

# Sentiment-based Features for Predicting Election Polls: a Case Study on the Brazilian Scenario

Diego Tumitan and Karin Becker

Instituto de Informática

Universidade Federal do Rio Grande do Sul, Brazil

Email: {dctumitan, karin.becker}@inf.ufrgs.br

**Abstract**—The success of opinion mining for automatically processing vast amounts of opinionated content available on the Internet has been demonstrated as a less expensive and lower latency solution for gathering public opinion. In this paper, we investigate whether it is possible to predict variations in vote intention based on sentiment time series extracted from news comments, using three Brazilian elections as case study. The contributions of this case study are: a) the comparison of two approaches for opinion mining in user-generated content in Brazilian Portuguese; b) the proposition of two types of features to represent sentiment behavior towards political candidates that can be used for prediction, c) an approach to predict polls vote intention variations that is adequate for scenarios of sparse data. We developed experiments to assess the influence on the forecasting accuracy of the proposed features, and their respective preparation. Our results display an accuracy of 70% in predicting positive and negative variations. These are important contributions towards a more general framework that is able to blend opinions from several different sources to find representativeness of the target population, and make more reliable predictions.

## I. INTRODUCTION

Opinions are key influencers of our behaviors. Governments, companies and organizations rely on public opinion to define strategies to improve the services they provide, or increase the success and visibility of the brands, entities and causes they represent. People increasingly share opinions on the Internet, through social networks (e.g. Facebook, Twitter), on-line newspapers, etc. Opinion mining [1] aims at automatically identify opinionated content, and determining people's sentiment, perception or attitude towards an entity or topic. Using opinion mining, it is possible to automatically analyze this vast and rich user-generated content, and develop less expensive and lower latency solutions for public opinion elicitation, with a reasonable degree of accuracy.

Assuming human sentiment can be characterized by automatic techniques on an acceptable accuracy level, so the next question is whether sentiment may be used to predict future behavior. Experiments of using sentiment expressed on Twitter have been developed for targets such as predicting stock market movement [2], [3], election or poll results [4], [5], sales performance [6], and movies' box-offices [7].

The representativeness of Twitter users for predictions in some domains, such as politics, have been severely questioned [8]. Certainly, using a single source of opinion may introduce a bias in the final prediction model because the opinion expressed is representative of a specific population. Thus, different sources of opinion must be investigated, in order to understand the underlying behavior and the public

they represent. For instance, in the political domain, we observed in [9] that authors of news' comments have a different behavior: they mainly expose their views and beliefs about politics in general. Understanding the role, the behavior, and how to weight opinions from different sources are fundamental steps towards a very challenging and real problem, which is how to combine different sources of opinions to constitute a significant and representative opinion sample.

Most works on sentiment-based prediction use long and daily time series for both the sentiment and the variable to be predicted. However, historical values for some types of variables to be predicted may be sparse. For example, in the United States of America, vote intention for the Presidential election are polled daily by many organizations (e.g. TV broadcasters, marketing companies, etc.). In Brazil, the scenario is completely different, as public vote intention polls can only be published by authorized research consulting companies, and they occur rather infrequently. Each authorized organization publishes about 12 polls every year, and most of them are concentrated during the month that precedes the election. The time elapsed between any two published polls varies enormously (from days to months). Besides vote intention, there exists many other contexts that present sparse data. For instance, one may wish to investigate whether public indicators about health, education or security may be predicted based on the public opinion expressed about these services. These indicators are infrequent, and usually are produced according to specific conditions (e.g. census).

In this paper, we develop a case study to investigate whether it is possible to predict variations in vote intention polls, based on the sentiment expressed on user-generated comments on newspapers. More specifically, we consider the Brazilian political scenario in which public election-related polls constitute sparse data, i.e., there are few data points and the time elapsed between two polling execution varies. Public polls are very important, as political parties explore their results as a major part of their campaigning strategy. Political parties can conduct their own polls, but results cannot be published. In addition, this possibility is subject to budgetary constraints, which may influence their precision. Our case study involves three elections, comments extracted from a major Brazilian on-line newspaper and the polls published by one of the most traditional consulting companies.

The contributions of this paper are: a) a case study evaluating two approaches for opinion mining in user-generated content in Brazilian Portuguese; b) the proposition of features to represent sentiment behavior towards political candidates

that can be used for prediction, and c) an approach to predict vote intention variations between polls, which is adequate for scenarios of sparse data. Works dealing with Portuguese language are scarce, and most efforts are directed towards the English language. With regard to sentiment features, we propose: a) metrics summarizing positive/negative sentiment about candidates, and b) existence of bursts of sentiment expression (i.e. more/less sentiment than usual). Finally, we simplify the prediction problem using the discrete variation of vote intention, in order to deal with sparsity.

We acknowledge that this source of opinions include bias, just as Twitter. However, these are important contributions towards a more general framework that is able to blend opinions from several different sources to find representativeness of the target population, and make more reliable predictions.

The remainder of this paper is organized as follows. Related work is discussed in Section 2. Section 3 presents an overview of the proposed approach. The case study containing the opinion mining process are described in Section 4. The part of the case study that concerns the prediction process and its results are detailed in Section 5. Conclusions and future work are addressed in Section 6.

## II. RELATED WORK

**Opinion Mining.** It aims to identify subjective content (e.g. web pages, products reviews, tweets), classify its polarity, and summarize the overall sentiment [1]. An opinion has two main attributes: a *target* (e.g. product, brand, event) and a *sentiment* towards this target, also referred to as its *polarity*. Most works are concerned only about positive and negative sentiment, disregarding other classes (e.g. neutral). Several approaches have been used in opinion mining to classify the polarity of subjective content, such as [10], [1]: dictionary-based, machine learning, and statistical. Dictionary-based approaches are very popular, and rely on the availability of generic or domain-dependent sentiment lexicons to provide sentiment words. In the supervised machine learning approach, an annotated corpus is submitted to a classifier. The quality of the results is dependent the availability of an extensive and domain-specific annotated corpus, but the annotation process is laborious and subjective. Sentiment classification can be developed in several levels: document, sentence and aspect. The latter two are more used when a document contains opinions about several targets.

There exist several resources for opinion mining targeted at the English language, including sentiment lexicons, natural language processing (NLP) tools, and annotated corpora. Resources for other languages, such as Portuguese, are scarce. The most complete sentiment lexicon for Portuguese is SentiLex-PT [11], which contains 7,014 lemmas and 32,347 inflected forms for Portugal Portuguese. Palavras [12] is the most complete NLP tool. A technique to create a reference corpus for opinion mining in Portuguese automatically is presented in [13], where the authors derive from comments to political news, syntactic-semantic patterns to identify the polarity of sentences. These resources do not always perform well with regard to Brazilian Portuguese, as well as with user-generated content, which is informal and often presents errors.

**Sentiment-based prediction.** Many works have addressed prediction based on sentiment, mostly using Twitter. A case

study is reported in [7], where the popularity of pre-defined movies on Twitter is used to predict their box-office, using both the volume and the Twitter sentiment. They consider two dozens of movies and daily time series of nearly three months each. Linear regression is used to correlate sentiment time series and indicators to be predicted. They conclude that prediction is more affected by the number of mentions, but sentiment can be used in combination to increase accuracy. Twitter sentiment is also used to predict the stock market movement in [2]. Two long daily time series are considered for prediction using linear and non-linear techniques: sentiment and stock market indicators. The authors stress the importance of smoothing and lags between events and stock movements. They also stress that sentiment polarity is not a good predictor, and make experiments with emotions (e.g. joy, calm).

The work that most resembles ours in terms of goal is the case study reported in [4], in which the sentiment expressed in tweets containing mentions to candidates are correlated with external indicators of consumer confidence (e.g. Index of Consumer Sentiment) and political opinion (presidential job approval and vote intention). Using daily time series, they try to predict the polls via linear regression, which performs poorly for all the indicators. None of these techniques can be applied to sparse data, and therefore are not suitable for our case study.

Also in the political context, a study [5] concludes that the German election result could be predicted using tweets involving mentions to running parties. The sentiment of the tweets was examined, but the mere number of mentions to the political parties strongly correlated with their respective share of votes in the election results. The use of Twitter as a major source for election results prediction is questioned in [8], with the main argument that Twitter users and opinion content may not be a representative sample of the target population.

**Our Work.** Our work complements the aforementioned works by taking into account a different source of opinions, namely comments on newspapers. Although the bias of a single opinion source is also present in our work, it provide insights how users issue opinions in this media. In a previous study [9], we developed an approach to identify, extract and summarize the opinion contained in comments about political news. We observed that newspaper readers have a different behavior compared to Twitter users, because they do not support nor detract candidates: instead, they comment about their frustration about politics in general, with a significant majority of negative comments over candidates and their parties. We also proposed summarization metrics inspired by related work [14], [2], and analyzed their correlation with acceptance and rejection indicators available in public polls. To the best of our knowledge, metrics that characterize burst of sentiment have not been proposed before. The work presented in this paper builds on this previous experience, to address the problem of vote intention prediction.

## III. APPROACH OUTLINE

In this paper, we investigate through a case study whether it is possible to predict vote intention variation based on a daily time-series of sentiment. The distinctive aspects of this case study are: a) we deal with users' comments written in Brazilian Portuguese as a reaction to political news; b) we compare the performance of dictionary-based and machine

learning opinion mining approaches for user-generated content in Brazilian Portuguese; c) we deal with the prediction problem as the classification of vote intention variation, in order to deal with the sparseness of vote intention data, compared to the sentiment time series; d) we propose features that characterize bursts of sentiment expression to be used for prediction, and; e) we compare the cumulative effect of the sentiment (since the start of the campaign), and its short-term effect (since the last poll). With regard to this last aspect, the cumulative effect translates the overall strategy of the campaign, whereas the short-term effect represents the course of actions taken in response to the results of the last poll.

An overview of the proposed approach for the case study is shown in Figure 1, which is briefly explained in this section.

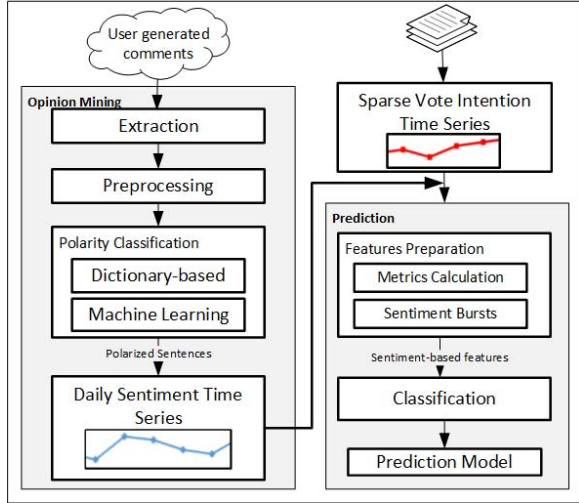


Fig. 1. Overview of the proposed approach.

**Opinion Mining:** this process aims at deriving a sentiment time series for each election candidate, based on daily comments about news. It is composed of the traditional steps of opinion mining [10]: a) extraction of comments on selected news about politics; b) preprocessing comments to handle noise, identifying the target entities, and separating comments in sentences; c) polarizing sentences with mentions to candidates of interest as positive/negative; and d) summarizing the polarized sentences in time series that quantifies the positive and negative sentiment for each candidate, together with the number of mentions. The sentence level of analysis was chosen because most comments refer to more than one candidate. Section IV details the opinion mining techniques used in the case study, the challenges and the results obtained. We complement our previous work [9] by comparing the dictionary and the machine learning approaches for polarity classification.

**Time Series:** the prediction process has as input two time series: a daily time series of sentiment and a vote intention time series. The sentiment time series is the result of the opinion mining process applied over daily news comments. The vote intention time series is extracted from public election polls, and it is sparse because it has very few elements when compared to the sentiment time series. About 12 polls are published every year, and the time elapsed between any two published polls varies enormously (from days to months).

Let  $T = [t_1, \dots, t_n]$  and  $K = [k_1, \dots, k_m]$  be time indexes, where  $K \subset T$  and  $k_1 = t_1$  and  $k_m = t_n$ . Let  $E = [e_1, \dots, e_i]$  be a set of observed candidates. Formally, the sentiment time series is defined as  $V = \{v_{jt} : t \in T, j \in E\}$ .  $v_{jt}$  is a quadruple  $\langle e_j, pos_{jt}, neg_{jt}, m_{jt} \rangle$ , where

- $e_j$  is an entity, i.e., an observed candidate;
- $pos_{jt}$  is the total positive sentiment towards the entity  $e_j$  at time  $t$ ;
- $neg_{jt}$  is the total negative sentiment towards the entity  $e_j$  at time  $t$ ;
- $m_{jt}$  is the total number of mentions to the entity  $e_j$  at time  $t$ ;

The vote intention time series is formally defined as  $P = \{p_{jt} : t \in K, j \in E\}$ .  $p_{jt}$  is a pair  $\langle e_j, int_{jt} \rangle$ , where:

- $e_j$  is an entity;
- $int_{jt}$  is the vote intention of the entity  $e_j$  at time  $t$ ;

**Prediction:** this process aims at developing a model that, given features extracted from a sentiment time series, can predict a variation in vote intention. Recall that the problem was simplified as the prediction of a discrete variation of vote intention (i.e. increased, decreased, unchanged). The prediction process is composed of two steps: the preparation of features, and the learning of the predictive model through classification.

In the *Features Preparation* step, the two time series  $V$  and  $P$  are transformed. For any two consecutive data points  $t_i, t_k \in K$  ( $t_i < t_k$ ), one record is prepared containing sentiment-based features derived from  $V$ , together with the corresponding discrete vote intention variation extracted from  $P$ . We trained the classifier with two types of sentiment-based features: a) *summarization metrics* aggregating in various ways positive and negative sentiment towards candidates, and b) *bursts of sentiment* expression towards candidates. We prepared these features to represent both the *cumulative effect* of the sentiment (since the start of the campaign), and its *short-term effect* (since the last poll was published).

In the *Classification* step, the prepared data is used as training/testing data for a classification algorithm. We developed experiments to verify which type of sentiment feature presents the best predictive behavior, using different algorithms. Section V details the features considered for the case study, and the experiments develop towards a prediction model.

#### IV. CASE STUDY: OPINION MINING

##### A. Dataset

The dataset is composed of comments from political online news. The news were extracted automatically from the political section (tagged *Poder - Power*) from Folha online, one of the most popular newspaper in Brazil. The dataset contains news and their respective comments referring to 3 elections, and the collected data encompasses the month that preceded each election date (first round), a period when the public opinion polls are published more frequently. It should be clear that in Brazil, elections occur every four years and vote is mandatory. Gubernatorial and Presidential elections occur in the same year, and the Mayoral elections two years after. For each election, we selected the candidates with the highest vote intentions (and therefore, the most commented ones). They are described below, together with the respective party:

- **2010 Gubernatorial Election:** Aloízio Mercadante (PT) and Geraldo Alckmin (PSDB);
- **2010 Presidential Election:** Dilma Rousseff (PT), José Serra (PSDB), and Marina Silva (PV);
- **2012 Mayoral Election:** Celso Russomanno (DEM), Fernando Haddad (PT) and José Serra (PSDB).

We used the same set of sentences to observe sentiment for both Gubernatorial and Presidential elections candidates. As they involve candidates from the same political party that go on campaign together, often news and comments refer to both of them. We used the target of the opinion in the sentence to distinguish between the two elections (e.g. a mention to Dilma refers to Presidential election). The profiling of each dataset is described in Table I. In the remaining of this section, we shall refer to the 2010 and 2012 datasets.

### B. Opinion Sentences Gold-Standard

Through a manual annotation process of randomly selected comments from each dataset, we built a set of opinion sentences gold-standard. This gold-standard set is used to assess the classification performance of the dictionary-based approach, and as a training corpus for the machine learning approach. We randomly selected 1,000 and 600 sentences from the 2010 and 2012 datasets, respectively. For each dataset, we used 3 annotators with a major in computer science, and no previous experience on corpus annotation. They were instructed to base their classification in the content that was explicitly written, disregarding any assumption about political entities or parties [13], so that their political background would not interfere in their judgment. Only sentences with at least two agreements were retained, representing 92.7% of the annotated sentences for 2010, and 97% for the 2012 election. For the 2010 election, we obtained 356 sentences annotated as negative, 154 as positive, and 417 as neutral. The annotated set for the 2012 election contains 482 sentences annotated as negative, 72 as positive, and 28 as neutral. Although the instructions were the same, we noticed some differences of the two annotation processes. Compared to 2012, the 2010 election set of annotated sentences revealed a higher proportion of positive sentences, and a significantly higher number of neutral sentences, which may have influenced the results. During the annotation process, we also identified regional and idiomatic expressions, nicknames, and informal sentiment words, which were included in a specialized sentiment lexicon.

### C. Opinion Mining Process

The precision of the opinion mining results is fundamental, because it influences the prediction model. Building on our previous experience of dictionary-based opinion mining [9], we developed additional experiments with machine-learning algorithms. This section describes the pre-processing applied over the corpora, and discusses the results of both approaches.

1) *Extraction:* the datasets mentioned in Section IV-A were extracted according to the procedures detailed in [9].

2) *Pre-processing:* considering a set of user comments, this step is responsible for: a) handling noisy data, b) breaking each comment into sentences, c) identifying the sentences with mentions to candidates, and d) extracting features from the sentences (e.g. unigrams) and transforming them according to

each specific polarity classification technique. Below, we detail how the most important issues were handled.

We noticed the existence of a great amount of duplicated or near-duplicated comments. To remove this spam and avoid bias, we used the Cosine Similarity to obtain a similarity score between all pairs of comments of a same dataset. All comments with a similarity above 85% were eliminated. We also removed comments that were too short (less than 4 words). The results of the pre-processing are displayed in Table I.

Another important noise was the disguise of cursing words using special characters, possibly due to newspaper moderation (j@ck@ss → jackass). To solve that, we replaced pre-selected special characters for the corresponding letters, possibly introducing errors that did not exist previously. However, a manual analysis revealed that, in most cases, it referred to disguised negative sentiment words. To handle the misuse of accentuation, we substituted all accentuated letters for plain ones, both on the sentences and the adopted sentiment dictionary. In addition, to find specific vocabulary for the dictionary approach, we manually analyzed the top 1,000 more frequent words that were not found in the used dictionary.

To identify mentions to candidates, we used their names and possible alternative terms. We compiled a set of alternative mentions using regular expressions based on variations of their names (e.g. José Serra was mentioned as “zehserra”, “serrinha”, etc). Some of these mentions imply sentiment (e.g. malhaddad - Mean Haddad, vamp Serra - Vampire Serra, Dilmais - Super Dilma). These terms were also added to a domain specific sentiment lexicon. The lack of proper NLP tools did not allowed us to resolve co-references (e.g. anaphora). Finally, to break comments into sentences, we used *punkt*, a specific NLTK<sup>1</sup> module trained for Portuguese. As a result, we obtained 80,469 sentences with mentions (69,490 for the 2010 dataset, and 10,979 for the 2012 dataset).

We developed many experiments that are not reported here, including the removal of stop words, variations of unigrams (n-grams), stemming, and handling of negation using a proximity window. None of these actions yielded good results. Using Palavras, we developed experiments for breaking sentences into clauses, handling negation and discovering the actual target of sentiment words. However, it also did not present good performance due to the excess of syntactical and structure errors, as well as use of informal language. This paper concentrates on reporting only most successful experiments.

3) *Polarity Classification:* This step aims at polarizing each sentence, and assigning this polarity to a target candidate. Again, the Portuguese language was the biggest challenge for this step. For the dictionary approach, the challenge was the lack of a good sentiment lexicon for Brazilian Portuguese. For the machine learning approach implies significant work for annotating the corpus for quality results. We experimented with both approaches and compared their results.

In the dictionary-based approach, the sentiment lexicon SentiLex-PT [11] was used to polarize the sentences into positive, negative or neutral (1, -1 and 0). To compensate its limitations with regard to the Brazilian language and domain-dependent informal vocabulary, we created a specialized lex-

<sup>1</sup><http://nltk.org>

TABLE I. DATA PROFILE FROM THE DATASETS.

	2010 Election		2012 Election	
	Raw	Preprocessed	Raw	Preprocessed
Number of news	2,232	1,763	583	340
Number of comments	225,217	190,975	36,108	25,115
Mean of comments per news (std)	98.59 ( $\pm 235.6$ )	86.09 ( $\pm 206.5$ )	61.93 ( $\pm 142.5$ )	44.06 ( $\pm 92.5$ )
Number of sentences	-	673,146	-	79,752
Mean of sentences per comment (std)	-	3.05 ( $\pm 2.06$ )	-	3.17 ( $\pm 2.34$ )
Comments with less than 4 words	5,148	0	7,185	0
Comments with similarity greater than or equal to 85%	29,094	0	3,808	0
Period	2010-09-01 to 2010-10-03		2012-09-01 to 2012-10-07	
Entities	5 candidates		3 candidates	

icon containing sentiment words of the Brazilian political scenario, slang, idiomatic and regional expressions [9]. In this approach, for every word in a sentence, we look up in both dictionaries if the word has an associated polarity. We then aggregate the polarity of all words in the sentence, where the positive terms are added, and negative terms are subtracted.

For the machine learning approach, we tested with different classifiers available in Weka [15]. The best results were obtained using the Sequential Minimal Optimization (SMO) algorithm to train a Support Vector Machine (SVM) classifier. We trained the SVM to classify sentences into positive and negative. As features, we used unigrams of the documents with their respective word count, and applied TF-IDF transformation to the analyzed text.

We experimented different feature transformations to improve the performance in this approach, including: a) feature reduction (simple count cutoff and information gain), b) use of binary feature representation (present/absent), c) bigrams and trigrams, etc. These variations did not yield better results compared to the features configuration described above, and are not reported in this paper.

Regardless the polarization approach, the result is a set of polarized sentences, associated with the respective candidate and timestamp. Formally, the set of sentences is defined as  $S = \{s_i : i \in \mathbb{N}\}$ , where  $s_i$  is a quadruple  $\langle text_i, pol_i, e_i, t_i \rangle$ , and:

- $text_i$  is the pre-processed sentence;
- $e_i$  is the entity mentioned in the sentence ( $e_i \in E$ );
- $pol_i$  is the sentiment polarity towards the entity  $entity_i$ ;
- $t_i$  is the time when the comment to which the sentence belongs was written.

4) *Sentiment Summarization*: This step considers all the polarized sentiment sentences of the set  $S$ , and use the time stamps to aggregate them per day and candidate. Then, it builds for each candidate a daily sentiment time series to compose  $V$ , as defined in Section III. Every element in a sentiment time series ( $v_{jt} \in V$ ) aggregates per candidate and per day, the amount of all positive sentences, amount of negative sentences, and the total number of mentions to candidate (regardless the sentiment). Only sentences polarized as positive or negative are considered for the sentiment summarization, and all sentences are considered for the mentions summarization.

#### D. Sentiment Classification Results Assessment

Using the opinion sentences gold-standard, we compared the performance of dictionary-based and machine learning experiments using standard measures (accuracy, precision, recall

and F-measure). For the experiments with SVM, we developed two variations: a) use of both 2010 and 2012 sentence sets with 10 fold cross-validation, and b) use of the 2010 set of sentences as a training set, and the 2012 set as a testing set. We focus on the results of the most successful experiments only.

The dictionary-based approach presented nearly the same performance for both 2010 and 2012 datasets, keeping roughly the same precision, recall and F-score. However, the precision for negative sentences of the 2012 election (90.18%) was substantially higher compared to the 2010 dataset (54.79%).

Among the classifiers, SVM consistently presented the best performance. In the 2010 Election set of sentences, the accuracy was 81.37%. Compared to the dictionary-based approach in the same dataset, the precision, recall and F-measure were better for both classes (positive and negative). For the 2012 election, the accuracy of the SVM with cross-validation (83.24%) was also better, though mainly due to the results for the negative class.

When considering the use of the 2010 sentences as a training set, and 2012 as a testing set, SVM accuracy drops significantly (77.4%), and we believe this difference is due to overfitting. To investigate this hypothesis, we applied an Information Gain feature selection function to both training and testing sets. We noticed a completely different influential vocabulary in both elections. For instance, the 2010 presidential election made several references to the first female presidential candidates, and the respective features were determinant for the positive class. On the 2012 elections, the vocabulary refers to the past history of each candidate, and their involvement with scandals.

Consistently, we faced bad results for positive class with all methods. One of the reasons is that positive sentences are really scarce in the dataset, so both training and validation using our set of gold-standard opinion sentences may be biased. Another explanation is the extensive use of irony in the comments.

The prediction experiments described in the next section were developed using sentiment classified using the SVM approach with cross-fold validation.

## V. CASE STUDY: SENTIMENT-BASED PREDICTION

### A. Public Opinion Polls

We used the public polls of vote intention published by Datafolha<sup>2</sup>, one of the most traditional and respected research consulting companies. All the polls correspond to the first round of their respective election. The polls dates from each

<sup>1</sup><http://datafolha.folha.uol.com.br>

TABLE II. EXPERIMENTS RESULTS ACCORDING TO ACCURACY (A), PRECISION (P), RECALL (R) AND F-SCORE (F).

Election	Approach	A(%)	Polarity	P (%)	R (%)	F(%)
2010	Dictionary-based	50.39	Positive	29.39	62.99	40.08
			Negative	54.79	44.94	49.38
	SVM	<b>81.37</b>	Positive	<b>70.63</b>	<b>65.58</b>	<b>68.01</b>
			Negative	<b>85.56</b>	<b>88.20</b>	<b>86.86</b>
2012	Dictionary-based	52.14	Positive	26.99	<b>56.41</b>	<b>36.51</b>
			Negative	<b>90.18</b>	51.45	65.52
	SVM	<b>83.24</b>	Positive	<b>27.59</b>	10.53	15.24
			Negative	86.45	<b>95.38</b>	<b>90.70</b>
	SVM (2010 as training set)	77.40	Positive	25.56	30.26	27.71
			Negative	87.98	85.27	86.61

election are different from one another. We used as first data point in all vote intention time series, the numbers available from the poll prior to September 1<sup>st</sup> of the respective year. The lag between any two polls ranges from 8 to 3 days.

### B. Predictive Sentiment Features

Two types of sentiment features are proposed in this case study to train the prediction classifier: summarization and bursts. Each type of feature is described in detail below, followed by the variations in how they are prepared (short term and cumulative).

**Summarization Metrics:** Table III describes six different metrics to summarize the sentiment over the time, which are variations of the ones proposed in related work [14], [7], [2]. We also included a metric to summarize mentions to candidates (metric  $s7$ ). In formulae 1-7,  $E$  corresponds to the candidate set of the election,  $j \in E$  corresponds to a candidate, and  $Q \subseteq T$  refers to a time period in which the sentiment is aggregated. All metrics are represented by ratios in order to normalize the data considering all three elections, and render them comparable. Indeed, the volume of comments is different according to the audience concerned by the news, namely inhabitants of the city of São Paulo (municipal election), state of São Paulo (gubernatorial election) and Brazilians (Presidential election).

**Sentiment Bursts:** this type of feature indicates that people are issuing significantly more opinionated comments than usual. Bursts reveal heated reactions to news. We created this feature based on the assumption that bursts express reactions to events that may influence vote intentions, and thus, may be used to predict their variation.

For instance, if there are bursts of negative sentiment (e.g. a reaction to a scandal), the vote intention for a candidate targeted of that sentiment may decrease. As a concrete example, we can mention the burst of comments related to the news that a drag queen kissed a candidate, who is perceived as too serious. His naive reaction to the kiss generated a burst of positive comments about his image, and decreased his rejection rate.

To identify bursts, we adopted the quantization process described in [16], which aims at detecting events of notice in sensor-produced time series data. It identifies “peaks”, “valleys” and “plateaus” in a time series. Given a threshold, peaks correspond to values that are considered much greater than expected, whereas valleys are much lower than expected.

Considering the time series set  $V$  resulting from the opinion mining process, we prepared three time series for each entity  $j \in E$ : positive sentiment ( $POS_j$ ), negative

TABLE III. DESCRIPTION OF SUMMARIZATION METRICS

Description	Metric
Ratio of positive sentiment towards an entity to the negative sentiment towards the same entity	$s1_{jQ} = \frac{\sum_{t \in Q} pos_{jt}}{\sum_{t \in Q} neg_{jt}} \quad (1)$
Ratio of positive sentiment towards an entity to the total sentiment towards the same entity	$s2_{jQ} = \frac{\sum_{t \in Q} pos_{jt}}{\sum_{t \in Q} (pos_{jt} + neg_{jt})} \quad (2)$
Ratio of negative sentiment towards an entity to the total sentiment towards the same entity	$s3_{jQ} = \frac{\sum_{t \in Q} neg_{jt}}{\sum_{t \in Q} (pos_{jt} + neg_{jt})} \quad (3)$
Ratio of the difference between positive and negative sentiment towards an entity, to the total sentiment towards the same entity	$s4_{jQ} = \frac{\sum_{t \in Q} (pos_{jt} - neg_{jt})}{\sum_{t \in Q} (pos_{jt} + neg_{jt})} \quad (4)$
Ratio of positive sentiment towards an entity to the total positive sentiment (towards all entities)	$s5_{jQ} = \frac{\sum_{t \in Q} pos_{jt}}{\sum_{c \in E} \sum_{t \in Q} pos_{ct}} \quad (5)$
Ratio of negative sentiment towards an entity to the total negative sentiment (towards all entities)	$s6_{jQ} = \frac{\sum_{t \in Q} neg_{jt}}{\sum_{c \in E} \sum_{t \in Q} neg_{ct}} \quad (6)$
Ratio of mentions to an entity to the total of mentions (of all entities)	$s7_{jQ} = \frac{\sum_{j \in Q} m_{jt}}{\sum_{c \in E} \sum_{t \in Q} m_{ct}} \quad (7)$

sentiment ( $NEG_j$ ) and ratio of positive/negative sentiment ( $R_j$ ). Formally, these time series are defined as  $POS_j = \{pos_{jt} : t \in T, j \in E\}$ ;  $NEG_j = \{neg_{jt} : t \in T, j \in E\}$ ; and  $R_j = \{\frac{pos_{jt}}{neg_{jt}} : t \in T, j \in E\}$ . Applying the aforementioned quantization technique, we identified the peaks and valleys for each time series. Figure 2 exemplifies the peaks and valleys of a time series of sentiment ratio. We had to define different thresholds for each type of time series, as there are significantly more negative sentences than positive ones. The thresholds were set empirically. We propose four types of bursts, considering a given time period  $Q \subseteq T$ :

- *PositiveBurst<sub>jQ</sub>*: given  $POS_j$ , whether in the considered time interval  $Q$  there was at least one peak of positive sentiment;
- *NegativeBurst<sub>jQ</sub>*: given  $NEG_j$ , whether in the considered time interval  $Q$  there was at least one peak of negative sentiment;
- *DeltaBurst<sub>jQ</sub>*: given  $POS_j$  and  $NEG_j$ , the difference between the number of peaks and valleys identified in

- $RatioBurst_{jQ}$ : given  $R_j$ , whether there was a predominance of mountains or valleys observed in the time series, considering the time interval  $Q$ .

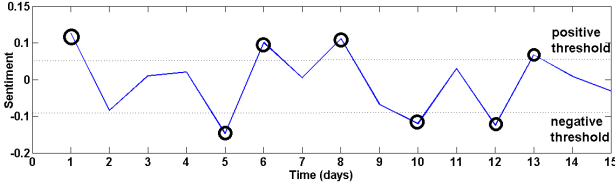


Fig. 2. Sentiment ratio time series with peaks and valleys

**Feature preparation variations:** We prepared each type of feature described above according to two variations, to represent both the cumulative effect of the sentiment (since the start of the campaign), and the short-term effect (since the last poll was published). Thus, given a timestamp  $t_k \in T$ ,  $k > 1$  and the time serie  $V$ , for each feature we calculate two variations, by considering different time intervals for  $Q$ , as follows:

- **cumulative:** this variation takes into account all sentiment expressed in  $S$  from the beginning of the observed period, i.e.  $Q = [t_1, \dots, t_k]$ .
- **short-term:** this variation takes into account all the sentiment expressed in  $S$  since the last poll, i.e.  $Q = [t_{k-1}, t_k]$ .

### C. Prediction Model Assessment

We developed a set of experiments using as target class to be predicted the discrete variation of vote intention extracted from the public polls. The predicted attributes were the sentiment-based features, prepared as described in Section V-B. By considering the variations between any two consecutive polls for each candidate, the training set contains 51 records for the “increased” class, 16 for the “decreased” class, and 10 for the “unchanged”.

We experimented with many different classification algorithms, but the OneR, available in Weka, yielded the best results. The experiments approached the prediction both as a three class problem, and as a binary class problem (i.e. increased and decreased). For the latter, all records for the unchanged class were disregarded. In general, classification performed substantially better in forecasting only two classes. The best result for predicting 3 classes was of 54.90%, compared to 70.74% for the binary classes. The best precision was always obtained for the “increased” class.

The experiments aimed at comparing the contribution of each type of feature for prediction accuracy. We used three criteria for this assessment: 1) the effect of cumulative versus the short-term features preparation; 2) comparison of the predictive power of metrics based on sentiment and candidate mentions, and 3) the effect of burst of sentiment expression. To measure this effect, each feature was submitted to the classification algorithm as a predictive attribute in isolation. Table IV and V shows the results for the summarization metrics and the sentiment bursts features, respectively, using short-term (ST) and cumulative (C) variations.

TABLE IV. ACCURACY OF SENTIMENT-BASED METRICS BASED ON THE SHORT-TERM (ST), AND CUMULATIVE (C) METRICS.

Features	No. classes	ST (%)	C(%)
s1 (Formula 1)	2 classes	43.90	48.78
	3 classes	35.29	41.17
s2 (Formula 2)	2 classes	43.90	48.78
	3 classes	35.29	41.17
s3 (Formula 3)	2 classes	43.90	53.65
	3 classes	43.13	47.05
s4 (Formula 4)	2 classes	53.65	58.53
	3 classes	39.21	33.33
s5 (Formula 5)	2 classes	53.65	56.09
	3 classes	45.09	43.13
s6 (Formula 6)	2 classes	<b>56.09</b>	<b>70.73</b>
	3 classes	<b>45.09</b>	<b>54.90</b>
s7 (Formula 7)	2 classes	43.90	39.02
	3 classes	41.17	43.13
Combined	2 classes	51.21	
	3 classes	39.21	

TABLE V. ACCURACY OF FEATURES BASED ON SHORT-TERM (ST) AND CUMULATIVE (C) BURSTS OF SENTIMENT.

Features (%)	No. classes	ST (%)	C(%)
RatioBurst	2 classes	60.97	56.09
	3 classes	49.01	41.17
PositiveBurst	2 classes	53.65	53.65
	3 classes	41.17	37.25
NegativeBurst	2 classes	60.97	60.97
	3 classes	49.01	41.17
DeltaBurst	2 classes	<b>60.97</b>	<b>65.85</b>
	3 classes	<b>49.01</b>	<b>41.17</b>
Combined	2 classes	43.90	
	3 classes	37.25	

1) *Cumulative effect vs. Short-term effect:* Among all proposed features we observed that, in general, the summarization metrics had a better performance considering the cumulative effect. However, sentiment bursts consistently displayed a slightly better performance for the short-term variation, except for the cumulative DeltaBurst. This provides evidences towards our assumption that bursts of sentiment may influence vote intention, particularly in a short-term scenario, i.e. people be influenced by news that were recently published on the media, when answering a poll.

2) *Sentiment vs. Mentions Summarization Metrics:* We also compared how the metrics based on variations of totals of sentiment (Formulae s1-s6) performed, when compared to the metric based on mentions to the candidates (Formula 7). Sentiment metrics performed significantly better. The best predictive performance was presented by metric s6 for a classification based on the cumulative metric ((70% of accuracy, against 43.90% of accuracy yielded by the short-term s7 mention metric). Differently from previous works on sentiment-based prediction, in our work, the simply mention to the observed entities did not overcome the predictive behavior of sentiment-based metrics [7], [2], [5].

3) *Sentiment Burst vs. Summarization Metrics:* Considering the best results for each type of feature (s6 and DeltaBurst), it is possible to see that each one performed better in a specific scenario (short and long term). Roughly, all burst features presented a similar performance, with a slight advantage towards DeltaBurst and NegativeBurst. The most common bursts of sentiment were comments generated by public attacks of running candidates to their competitors that were published on the news.

Finally, we considered combinations of these features. We

combined all sentiment metrics features, all burst features, all features, and the best sentiment and burst features. The first combination performed very poorly, displaying 51.21% and 37.25% of accuracy for predicting 2 and 3 classes, respectively. The same happened for the second combination, displaying 43.90% of accuracy for predicting 2 classes. Considering the combination of all features, the accuracy was 51.21% for 2 classes, and 39.21% for 3 classes. Finally, we submitted to the classifier only the best feature of each type (i.e. cumulative s6 and cumulative DeltaBurst), with a resulting accuracy of 70.73% and 54.90%, respectively. We also applied a feature selection algorithm based on information gain, and not surprisingly, it selected the features that had the best results when submitted individually.

## VI. CONCLUSION AND FUTURE WORK

In this paper, we examined whether the sentiment extracted from user-generated content with regard to political news could be used to forecast variations in vote intention polls. Although the problem is not new, the distinctive features of our case study were: the opinion data source (comments about news); comments were in Brazilian Portuguese, for which resources are scarce; and variables to be predicted were sparse, because they correspond to public vote intention polls that are published infrequently. We developed an approach for extracting sentiment of user-generated comments in Portuguese and examined two methods for opinion mining. We also proposed two types of features to represent the sentiment, summarization and bursts, for which we examine the predictive power through experiments. We addressed the sparseness of the to-be predicted data by considering vote intention variations over multiple elections.

We reached an accuracy of 70% of prediction for the binary class problem, mainly based on negative sentiment, which we are able to detect with significant confidence. Unlike other works, mentions to candidates revealed very poor predictive power, compared to sentiment-based features. Nevertheless, these results should be handled with caution, because the sentiment intrinsic in the comments needs to reflect the general public sentiment, otherwise the results may be biased by the group of comment's authors. Indeed, newspaper comments may be as limited as tweets. So this work should be regarded as a step towards a more general framework for analyzing behavior and examining sentiment-based prediction. Our experiments have revealed a different opinion expression behavior if compared to Twitter, and possibly, they represent a different population [9].

Opinion mining results are still not satisfactory. The unavailability of good lexicons is the main issue with the dictionary-based approach, and the annotation process for training data, an impediment for the machine learning approach in real settings. We need to improve the process by handling indirect mentions (e.g. pronouns), clauses, co-reference, negation and irony.

We are currently developing more experiments with new testing data, among them the second round of the 2010 and 2012 elections. We also plan to use the forthcoming 2014 Presidential and Gubernatorial elections to validate our results. In addition, we could experiment the proposed approach for other sparse governmental indicators, such as popularity or

government approval, census data per critical area (e.g. health, education), etc. As future work, we need to develop mechanisms to integrate sentiment expressed in various medias, each one with its own form of expression and media representativeness. Some of the issues that need to be solved are: what techniques are suitable to each media, and its underlying expression behavior; how to discover the representativeness of the population interacting through the media, and in which proportion the opinion should account for prediction; among others. Another important line of work is to address reactions to news beyond direct comments on the newspapers, such as repercussion on Facebook or Twitter.

## REFERENCES

- [1] B. Liu, *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers, 2012.
- [2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.
- [3] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," in *ICWSM*, 2010, pp. 59–65.
- [4] B. O'Connor, R. Balasubramanyam, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," *ICWSM*, vol. 11, pp. 122–129, 2010.
- [5] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the Fourth International aaai conference on weblogs and social media*, 2010, pp. 178–185.
- [6] Y. Liu, X. Huang, A. An, and X. Yu, "Arsa: a sentiment-aware model for predicting sales performance using blogs," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 607–614.
- [7] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2010 *IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2010, pp. 492–499.
- [8] H. Schoen, D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, M. Strohmaier, and P. A. Gloor, "The power of prediction with social media," *Internet Research*, vol. 23, no. 5, pp. 528–543, 2013.
- [9] D. Tuminan and K. Becker, "Tracking sentiment evolution on user-generated content: A case study on the brazilian political scene," in *Simpósio Brasileiro de Banco de Dados*. Anais do 28 Simpósio Brasileiro de Banco de Dados, 2013, pp. 139–144.
- [10] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 478–514, 2012.
- [11] M. Silva, P. Carvalho, and L. Sarmento, "Building a sentiment lexicon for social judgement mining," *Computational Processing of the Portuguese Language*, pp. 218–228, 2012.
- [12] E. Bick, *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press Aarhus, 2000, vol. 202.
- [13] L. Sarmento, P. Carvalho, M. Silva, and E. de Oliveira, "Automatic creation of a reference corpus for political opinion mining in user-generated content," in *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*. ACM, 2009, pp. 29–36.
- [14] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, vol. 2, 2007.
- [15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- [16] L. A. S. Romani, A. M. H. De Avila, D. Y. Chino, J. Zullo, R. Chbeir, C. Traina, and A. J. Traina, "A new time series mining approach applied to multitemporal remote sensing imagery," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 1, pp. 140–150, 2013.