which are identical or similar to the graphemes or phonemes that make the OOV word. These candidates are elements of the union of a dictionary of Spanish standard word forms and a gazetteer of proper nouns. The best normalization candidate for the OOV word is which best fits a statistical language model. The language model was estimated from the Spanish Wikipedia corpus. Lastly, the selected candidate is capitalized according to the capitalization rules of the Spanish language. Extensive research on lexical normalization of Spanish tweets can be read in [45].
- **Negation handling.** Inspired by the approach proposed by Pang et al. [13], this research defined a negated context as a segment of the tweet that starts with a (Spanish) negation word and ends with a punctuation mark (i.e., "!", ",", ":", "?", ".", ";"), but only the first token (from left to right) labeled with a specific POS tag (i.e., verb, adjective, or common noun) is affected by adding it the "_NEG" suffix. This definition was result of experimentation on the training set.

2) Feature Extraction: In this stage, the normalized tweet text is transformed into a feature vector that feeds the machine learning classifier. The features are grouped into basic features and word-based features.

Basic Features:[2]

- All-caps (1): the number of words completely in uppercase.
- Elongated words (1): the number of words with more than two consecutive repetitions of a same character.
- Punctuation marks (2): the number of consecutive repetitions of exclamation marks, question marks, and both punctuation marks (e.g., "!!", "??", "?!") and whether the text ends with an exclamation or question mark.
- Emoticons (3): the number of occurrences of each class of emoticons (i.e., positive and negative) and whether the last token of the tweet is an emoticon.
- Lexicons (3): the number of positive and negative words, relative to the ElhPolar lexicon [22]. In a negated context the label of a polarity word is inverted (i.e., positive words become negative words, and vice versa). Additionally, a third feature labels the tweet with the class whose number of polarity words in the text is the highest.

[2] Some of these features are computed before the process of text cleaning and normalization is performed.

Table IV
DISCRIMINATIVE POWER OF THE SYSTEM FOR EACH CLASS

Class	Precision	Recall	F1-score
Positive	0.65	0.43	0.52
Negative	0.62	0.74	0.67
Neutral	0.56	0.55	0.55

- Negation (1): the number of negated contexts.
- POS (13): the number of occurrences of each Part-of-Speech tag.

Word-based Features. The fixed-length set of basic features is always extracted from tweets. However, the tweet text varies from another in terms of length, number of tokens, and vocabulary used. For that reason, a process that transforms textual data into numerical feature vectors of fixed length is required. This process, known as vectorization, is performed by applying the tf-idf weighting scheme [18]. Thus, each document (i.e., a tweet text) is represented as a vector $d = \{t_1, \ldots, t_n\} \epsilon \mathbb{R}^V$, where V is the size of the vocabulary which was built by considering unigrams in the collection (i.e., the training set).

3) Machine Learning Classification: At the last stage, the sentiment analysis system classifies a given tweet as either positive, negative, or neutral. It is important to note that the system deals with a multiclass classification task, and therefore it assigns only one label to the tweet. After receiving as input the feature vector, a L2-regularized Logistic Regression classifier assigns a label class to the tweet. The classifier was trained on the training set via cross validation, using the Scikit-learn [46] implementation of the Logistic Regression algorithm.

B. Experiments

The sentiment labeled dataset was splitted into two sets: 80% of tweets were used as the training set and the remaining as the test set. The splitting of the data was performed in a stratified way. In terms of overall performance, the system achieved a macro-averaged F1-score of 58.24% and an accuracy of 60.19% on the test set.

Subsequently, the discriminative power of the system for each class was evaluated. The standard information retrieval metrics of precision, recall, and F1-score were used to perform the evaluation. Table IV shows the results.

As a final point, it is hypothesized that the low performance of the system is due to the process of creating the sentiment labeled dataset. Therefore, it is proposed as future work to train and evaluate a system on the general corpus provided by the organizing committee of the TASS workshop [32] in order to evaluate the previous hypothesis. As a result, it might be determined that a labeling process where a small number of volunteers participate, and it is also assisted by a classifier [32], produces a subjectivity easier to learn by the system, instead of a process where a large number of volunteers are involved, thus producing a subjectivity from different tendencies which is harder to learn.

VI. VOTING INTENTION INFERENCE IN THE COLOMBIAN ELECTION

A. Features and Method

A common denominator in the literature on voting intention inference from Twitter data is either treat the proportion of tweets mentioning an electoral option as the reflection of its voting share [26], or employ the simplest of sentiment analysis methods (e.g., a lexicon-based classifier) and assume that the candidate with the highest sentiment score would result to be the chosen [12]. However, both inference methods have proven to be inconsistent [10]. Therefore, the feature selection has followed recommendations discussed in [12], [10] to deal with these problems.

1) Features: The following features, which are the independent variables used by the inference method, are computed from the COpres14 set in a daily basis per candidate to correlate them with the polling data.

1) Tweet volume: the number of tweets mentioning candidate c on day d.
2) Unique tweet volume: the number of tweets that only mentions candidate c on day d.
3) Twitter user count: the number of different Twitter users with at least one tweet mentioning candidate c on day d.
4) Unique Twitter user count: the number of different Twitter users whose tweets only mention candidate c on day d.
5) Positive (negative) tweet volume: the number of positive (negative) tweets that mentions candidate c on day d.
6) Positive- (negative-) based Twitter user count: the number of different Twitter users with at least one positive (negative) tweet mentioning candidate c on day d.

In total, 14 features were proposed by additionally taking into account the *sentiment score* [9] and other ratios such as *tweets per user*. Tweets were classified by the sentiment analysis system to compute the sentiment-based features. Finally, the features are normalized by applying the moving average smoothing technique over a window of the past seven days, as it was proposed by O'Connor et al. [9].

2) Inference Method: The voting intention inference was approached as a multiple linear regression analysis. In this way, several regression models were built to infer the vote of each candidate in the two electoral rounds, using the aggregated polling as the output variable of the models. In total, nine models were built, six of which were used to infer the voting in the first round election (one model per electoral option).

In order to choose the best setting of each model, the first 80% observations were used as the training set and the remaining as the test set. Under the proposed method, a regression model built to infer the voting of candidate c in either the first round or the run-off of the election, receives the feature vector computed for candidate c on day d and produces the voting share candidate c would receive if the election was held on day d. The Scikit-learn [46] implementations of the Ordinary Least Square, Ridge Regression, Lasso, and

Table V
RESULTS AND VOTING INFERENCE PER METHOD IN THE FIRST ROUND ELECTION. NUMBERS IN BOLD SHOW THE INFERENCE METHOD WITH THE LOWEST ABSOLUTE ERROR THAT CORRECTLY RANKED A CANDIDATE

Candidate	Result	Polls	Twitter volume	Proposed method
Zuluaga	29.28%	27.53%	24.10%	**29.21%**
Santos	25.72%	28.99%	35.12%	**28.34%**
Ramírez	15.52%	9.43%	8.99%	9.23%
López	15.21%	10.56%	12.09%	10.15%
Peñalosa	8.27%	11.25%	8.13%	11.54%
Blank vote	5.98%	12.24%	11.65%	12.87%

Table VI
RESULTS AND VOTING INFERENCE PER METHOD IN THE RUN-OFF ELECTION. NUMBERS IN BOLD SHOW THE INFERENCE METHOD WITH THE LOWEST ABSOLUTE ERROR THAT CORRECTLY RANKED A CANDIDATE

Candidate	Result	Polls	Twitter volume	Proposed method
Santos	50.98%	44.97%	**52.37%**	43.25%
Zuluaga	44.98%	**43.74%**	32.27%	46.29%
Blank vote	4.02%	**11.29%**	15.36%	12.94%

Support Vector Regression were used to train the different model settings via cross validation on the training set. Based on the performance of the settings on the test set, in terms of *mean absolute error* (MAE), the best one was chosen.

B. Results

Tables V and VI show the official results and the voting intention inference in the first round and the run-off of the Colombian election, respectively. The *Result* column contains the official results of the election. Instead, the *Polls* and *Twitter volume* columns show the inferences from two baseline methods: the first one corresponds to the aggregated poll reports and the second one is based on Twitter volume [26]. The results of the proposed method are shown in the last column. Considering the Colombian law regulating the electoral polling in presidential elections [47], the results of Twitter volume and the proposed method for the ninth day before the election dates were used as the final inferences.

The inference results of the proposed method were good enough in the first round election, with a MAE of four percentage points (the lowest one) and the highest-polling candidates correctly ranked. However, the results of the proposed method in the run-off election were worse than those of the baseline methods, being the poll-based inference the one that correctly ranked all the candidates with the lowest MAE (4.84%). At last, although the Twitter volume method obtained the highest MAE in both rounds of the election, it correctly ranked the contenders in the run-off.

The obtained results show that inference methods based on Twitter data are not consistent, and therefore, more effort needs to be put into automated identification of demographic data (e.g., sex, age, and geographic location) in order that the methodology of voting intention inference be consistently competent against the statistical sampling methods employed in professional polling [12]. Twitter demographics are, hence, major issues to be addressed in future research in order to extract useful insights from the large amount of user-generated content; the insights could lead to remove demographic bias in Twitter data.

VII. CONCLUSION AND FUTURE WORK

Social media analysis represents a prolific research trend that demands a cautious handling. Its potential partly depends on the acknowledgment of its particularities and of the appropriate selection of the data it provides. In this research, two major issues in the search for more reliable public opinion measurements from Twitter data regarding an electoral race have been dealt with. Firstly, spammer accounts and their tweets were removed from the data used. Secondly, state-of-the-art approaches to sentiment analysis in Twitter were implemented to rate Spanish political tweets. Then, the potential of social media to infer voting intention was investigated. The obtained results showed that inference methods based on Twitter data are not consistent, reason why emphasis is placed on demographic characterization of users in order to set consistently competent inference methods.

There are several potential future directions based on this work. Firstly, the sentiment analysis system should learn to deal with figurative language. Secondly, because the voting intention inference was based on the aggregation of opinion orientation, sentiment analysis should be tackled as a quantification problem instead of a classification problem. Lastly, demographic characterization of Twitter users should be addressed in order to extract insights that lead to remove demographic bias in the data.

ACKNOWLEDGMENT

The authors sincerely thank Camilo López, Róbinson Alvarado, and anonymous reviewers for their valuable comments and suggestions. Thanks also to the volunteers involved in the labeling process.

REFERENCES

[1] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,” in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Springer US, 2012, pp. 415–463.

[2] Internet World Stats, “Internet world users by language – top 10 languages,” http://www.internetworldstats.com/stats7.htm, 2015, (accessed: February 01, 2016).

[3] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, “Tedas: A Twitter-based event detection and analysis system,” in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, April 2012, pp. 1273–1276.

[4] J. Stecyk, “Study: Twitter users love mobile apps,” http://www.webcitation.org/6g3pmjP9k, 2015, (accessed: November 10, 2015).

[5] H. Schoen, D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor, “The power of prediction with social media,” *Internet Research*, vol. 23, no. 5, pp. 528–543, 2013.

[6] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, “The rise of social bots,” *CoRR*, vol. abs/1407.5225, 2014.

[7] B. Han and T. Baldwin, “Lexical normalisation of short text messages: Makn sens a #Twitter,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 368–378.

[8] J. M. Cotelo, F. L. Cruz, J. A. Troyano, and F. J. Ortega, "A modular approach for lexical normalization applied to Spanish tweets," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4743–4754, 2015.

[9] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *International AAAI Conference on Weblogs and Social Media*, 2010.

[10] 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, 2013.

[11] A. Tsakalidis, S. Papadopoulos, A. I. Cristea, and Y. Kompatsiaris, "Predicting elections for multiple countries using Twitter and polls," *Intelligent Systems, IEEE*, vol. 30, no. 2, pp. 10–17, 2015.

[12] P. T. Metaxas, E. Mustafaraj, and D. Gayo-Avello, "How (not) to predict elections," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, Oct 2011, pp. 165–171.

[13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, ser. EMNLP '02. Association for Computational Linguistics, 2002, pp. 79–86.

[14] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Association for Computational Linguistics, 2002, pp. 417–424.

[15] S. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," in *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[16] M. Hagen, M. Potthast, M. Büchner, and B. Stein, "Webis: an ensemble for Twitter sentiment detection," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, 2015, pp. 582–589.

[17] L.-F. Hurtado, F. Pla, and D. Buscaldi, "Elirf-upv en tass 2015: Análisis de sentimientos en Twitter," in *Proceedings of the Sentiment Analysis Workshop at SEPLN (TASS2015)*, September 2015, pp. 75–79.

[18] C. D. Manning, P. Raghavan, and H. Schütze, "Scoring, term weighting and the vector space model," in *An Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[19] Y. Miura, S. Sakaki, K. Hattori, and T. Ohkuma, "Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data," in *Proceedings of the eighth international workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, August 2014.

[20] S. Amir, M. Almeida, B. Martins, J. Filgueiras, and M. J. Silva, "Tugas: Exploiting unlabelled data for Twitter sentiment analysis," in *Proceedings of the eighth international workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, August 2014.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[22] X. Saralegi and I. S. Vicente, "Elhuyar at tass 2013," in *Proceedings of the Sentiment Analysis Workshop at SEPLN (TASS2013)*, September 2013.

[23] M. Díaz-Galiano and A. Montejo-Ráez, "Participación de sinai dw2vec en tass 2015," in *Proceedings of the Sentiment Analysis Workshop at SEPLN (TASS2015)*, September 2015, pp. 59–64.

[24] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, June 2014, pp. 1188–1196.

[25] M. Huberty, "Can we vote with our tweet? on the perennial difficulty of election forecasting with social media," *International Journal of Forecasting*, vol. 31, no. 3, pp. 992–1007, 2015.

[26] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *International AAAI Conference on Weblogs and Social Media*, 2010.

[27] A. Jungherr, P. Jürgens, 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., &amp; 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.

[28] L. Shi, N. Agarwal, A. Agrawal, R. Garg, and J. Spoelstra, "Predicting us primary elections with Twitter," 2012.

[29] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans. Dependable Secur. Comput.*, vol. 9, no. 6, pp. 811–824, 2012.

[30] J. A. Cerón-Guzmán and E. León, "Detecting social spammers in Colombia 2014 presidential election," in *Advances in Artificial Intelligence and Its Applications: 14th Mexican International Conference on Artificial Intelligence, MICAI 2015, Cuernavaca, Morelos, Mexico, October 25-31, 2015. Proceedings, Part II*, O. P. Lagunas, O. H. Alcántara, and G. A. Figueroa, Eds. Springer International Publishing, 2015, pp. 121–141.

[31] G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Trevino, and J. Gordon, "Empirical study of machine learning based approach for opinion mining in tweets," in *Advances in Artificial Intelligence: 11th Mexican International Conference on Artificial Intelligence, MICAI 2012, San Luis Potosí, Mexico, October 27 – November 4, 2012. Revised Selected Papers, Part I*, I. Batyrshin and M. G. Mendoza, Eds. Springer Berlin Heidelberg, 2013, pp. 1–14.

[32] J. Villena-Román, J. García-Morera, S. Lana-Serrano, and J. C. González-Cristóbal, "Tass 2013 - a second step in reputation analysis in Spanish," *Procesamiento del Lenguaje Natural*, vol. 52, no. 0, pp. 37–44, 2014.

[33] Infométrika, "Encuesta de intención de voto elecciones presidenciales Colombia 2014," http://www.webcitation.org/6g3mVY6pk, 2014, (accessed: December 15, 2015).

[34] El Tiempo, "Zuluaga ganaría en primera y segunda vuelta, dice encuesta del cnc," http://www.webcitation.org/6g3nwoz3z, 2014, (accessed: December 15, 2015).

[35] Cifras y Conceptos, "Polimétrica," http://www.webcitation.org/6g3oHiI58, 2014, (accessed: December 15, 2015).

[36] El Tiempo, "Santos y Zuluaga disputarían la presidencia en una segunda vuelta," http://www.webcitation.org/6g3oqbXc2, 2014, (accessed: December 15, 2015).

[37] N. Arteaga and E. Guerra, "Las encuestadoras se lavan las manos en la recta final de la campaña presidencial," http://www.webcitation.org/6g3p9G1lE, 2014, (accessed: December 15, 2015).

[38] La FM, "Última gran encuesta: Óscar Iván Zuluaga 29.5 %, Juan M. Santos 28.5 %," http://www.webcitation.org/6g3oYtJ2k, 2014, (accessed: December 15, 2015).

[39] El Tiempo, "Óscar Iván Zuluaga, 47 %; Juan Manuel Santos, 45 %," http://www.webcitation.org/6g3p1tS1z, 2014, (accessed: December 15, 2015).

[40] ——, "Leve ventaja de Santos en carrera con Zuluaga," http://www.webcitation.org/6g3otoW9E, 2014, (accessed: December 15, 2015).

[41] Colprensa, "Resultados de la gran encuesta de los medios de junio de 2014," http://www.webcitation.org/6g3pXRoUX, 2014, (accessed: December 15, 2015).

[42] La FM, "Última gran encuesta: Óscar Iván Zuluaga (49%) y Juan M. Santos (41%)," http://www.webcitation.org/6g3oiQRtJ, 2014, (accessed: December 15, 2015).

[43] D. Vilares, M. A. Alonso, and C. Gómez-Rodrıguez, "On the usefulness of lexical and syntactic processing in polarity classification of twitter messages," *Journal of the Association for Information Science and Technology*, 2014.

[44] L. Padró and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA, May 2012.

[45] J. A. Cerón-Guzmán and E. León-Guzmán, "Lexical normalization of Spanish tweets," in *Proceedings of the 25th International Conference Companion on World Wide Web*, ser. WWW '16 Companion. International World Wide Web Conferences Steering Committee, 2016, pp. 605–610.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[47] Congreso de la República de Colombia, "Ley 996 de 2005," http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=18232, 2005, (accessed: January 25, 2016).