PMKT
BRAZILIAN JOURNAL OF MARKETING, OPINION, AND MEDIA RESEARCH

Chief Editor: Fauze Najib Mattar | Associate Editor: Claudio Alberto de Moraes
Evaluation System: Triple Blind Review
Published by: ABEP - Brazilian Association of Research Companies
Languages: Portuguese and English

Index of candidates and voting intention. Can these indicators coexist?

Índice de sentimento de candidatos e intenção de voto. Podem esses indicadores coexistirem?

Rodrigo Otávio de Araújo Ribeiro*, Reinaldo Gomes Morais
IBOPE DTM, Rio de Janeiro, RJ, Brazil

Patrícia Pavanelli, Bruna Suzzara Bueno de Miranda
IBOPE Intelligence, São Paulo, SP, Brazil

ABSTRACT

This study aims to evaluate the relationship between the Sentiment index of the leading candidates on Twitter during the election campaign for the presidency in 2014 and the voting intention of Brazilians captured by IBOPE surveys on the same period. The use of more than one source of information in data analysis is one of the foundations of Big Data. It was found that metrics related to the volume of posts have higher correlation with the results of surveys than those based on Sentiment analysis. A Topic Analysis was also performed considering the periods immediately prior to the days of the elections, enabling a faster identification of subjects posted on Twitter about the campaign.

KEYWORDS: Presidential elections; Sentiment analysis; Twitter.

RESUMO

Este estudo tem como objetivo avaliar a relação existente entre o índice de sentimento dos principais candidatos no Twitter, durante a campanha eleitoral para presidência de 2014 e a intenção de voto dos brasileiros, captada pelas pesquisas realizadas pelo IBOPE no mesmo período. A utilização de mais de uma fonte de informação em análises de dados constitui um dos alicerces do Big Data. Foi verificado que índices relacionados ao volume de postagens possuem maior correlação com os resultados das pesquisas realizadas do que os que se baseiam na avaliação do sentimento. Uma análise de tópicos complementar também foi realizada em períodos imediatamente anteriores aos turnos da eleição, possibilitando a rápida identificação dos assuntos postados no Twitter sobre as candidaturas.

PALAVRAS-CHAVE: Eleições presidenciais; Análise de sentimento; Twitter.

Submission: May 19, 2016
Approval: August 23, 2017

***Rodrigo Otávio de Araújo Ribeiro**
Ph.D. in Industrial Engineering from Universidade Federal Fluminense. Marketing Intelligence Director at IBOPE DTM. Professor at the Instituto de Matemática e Estatística (IME) of the Universidade do Estado do Rio de Janeiro (UERJ).
(CEP 22270-000 - Botafogo, Rio de Janeiro, RJ, Brazil).
E-mail: rodrigo.ribeiro@ibopedtm.com
Address: Rua Voluntários da Pátria 45, sala 1308 - 22270-000 - Botafogo, Rio de Janeiro, RJ, Brazil.

Reinaldo Gomes Morais
Master's degree in Engenharia Eletrônica from the Universidade do Estado do Rio de Janeiro. Marketing Intelligence Analyst in the IBOPE TMD.
E-mail: reimorais1986@hotmail.com

Patrícia Pavanelli
Post-Graduate degree in Fundação Escola de Sociologia e Política de São Paulo. Director of accounts, public opinion, political communication and IBOPE Inteligência.
E-mail: patricia.pavanelli@ibopeinteligencia.com

Bruna Suzzara Bueno de Miranda
Graduated in Statistics from the Universidade Estadual de Campinas (UNICAMP). Statistical Coordinator in IBOPE Inteligência.
E-mail: bruna.suzzara@ibopeinteligencia.com

1 INTRODUCTION

This study aims to evaluate the relationship between the Sentiment index of the leading candidates on Twitter during the election campaign for the presidency in 2014 and the voting intention of Brazilians captured by IBOPE surveys on the same period. It was observed a greater relevance of the volume of posts in relation to the sentiment of the comments in the correlation generated with the rates of voting intentions from the surveys. This was a surprising result since a high correlation between the sentiment index and the voting intention rates was expected. Perhaps, this phenomenon could be explained by the lack of adherence of the Sentiment Model. However, it did not occur, since the performance of the model was satisfactory.

In the last few years, many studies based on information from social networks have been developed. However, there are few studies comparing results using social networks and traditional quantitative research. The use of more than one source of information in data analysis is one of the foundations of Big Data, in which not only the volume and speed of information are important, but variety also plays a key role for a clearer view of the subject of interest.

In The United States, O'Connor, Balasubramanyan, Routledge e Smith (2010) conducted a study with close characteristics. However, the same authors obtained a positive correlation between sentiment metrics captured via Twitter and the result of approval surveys in the Obama administration.

The results of this study serve as an aid to professionals and research companies that work in the political sector, in order to allow a richer assessment of the Brazilian electoral scenario in times of election. It was possible to understand the limitations and benefits of the information generated through the analysis of data from the social network Twitter in the electoral context, as well as the existing correlation with the traditional electoral indicators.

The results were obtained based on the information of the quantitative surveys of voting intentions captured by IBOPE and on the monitoring of the main candidates for the presidency of the republic in the 2014 elections: Dilma Rousseff, Aécio Neves, Eduardo Campos and Marina Silva (Marina Silva replaced Eduardo Campos after his death).

The sentiment analisys of the candidates on Twitter was made based on the methodology developed by IBOPE DTM. The algorithm performs the reading of the post containing the name of the candidates, to classify it as positive, neutral or negative.

IBOPE develops surveys of national voting intention with samples proportional to the number of voters from each region of the country. In this way, the voting intention for a given candidate is calculated by the proportion of people who declared that they would vote in the same candidate, if the election was the date of the research.

It is worth mentioning that IBOPE DTM has all the posts made on Twitter, made in Portuguese language with the names of the candidates, during the period studied. Posts were captured using the GNIP tool. The analysis of the main topics related to the candidates that were featured on Twitter, in each of the moments of the campaign in which the surveys took place, was also carried out.

2 ELECTORAL POLLS

2.1 Objectives and history in Brazil

Since Brazil's redemocratization in the 1980s, which allowed Brazilians to vote for a choice of their governors after years of military dictatorship, Brazil held seven presidential elections, the last in 2014. As in other democratic countries, the conduct and dissemination of voting intention quantitative surveys has become part of the context of the country's elections.

Opinion survey is a source of information about a population's general thinking about a country's social and political issues. In this context, voting intentions (political, electoral) are important and effective tools for knowledge of voters' opinion and behavior, and make it possible to understand how voters intend to vote within the social group.

In general, electoral polls always represent an instant of reality, a portrait of the moment. Like photography, the result of a research is an inert image of something that is constantly moving: opinion. Survey does not predict the future, it indicates trends that can be altered if something interferes with measured reality, causing it to change public opinion. Public opinion in this article is understood as the result of answers to interview questions (Lane & Sears, 1964, Converse, 1987, Price, 1992, Boyte, 1995, Worcester, 1997) in contrast to the concepts that show it as a deliberative process promoted by informed citizens and active participants in democratic life, as Speier (1950), Habermas (1998) and Bourdieu (1973) propose.

According to Marcia Cavallari, Executive Director of IBOPE Intelligence, in an interview with the site Congress in Focus:

> Survey is not infallible, it does not dictate the last word. It is one more information that the voter has among so many others for the decision making. More and more survey has to be interpreted as a diagnosis of the moment. They are a photograph of the moment. The sequence of these photographs is building a film with the trends. When we publicize the poll the day before, it does not mean that the process of vote consolidation freezes, that no one changes anymore. (Cavallari, 2016)

Currently, there are several proposals for laws to control the conduct of polls and their dissemination in the media. Since 1997, the Superior Electoral Court has regulated the dissemination of the results of electoral surveys, requiring that any survey to be disclosed be registered for publication after the deadline stipulated by the body. Throughout the year, the TSE registered 2,411 electoral surveys (Gramacho, 2014).

A pioneer in this type of research in Brazil, having begun the realization and dissemination of electoral surveys in 1945, IBOPE followed the seven post-dictatorship elections and, therefore, can be considered one of the most important and one of the most knowledgeable of Brazilian voter behavior.

IBOPE has been the country's research institute for a long time, responsible for measuring and disseminating the largest volume of electoral surveys in Brazil, most of which are reported in the most expressive Brazilian communication vehicle. With this importance and scope, the results are reflected by all other vehicles and commented by both the specialized critics and the general population.

2.2 Sample construction process

The national samples that are made by IBOPE Intelligence aim to reflect the opinion of the Brazilian electorate who voted in the last elections (voters).

When planning this type of study, we face the limitation / no update of the information that exists in the Superior Electoral Court registers. This information mostly reflects the characteristics of the voters at the time they obtain their voters' identification. Information such as age and level of education is not updated in these official databases.

In order to update the current profile of the electorate, we aggregate to data from the Superior Electoral Court, population estimates made by IBOPE Intelligence based on official data (the most recent is from Census and PNAD), as well as internal studies. This information helps when preparing the sample quotas.

The universe of voters is stratified by state, with the exception of the states of Acre, Amapá and Roraima, which together constitute only one stratum. Since the State has a Metropolitan Region, its universe is stratified in the Metropolitan and Interior Regions. Next, a sample of conglomerates is selected in three stages:

- At the first stage, municipalities are probabilistically selected through the Proportional Probability to Size (PPT) method, taking voters who voted in the last (voter) election as the basis for such selection;
- At the second stage, the conglomerates are selected: census tracts, with systematic PPT. The measure of size is the number of voters of the sectors;
- In the third stage, in each conglomerate, a fixed number of voters according to gender, age, education and activity condition are selected.

3 TWITTER

Twitter was created in 2006 by partners Jack Dorsey, Evan Williams, Biz Stone and Noah Glass, in San Francisco - USA. The service is a social network that allows users to post and read tweets, which are nothing more than messages of up to 140 characters. Its access can be done directly in some internet browser, by mobile applications and, in some countries, the posts can be made by through SMS. The idea quickly spread and gained worldwide popularity: in 2012, there were over 500 million registered users posting 340 million tweets per day (Lunden, 2012).

Once registered, the user sets an address on the site that is not yet in use; from then on, it will always be known by that address preceded by the @ symbol by other users.

Once this address is defined and the account is registered, the user can follow or be followed by other accounts. This means that, whenever users post in a row, the message appears directly on your page (also called a timeline). By default, tweets are publicly viewable, however, you can restrict viewing of messages to only your followers. Another possibility of sending a message is to repost what has already been posted by someone, a practice also known as retweet, and which is characterized by the acronym RT. The purpose, in this case, is for the user to pass on that particular text to all who follow it (Strachan, 2009).

When a posting is made about an specific topic, the user can make use of a technique called *hashtag* - phrases or words that begin with the # symbol (Strachan, 2009). Likewise, if the interest is only to view messages from that topic, a search can be done using the same *hashtag* term.

4 2014 PRESIDENTIAL CAMPAIGN

The 2014 election, the seventh Brazilian presidential race since redemocratization in the 1980s, was the most fierce the country ever had. Dilma Rousseff was re-elected with 51.6% of the valid votes, in the second round, which represents the tightest victory since the end of the Military District, evidencing the latent desire of the voters for the change in the conduct of the Brazilian government.

Surveys made by IBOPE indicated that by 2014, about 70% of voters wanted the next president to completely change the programs and measures of the Federal Government or to keep only some of them. This index is only lower than that observed in the 2002 surveys, the year of Lula's election. This scenario was quite different in the 2010 election, in which the desire for continuity prevailed for six out of ten voters.

After years of economic growth during Lula's administration, Dilma Rousseff's first mandate reflects a period of stagnation rather than progress, marked by a combination of rising prices, falling purchasing power, household indebtedness and tariff readjustments public policies. The general discontent of millions of Brazilians was marked by the demonstrations of July 2013 and accentuated by the cost of constructions for the World Cup in Brazil, occured in July 2014.

After the protests in 2013 against the increase of bus fares, the 2014 World Cup, the low performance of the Brazilian economy and for improvements in public services, the electoral atmosphere of the presidential race was, for voters, not very interesting and a intense disillusionment with politics.

These aspects can be observed in the results of Figure 1, which shows a significant increase in the percentage of Brazilians who declare that they would not vote if the vote was not mandatory (from 35% in 2010 to 50% of voters in 2014).

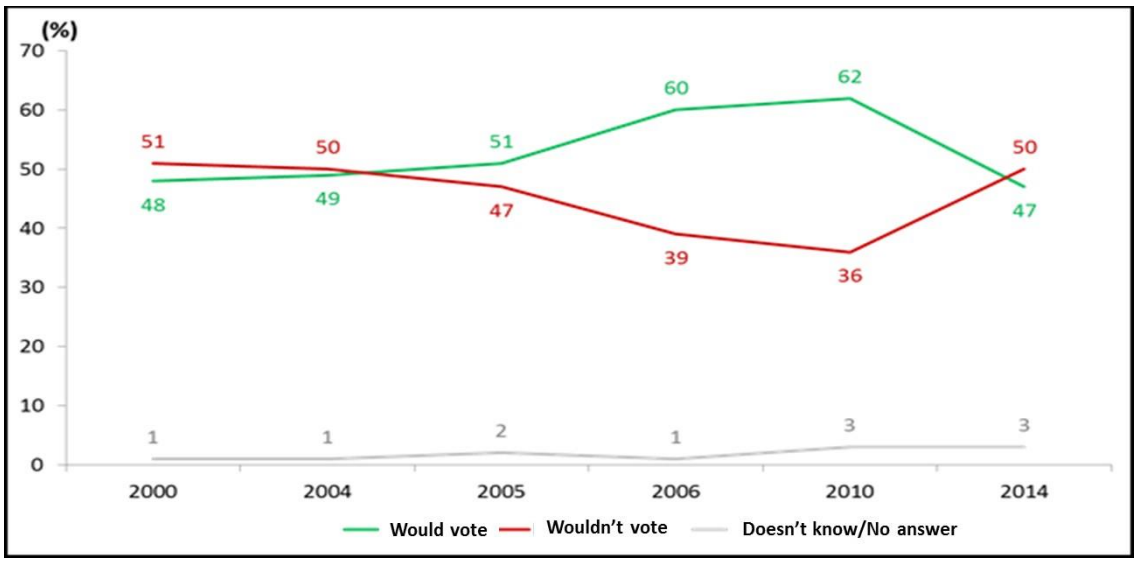

Figure 1 - Evolution of the opinion on the voting condition, if it was not mandatory

With the officialization of candidacies, the campaign was consolidated with three main candidates: President Dilma Rousseff for the Workers' Party (PT), Senator Aécio Neves for the Brazilian Social Democracy Party (PSDB) and the former Governor of Pernambuco Eduardo Campos by the Brazilian Socialist Party (PSB). However, on August 13, the electoral campaign registered a major tragedy: the death of PSB candidate Eduardo Campos. The candidate was aboard his campaign plane, along with four assistents and two pilots who also lost their lives. The accident moved the country and transformed the dynamics of the electoral dispute. In third place in the polls, Campos was replaced by Marina Silva, with deputy Beto Albuquerque from PSB of Rio Grande do Sul as deputy.

As can be seen in Figure 2, Dilma Rousseff led the way in all IBOPE surveys in the first round. From the moment she became a candidate, Marina Silva, who finished the election in 3rd place, performed well until the eve of the election. When attacked by opponents for her contradictory statements and proposals, she lost her voting intentions and could not sustain herself until the end. Aécio Neves, who failed to capture Marina's votes as fast as she fell, was repositionated in second place after the last election campaign debate on the eve of the election. The official result of the first round established, once again, the dispute for the Presidency of the Republic in the second round between PT and PSDB.

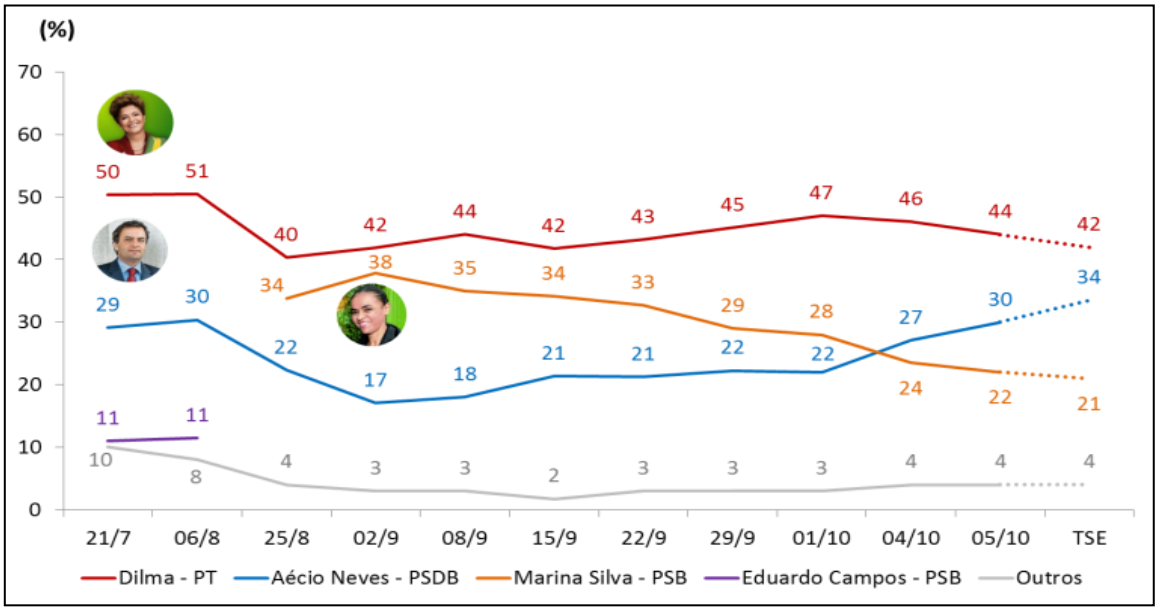

Figure 2 - Evolution of voting intentions for president and Official Result of the 1st Round of Presidential Elections - Valid votes

It was in the second round of the campaign that the intensification of the dispute between the candidates became more potent. The first surveys presented Aécio Neves numerically ahead of Dilma, but five days after the election the situation was reversed. President Dilma Rousseff surpassed the PSDB candidate and won the election with 51.6% of the valid votes, as mentioned earlier, the tighter victory since the country's redemocratization.

5 ANALYTICAL METHODOLOGY

The analytical methodology applied in this article consists of the execution of four steps: the first refers to the analysis of the adjustment quality of the Sentiment Model (in relation to the candidate) developed; the second seeks to evaluate the general behavior of users and their profile regarding the use of Twitter to make postings about politics, and also a brief description of the profile of the Brazilian voter; the third concerns the analysis of the correlation between voting intentions and indexes from Twitter; in the fourth step, a semantic analysis is made, which is based on the use of Text Mining techniques to identify the most pertinent topics within the conversation environment of the presidential elections.

5.1 Sentiment model (in relation to the candidates)

5.1.1 *Text Mining*

Text Mining is the process of extracting useful information or knowledge from structured or non-structured text documents (Barion & Lago, 2008). In the context of this article, this technique will be applied to identify patterns of comments and opinions from Twitter users about the presidential candidates of the 2014 National election.

The first step of mining is indexing, a process that stores an index structure from the words of the texts and enables the search for documents through all the terms contained therein (Salton & McGill, 1983). According to Barion and Lago (2008), some important steps for a Text Mining analysis must be followed and are:

- Lexical Analysis: converts a sequence of characters into a sequence of words that will be considered to be used as a term for the index. At this stage, the input alphabet is separated into word characters and word separators;
- Stopwords removal: removes a set of words that appear frequently in texts, but do not have semantic value, such as: prepositions, articles and conjunctions. This stage is extremely important because it lowers the base to be indexed and facilitates mining;
- Stemming: removes all variations of words, leaving only the essential, for example, the word "we love" (“amamos” in Portuguese) is identified as the radical "loves" (“ama” in Portuguese);
- Selection of index terms: determines which words or radicals will be used as indexing elements. These words are selected according to the weight assigned to them;
- Bag of words (BOW): consists of a matrix in which each different term present in the collection of documents is indexed. From this indexation, each document can be represented by a vector *1 x n*, where *n* is the total number of terms, each entry of that vector will be the number of times the terms appear in that document (Sivic, 2009);
- Determination of Weights: BOW matrix padding is based on metrics that weight the frequency of terms in documents and in the total collection (set of all documents). The metric most commonly used for this purpose is called *tf-idf (term frequency - inverse document frequency).*

5.1.2 Sentiment identification

The Sentiment Index serves as a thermometer of the polarity of messages posted about politics on Twitter. It is a comparative metric that allows the evaluation of candidates at a given moment, as well as their evolution over time.

The elaboration of the algorithm occurred in four stages:

- Step 1 – Development of a polarization methodology made by specialists in Portuguese language;
- Step 2 - Manual polarization in the sample following the methodology developed in step 1;
- Step 3 - Development of the algorithm for automatic polarization of policy postings based on the polarized sample from step 2;
- Step 4 - Implementation, error evaluation and recalibration based on new collected samples.

The developed method combines human knowledge with sophisticated statistical models to evaluate the polarity of a particular post. Sentiment analysis models have the following characteristics:

- Ability to handle large volumes of data with high speed of response;
- Guarantee of the same evaluation criteria and the same rule applied to all posts.

The manual marking of polarity by human agents has the following characteristics:

- When well trained, the classifier can capture the interpretive nuances of the text;
- Inability to handle large volumes of data at high speed;
- Difficulty to traine polarizers team to maintain a homogeneous classification pattern in markings.

A study conducted by TOPSY in the USA, with three classifiers polarizing a sample of 10,000 political-related posts, obtained a concordance of 73% (Gosh, 2016). A similar study was carried out in IBOPE DTM comparing the same postings by different people and a 59% agreement was found in the classifications, considering the classification of the specialists with that of a higher level classifier. The importance of the consistency of the manual marking that will feed the model is an essential and critical aspect of the sentiment analysis methodology developed.

Sentiment analysis algorithms attribute different degrees of importance to the words that appear in the posts, symbols or expressions present according to their frequency of occurrence in the evaluated messages. These degrees of importance or weights are obtained during the process of statistical modeling that aims to elaborate rules for a correct classification of manually polarized messages.

The developed algorithm considers, for its calibration, only posts classified by specialists that are related to the candidates, therefore, the linguistic terms have their measured importance considering the correct connotation. To illustrate this point, observe the different polarities that can be attributed to the word *encoberto* (covered) in the following sentences: "*Foi evidente que o governador havia encoberto o esquema de lavagem de dinheiro*" ("It was evident that the governor had covered the money laundering scheme"); *"Mesmo com o céu coberto de nuvens nesta manhã de terça-feira, não há a possibilidade de chuvas fortes*" ("Even with the sky covered by clouds this Thursday morning, there is no possibility of heavy rains.")

In the first sentence, it can be observed that the word "covered" associated with the "political" subject has a negative polarity. Already when it is related to the subject "meteorology", it can present positive polarity. So if a statistical model were designed to mark the polarity in the "political" chat environment it was adjusted based on a sample of posts related to "weather", it would not generate consistent information.

For this reason, it is essential that the models of Sentiment analysis be calibrated taking into account samples associated with the context that one wishes to model.

The algorithm developed marks a Sentiment score on all postings made on the considered candidates. The Sentiment score is a measure that ranges from 0 to 100, the closer to 0, the more negative the message is, and the closer to 100, the more positive it is. When the score assumes values over 60, the postage is considered positive, neutral between 40 and 60 and negative under 40.

5.1.3 *Random Forest*

The developed Sentiment classification method uses as a classifier the Random Forest model, developed by Breiman (2001). It consists of the joint use of multiple random decision trees (Random Trees) aiming the improvement of the classification. According to the author, among the positive characteristics of this method are:

- Excellent accuracy when compared to other classification algorithms;
- Good performance in large databases;
- Ability to deal with thousands of predictor variables, without the need for deletion;
- Ability to tell which variables are most important for classification;
- Generation of an uncorrupted estimate of the error through the process of forest construction;
- Efficient proposal of estimation with missing information, when this problem exists in the database to be analyzed;
- Possibility of balancing the error in unbalanced databases;
- Possibility of using forests to forecast future samples;
- Calculation of the proximities between pairs of cases, information that can be used in outliers detection or in data clustering.

In order to verify the accuracy of the markings made by the Sentiment Analysis Model, post samples about the candidates were extracted (during the election period) to be manually polarized by trained professionals in the manual polarization methodology developed by the Portuguese language specialists. Their job consisted in evaluating whether or not the automatic model marking was right.

5.2 Evaluation of the evolution of posts and profile of users who interact with the campaign by Twitter

In the first analytical stage, the objective is to evaluate the main added metrics, present in the work, grouped in time. The most importants were:

- Total number of posts: evaluates the total number of posts made by time interval;
- Number of posts per user;
- Post Penetration: percentage of posts with a given characteristic of interest.

The analysis of the total number of posts makes possible the evaluation of the intensity of the impacts occurred during the observed period. The penetration of posts evaluates the weight of the existence of a certain characteristic in the universe of considered posts.

The evaluation of these metrics serves to understand the general behavior of users about a particular conversation environment. Peak identification is done by viewing the time series of the number of posts.

Twitter allows the use of specific metrics that denote the different types of behavior of its users, among them, penetration is emphasized. The evaluated characteristics are:

- RT - Tweets that passed on a message that had previously been posted by another user;
- @ without RT - Directing messages to another person, which was not a retweet;

- HTTP - Tweets that have information contained in websites;
- HASHTAG (#) – Discussion group about some specific subject.

5.3 Evaluating Correlations

The evaluation of correlations is an essential analysis to verify the relationships between different indicators. This study sought to use Pearson's correlation coefficient to identify the relationship between the voting intention and the Sentiment index about the candidates on Twitter.

O'Connor et al. (2010) suggest that the correlation between survey indexes and Twitter information should be evaluated considering the moving average in the information vector coming from the social network. Such a procedure is more justified by the greater variability of the temporal data coming from Twitter than by the information coming from the political survey. However, in this article, the correlation was performed directly, since the amount of points available for the evaluation is greatly reduced due to the limitation of the amount of surveys. The study sought to elaborate a Sentiment index on Twitter with an idea close to that of an election, considering as votes the positive posts about a candidate. Therefore, the proportion of positive posts was the first indicator of interest to be tested. The calculation was made considering the performance of the three main candidacies: PT, PSDB and PSB. It was considered, in this study, the candidacy of Marina Silva and Eduardo Campos, both of the alliance led by the PSB party, as a single candidacy. The intentions of the three candidacies total 100%. The same idea was applied in relation to the proportion of posts. Another important point is that, for the calculation of the correlations, only those postings that had only a single candidate mentioned were considered.

5.4 Semantic Analysis

For the analysis of the subjects commented by the users on politics, the analysis of topics was used. The Latent Dirichlet Allocation (LDA) model is a generalized probabilistic structure for the modeling of sparse matrices of counting data, such as the matrices used in Text Mining (Bag of words). The main idea behind this algorithm is that the words in each document are generated by a mix of themes (topics).

According to (Chein, 2016), the LDA represents the documents (in this case, posts) as a mixture of topics, in which each word is allocated to a topic with a defined probability. This model assumes that each post is produced in this way: when the author (user) writes a document he decides:

- First, the topics to be written;
- Then choose the words to write about the topics (according to a multinomial distribution);
- The number of N words that the document should contain (Poisson distribution);
- The amount of topic mixes that each document should contain (distribution of Dirichlet by the topical K). In this case, a document may contain more than one different topic;
- The generations of each word in the document.

The topic analysis model LDA seeks to recursively obtain the probability that a collection of topics has generated the collection's documents. The estimation of the parameters that compose the model is done by the Collapsed Gibbs Sampling method.

The analysis of the correlation between topics was done according to the following process: in the first moment, the lexical analysis was performed. In the second moment, the Stopword clean was made (words without semantic value), for later execution of the algorithm of stemming (extraction of radicals). After these steps, the BOW matrix was calculated. In this matrix, each term considered corresponds to one column, and each row to a document (tweet). The measure used for evaluation was the tf-idf (*term frequency inverse document frequency*).

The postings related to waves 9 and 14, were made immediately before the election shifts, evaluated through the Topic Analysis using the LDA model.

LDA modeling allows the identification of the main words related to each of the topics. By the interpretation of these main words belonging to each topic, the interpretation of its meaning is made.

6 Data Collection

For this article, we considered all Twitter posts about the candidates considered on the days on which the searches were conducted. A total of 14 waves were performed: 10 in the first round and 4 in the second. Table 1 shows the relation of days in which each one of the waves was made.

Table 1 – Considered time intervals

Round	Wave	Survey Application Date
1º	1	18 to 21/07/2014
	2	03 to 06/08/2014
	3	23 to 25/08/2014
	4	31/08 to 02/09/2014
	5	13 to 15/09/2014
	6	20 to 22/09/2014
	7	27 to 29/09/2014
	8	29/09 to 01/10/2014
	9	02 to 04/10/2014
	*10	5/10/2014
	11	07 to 08/10/2014
2º	12	14/10/2014
	13	20 to 22/10/2014
	14	24 to 25/10/2014

3,096,032 tweets were collected as a whole about candidates Dilma, Aécio, Eduardo Campos and Marina at the mentioned intervals. All of them were classified according to the Sentiment analysis model developed. However, for analysis of correlation were considered 2,388,300, since they contained the name of only one candidate.

As for the surveys, the data was collected following the sampling plan developed by IBOPE Intelligence.

7 DATA ANALYSIS

7.1 Sentiment Analysis Model Accuracy

In Table 2, the historical series of monthly monitoring of the adjustment metrics obtained for the Sentiment Analysis model can be seen.

Table 2 - Sentiment analysis model adjustment quality

Metric	jul-14	aug-14	sep-14	oct-14
Recall	0.77	0.72	0.75	0.79
Precision	0.77	0.78	0.81	0.84
Accuracy	0.73	0.71	0.72	0.74
F-measure	0.77	0.75	0.78	0.82

The Recall metric represents the percentage of true positives among the total of false negatives and true positives. The Precision is relative to the proportion of true positives among the total of false positives and true positives. The Accuracy metric represents the total hits among the total number of

possible cases. Finally, the F-measure is the harmonic average between Recall and Precision. Figure 3 shows the confusion matrix.

Confusion Matrix		Real	
		Yes	No
Predict	Yes	True Positive	False Positive
	No	False Negative	True Negative

Figure 3 – Confusion matrix

Comparing these results with those obtained by Araújo, Gonçalves and Benevenuto (2013), it can be concluded that the assertiveness levels of the elaborated Sentiment Analysis algorithm are even higher in relation to a large number of solutions available in the market for English language because it has a number of studies developed much superior to those of Portuguese language.

7.2 The evolution of posts about candidates

During the analyzed period, there was an increase in the number of postings between the first wave and the last, with the peaks occurring in waves 9 and 14, days immediately preceding the elections, as can be seen in Figure 4.

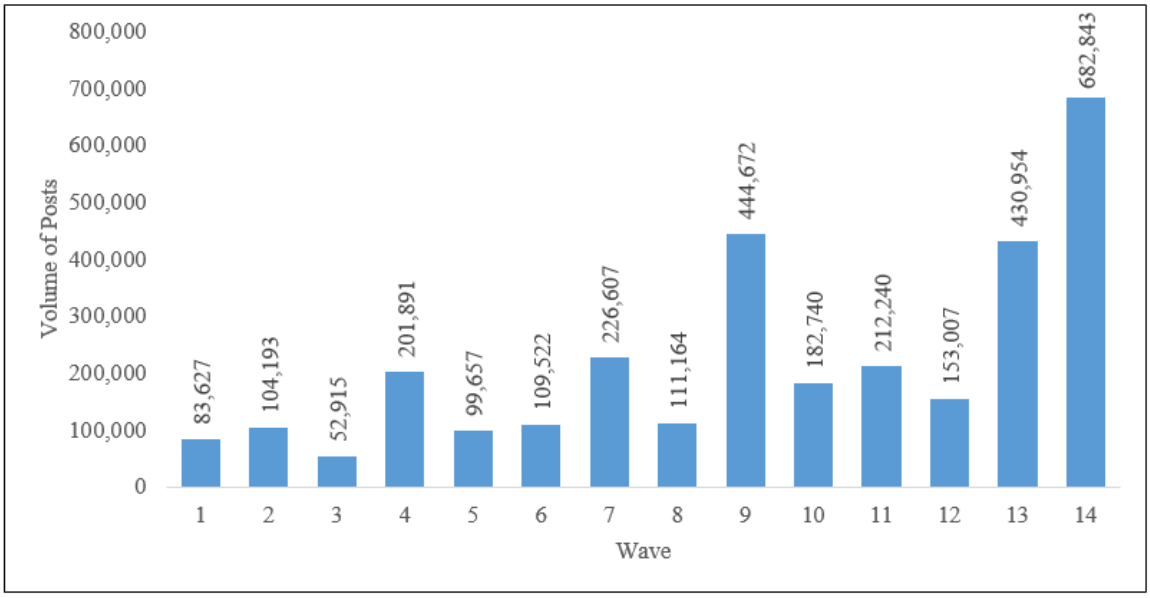

Figure 4 - Evolution in the number of posts about candidates

Figure 5 shows an increasing tendency in the proportion of retweets, 62% in wave 1, reaching 76% in wave 14. This high proportion throughout the historical series shows a strong transfer of information, which is a characteristic of the political subject on Twitter: few generate information and many pass on. The proportion of hashtags (#) also shows significant increases varying from 15% to 48%, showing the evolution of popularization of the electoral theme on Twitter. The proportion of HTTP is more representative in the first three weeks and then decreases. However, it never presents values below 26%, average of 43% between the 4th and 14th wave, a value that can be considered high, indicating the presence of posts that direct information to existing sites. The proportion of "@ without RT" shows a certain regularity throughout the period.

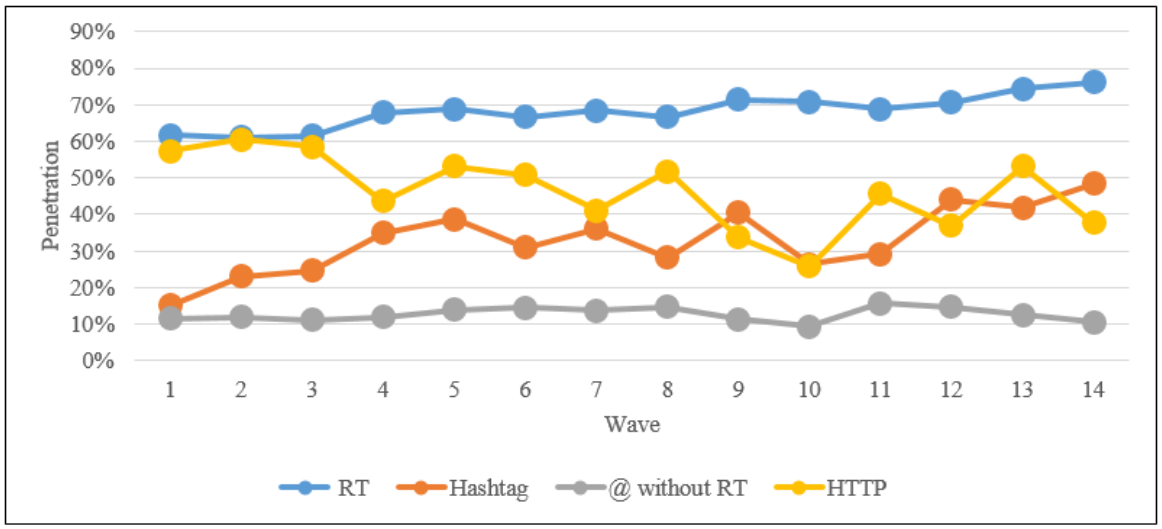

Figure 5 – Twitter metrics evolution

7.3 Profile of users interacting with the campaign by Twitter

Among the users who spoke about politics on Twitter during the considered period, it can be seen that 52% made a single post. Users who made more than 50 posts (9,420 users) were responsable for 48% of total posts and 70% of impressions (Table 3).

Table 3 – Post quality distribution per user

Number of Posts	Users	%	Posts	%	Impressions	%
1	243,127	52%	243,127	8%	260,616,867	3%
2	77,280	16%	154,560	5%	222,381,564	2%
3	37,898	8%	113,694	4%	170,847,033	2%
4	22,561	5%	90,244	3%	120,361,596	1%
5	14,835	3%	74,175	2%	95,079,530	1%
6 to 10	33,088	7%	248,606	8%	396,183,962	4%
11 to 20	19,271	4%	278,360	9%	622,388,884	6%
21 to 50	12,577	3%	393,974	13%	1,085,334,096	11%
51 or more	9,420	2%	1,499,292	48%	7,032,772,640	70%
Total	**470,057**	**100%**	**3,096,032**	**100%**	**10,005,966,172**	**100%**

By ranking the users by the volume of impressions (number of posts x amount of followers), as shown in Table 4, it can be seen that the user “dilmabr” (official of the PT candidacy) was the one that generated more impressions, having performed 436 Twitter postings during the considered period. News agencies users appear in fifth place, such as *G1*, *O Globo* newspaper, *R7* news portal and Veja Magazine.

The first celebrity appears in sixth place, presenter Danilo Gentili, who made 43 posts about the candidates in the period.

The user "silva_marina" (official of the Marina Silva’s candidacy) appears in the thirteenth position. The lack of engagement of the candidate Aécio Neves on Twitter as a content creator can be pointed out as one of the factors explaining the great difference between his and Marina’s posting proportion at the beginning of his campaign, but with the proximity of the elections his name gains a greater impact.

It is worth to notice that the number of followers considered in Table 4 refers to the considered period in the analysis. Currently, volumes are larger.

Table 4 - Top users by impression volume

Rank	UserID	User Name	Followers	Posts	Impressions
1	89826	dilmabr	2,335,703	436	1,018,366,508
2	11435	g1	2,801,001	141	394,941,141
3	12091	JornalOGlobo	1,493,103	198	295,634,394
4	14448	portalR7	3,070,130	88	270,171,440
5	20450	VEJA	3,322,907	79	262,509,653
6	19511	DaniloGentili	5,917,726	43	254,462,218
7	11926	Estadao	1,244,307	150	186,646,050
8	64886	PastorMalafaia	741,810	220	163,198,200
9	14937	folha_com	1,322,629	122	161,360,738
10	147778	DaviSacer	398,124	397	158,055,228
11	11423	Val_Ce1	152,675	740	112,979,500
12	11720	TerraNoticiasBR	634,838	172	109,192,136
13	16836	silva_marina	838,921	127	106,542,967
14	21531	drangelocarbone	1,627,457	59	96,019,963
15	269254	rodrigovesgo	5,049,476	19	95,940,044
16	368839	lobaoeletrico	262,346	296	77,654,416
17	14274	UOLNoticias	454,990	169	76,893,310
18	13806	cartacapital	476,360	118	56,210,480
19	47350	felipeneto	2,732,792	17	46,457,464
20	10771	massavcs	144,700	319	46,159,300

7.4 Voter profile

Brazilian voters who answered to IBOPE's surveys are mostly women, 52%. They have a higher concentration in the 25-34 age group, 25%. High school is the predominant degree of education, 39%. On the other hand, primary and elementary education together account for 42% of all voters. The surveys' samples have national coverage with a minimum size of 2,002 interviews, which corresponds to an estimated sampling error of 2 percentage points.

In relation to the distribution by region, the differences between the Brazilian voters and the posts on the three candidatures in the period are verified.

It is worth noting that not all postings captured contain their geolocation, only 36% have this information. After all, not every post is issued by means of portable devices that allow this identification.

Assuming that the geographical distribution of the geolocated posts is the same as those that were not geolocated, the comparison of Table 5 with the research's public was made.

It can be verified that the differences were not that discrepant. Twitter users who posted messages about the analyzed candidates are arranged next, in the same order of magnitude, in relation to the distribution of the number of voters. The largest difference, in terms of representativeness, are the proportions of the Northeast and Southeast region.

Table 5 – Posts and voters distribution by region

Region	Survey	Twitter	Total
Southeast	43%	51%	50%
Northeast	27%	20%	21%
South	15%	15%	15%
Midwest	8%	9%	9%
North	7%	5%	5%
Total	100%	100%	100%

7.5 Historical correlations

The correlations between the historical series of voting intentions and the proportion of positive posts for the three applications were evaluated and the positive correlations were verified in all cases. However, the same phenomenon is observed when the indicator is calculated by negative polarity

posts, which means the higher the negativity, the greater the candidate's intention to vote. We conclude that the correlation is high, regardless of the polarity of the post. Analyzing the historical series of the representativeness of the candidate Dilma in the realized postings, it is concluded that its proportion is practically the same, basing the aforementioned conclusion.

It is worth remembering that the representativeness shown in Figure 6 is the comparative among the candidates. For example, on wave 1, the candidate Dilma had 85.2% of the total positive posts, 84.2% of the total negative posts, even value among the neutral ones. It can be seen that close values occur in practically all waves. This same phenomenon occurred with the other candidates.

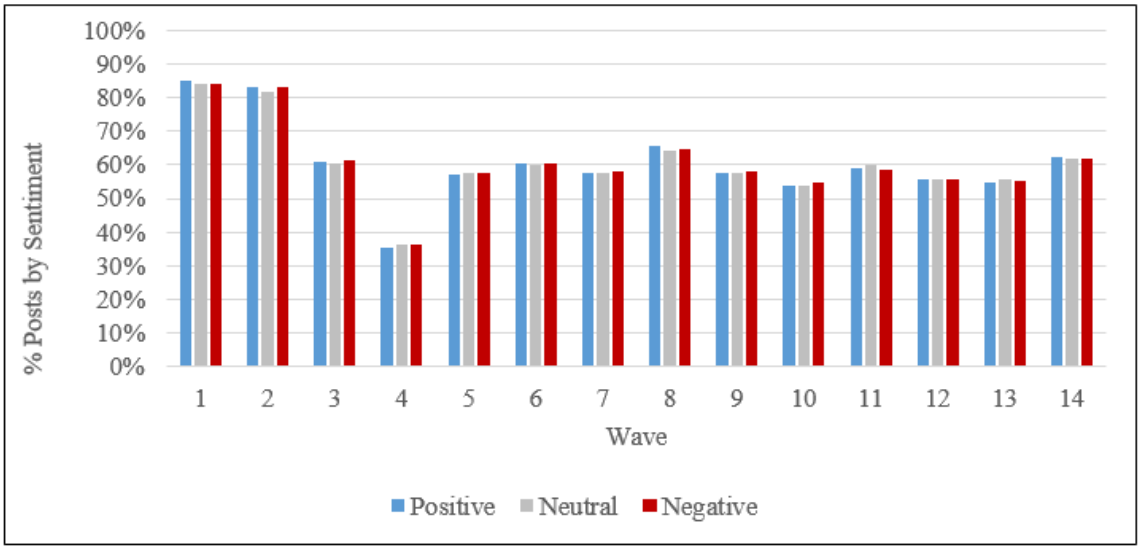

Figure 6 - Representativeness of the candidate Dilma in the total of posts by Sentiment

Based on the acquired knowledge, the correlations are evaluated based on the total number of performed postings, not only the positive ones. Initially, considering the 10 waves of surveys done in the first round, the correlation of the candidate Aécio approaches zero, as it had high rates of voting intention and low proportion of posts in the two initial waves, when the presidential campaign was still "tepid". However, if considered only the waves that occurred after the death of Eduardo Campos and the launch of Marina Silva's candidacy, it is verified that the correlation would increase to 0.65. Evaluating the first and second rounds together, with the proportions related only to Dilma and Aécio, there are high correlations. Considering all the history, the correlation was 0.66 for both (Table 6).

Table 6 - Pearson Correlation

Period	Aécio	Dilma	Eduardo/Marina
10 Round - 10 Waves	0.01	0.77	0.95
10 Round e 20 Turno -14 Waves	0.66	0.66	

Analyzing the first round historical, Dilma, the current president, presented a high proportion of postings and voting intentions in the first two weeks, possibly because she had the most well-known name, with posts proportions far higher than those of voting intentions. Already in the survey of election exit poll (wave 10) it is possible to see that the indicators are very close (Figures 7, 8 and 9).

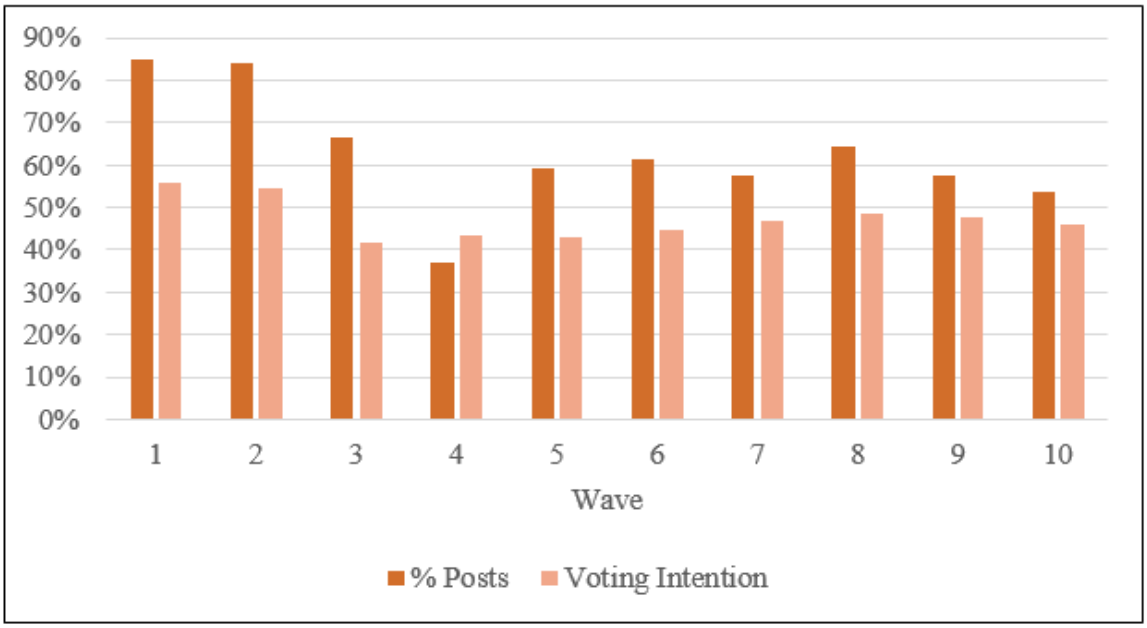

Figure 7 – Candidate Dilma's first round percentage of posts and voting

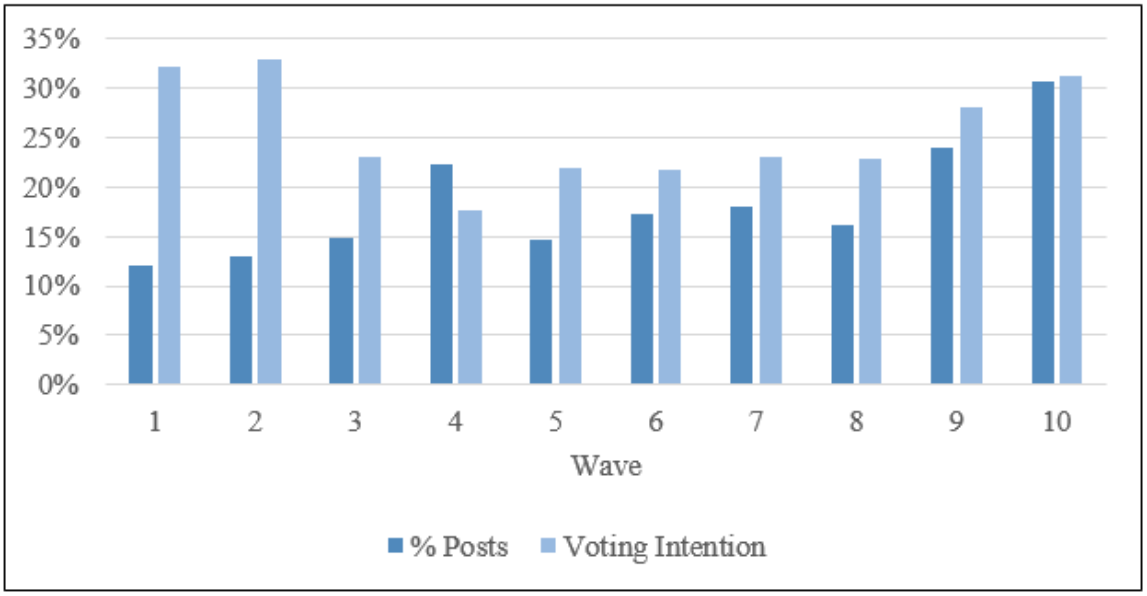

Figure 8 - Candidate Aécio's first round percentage of posts and voting

Aécio Neves showed a much lower participation than Dilma on Twitter in the first round of the elections. Especially in the first two weeks.

In Figure 9, that refers to the application of Eduardo Campos / Marina Silva for the first round, it is observed that the information of Twitter behaves close to that observed in the surveys, having its peak observed in Wave 4, the second survey after the launch of Marina's candidacy, which declined in the following weeks.

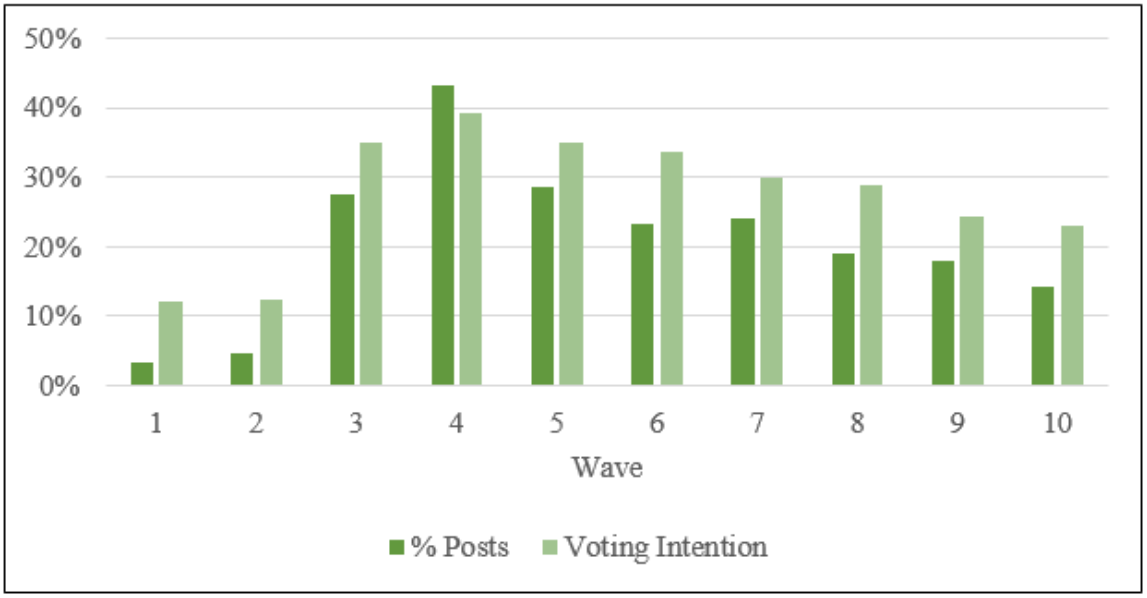

Figure 9 - Candidates Eduardo Campos/Marina Silva's first round percentage of posts and voting

Considering the statistics of the three applications together (10 x 3 = 30 points) in the first round, we obtain a correlation of 0.92 between voting intentions and proportion of Twitter posts Generating a simple linear regression model, it was possible to verify that the increase of 1 percentage point in the representativeness of the candidate in relation to the others on Twitter generated, on average, an increase of 0.4783 percentage points in the candidate's intention to vote during the first round. The regression was significant, with inner p-value at 0.001. Of course, this model is an approximation, not being necessary to capture small differences between candidates during the electoral process. However, it demonstrates the strong relationship between the metrics during the campaign.

By doing the same evaluation, but considering only Aécio and Dilma (together) for the 14 waves, we obtain similar results. A model with R2 of 0.83 (Figures 10 and 11).

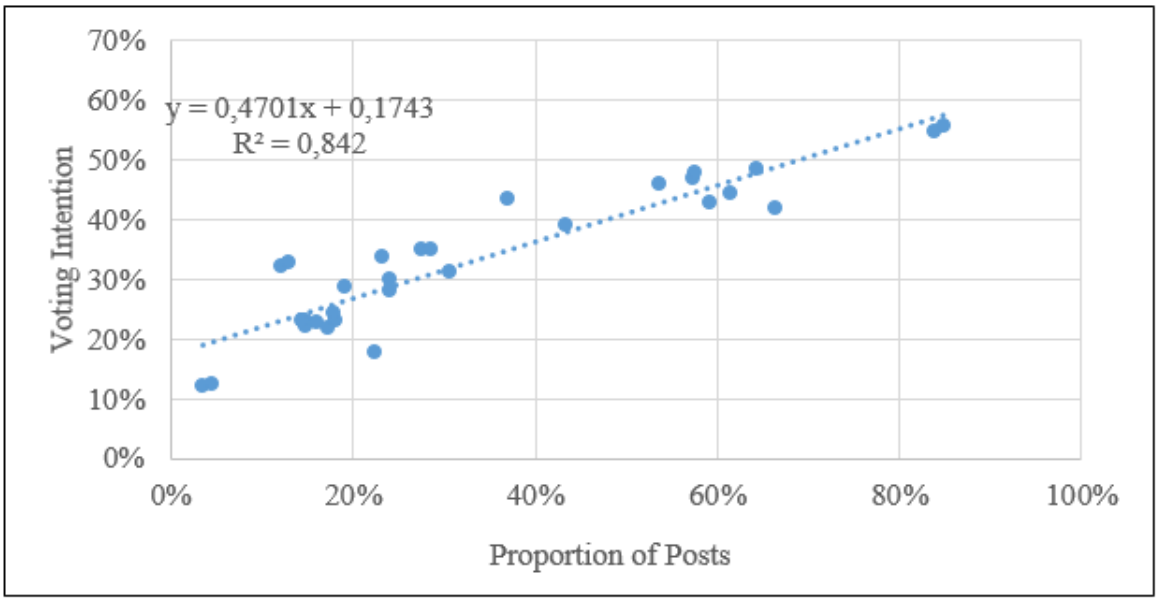

Figure 10 - Simple Linear Regression - First Round - Three candidacies

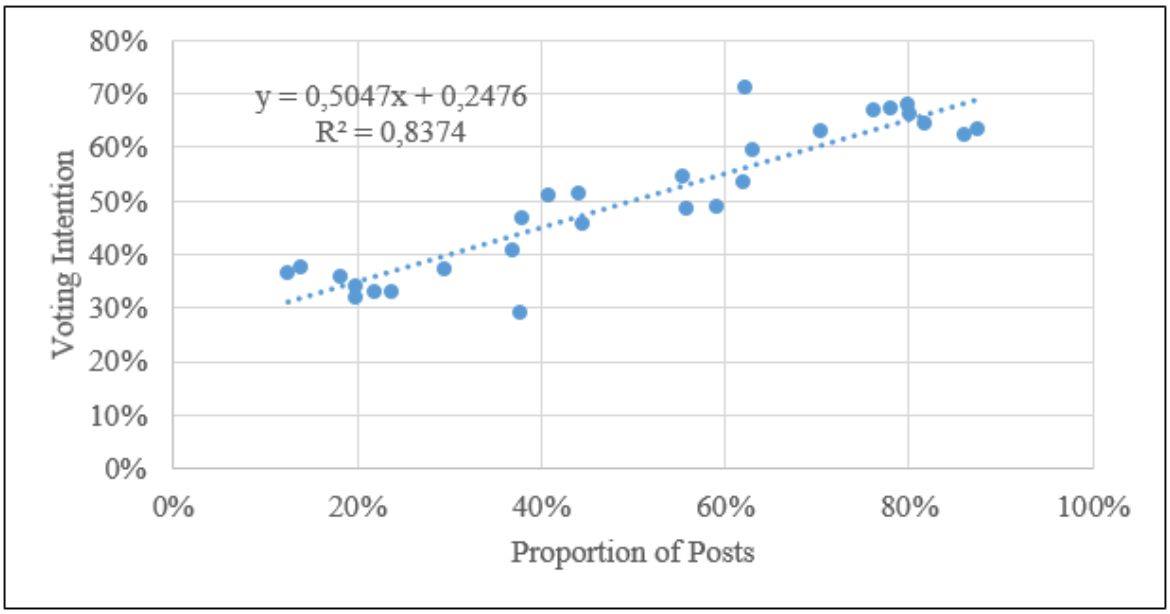

Figure 11 - Simple Linear Regression - First and Second Rounds - Two candidacies

The analysis of the graphs was made considering only the votes of Dilma and Aécio (obtained in the surveys) and posts of the same candidates in Twitter. It can be seen that the trend is the same, Dilma begins with a much larger proportion and this proportion tends to converge with that of Aécio in the second round. This result is another indication of the existing relationship (Figure 12).

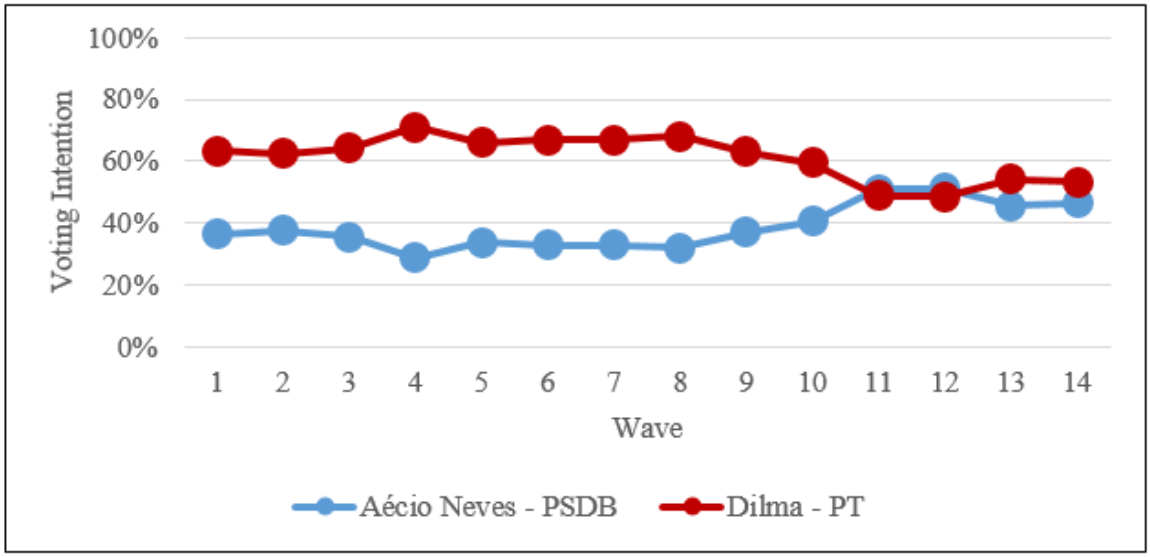

Figure 12 - Aécio and Dilma's evolution – Survey

The biggest difference in relation to reading occurs in wave 4, when the proportion of posts of Aécio rises on Twitter, but decreases in relation to its representativeness when compared to PT's candidate in the surveys of voting intentions (Figure 13).

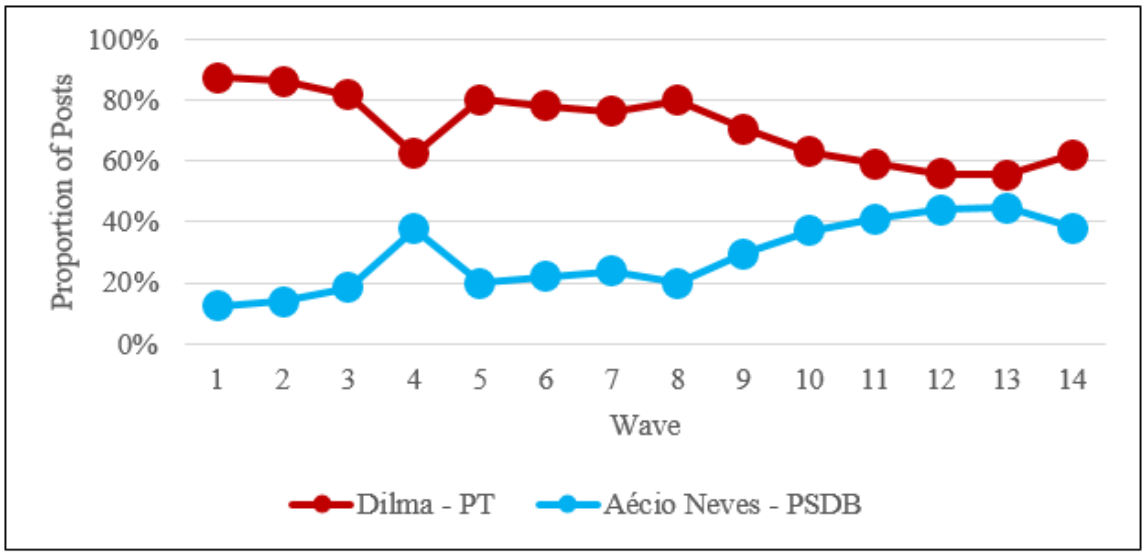

Figure 13 – Aécio and Dilma's evolution - Postings

7.6 Related topics

The analysis of topics was done considering a random sample of 20 thousand posts for waves 9 and 14 (10 thousand for each one), on the eve of elections' first and second rounds. It was decided to identify five topics in each one. The choice of topics was based on the interpretation of the results. For each topic, the 15 most relevant words were selected for identification.

Evaluating Figure 14, it is possible to verify the most representative words of the topics of wave 9. The first topic contains, predominantly, posts of support to the candidate Aécio Neves, denouncing scandals involving the Workers Party's (PT) candidate. In the second topic, the reasons for choosing the candidate Marina Silva were more present. In the third, Dilma appears as a central figure, with inflation control being the most prominent argumentation of her defense, and there are also strong negatives to her candidacy, such as the postings made by the member of the church Assembly of God, Silas Malafaia (#chegaderoubalheiraforadilma). In the fourth topic, appearances related to other candidates, such as Luciana, Levy and Pastor Everaldo appear. The fifth topic relates, mainly, speculative posts about the result of the elections, also mentioning Channel Globo's debate.

Wave 9: Topic				
1	**2**	**3**	**4**	**5**
aecioneves	**40**	**dilmabr**	**marina**	**marina**
dilmabr	**silvamarina**	**chegaderouba...**	silva	**turno**
45aecioconfir...	domingo	**presidenta**	**luciana**	**presidente**
corrupção	**votar**	vai	pra	neves
pt	conheça	novo	**levy**	aécio
sobre	**razões**	**marina**	neves	aecioneves
aécio	httptcoaz...	pastormalafaia	pergunta	ser
petrobras	neste	pra	candidato	silva
diz	vou	**pt**	**pastor**	**segundo**
correios	**brasilmarina40**	**inflação**	aécio	pode
minas	**marina40**	13	**everaldo**	**debate**
neves	dia	**controle**	corrupção	pois
educação	fazer	**dilma**	noite	**pesquisa**
fhc	dias	presidente	falar	**globo**
frase	econômica	**povo**	**vai**	qualquer

Figure 14 – Wave 9's topics

Wave 14, as shown in Figure 15, the first topic was related to posts from users who supported Aécio and those who supported Dilma. Aécio's defenders accused Dilma of irregularities related to the BNDES loan to the port of Cuba. However, the allies of Dilma, emphasized the construction of

schools to have been superior in PT's government in relation to the government of the PSDB. In the second topic, they highlighted posts related to the request for the right of reply from Dilma's alliance to Veja Magazine. In the third topic, posts from the user @ OGloboPolítica evaluate whether the candidate's statements in the debate were true or not. In the fourth, it was possible to verify postings of repudiation to the candidate and support to the candidacy of Aécio Neves. In the fifth topic, the support of the soccer player Neymar to the candidacy of Aécio Neves and the comparison of Aécio Neves to the ex-president Fernando Henrique were the highlights. One can observe the pertinence of the topics considered, through the interpretation of results. LDA analysis was performed using program R.

Wave 14: Topic				
1	**2**	**3**	**4**	**5**
governo	**veja**	vida	brasil	neves
anos	**tse**	vai	**mensalão**	presidência
quer	**revista**	tirar	**danilogentili**	candidato
aecioneves	eleitoral	pra	pastormalafaia	momento
brasil	lula	brasil	presidente	qualquer
vcs	jornaloglobo	oglobopolitica	é	ser
somostodos...	**pedido**	somostodos...	nunca	pode
psdb	**dilma**	eleição	pra	neymar
cuba	**resposta**	**pretonobranco**	**elessabiam...**	preso
escolas	**direito**	governo	mineiro	aécio
maior	nega	corrupção	corrupto	chamo
brasileiros	critica	eleitor	mudança	fernando
educação	terrorismo	checa	foradilma	henrique
esconder	fundadora	debate	aecio45pelo...	aecio
porto	justiça	debatenaglobo	pq	eaecio45...

Figure 15 – Wave 14's topics

8 CONCLUSIONS

The pertinence of the information from Twitter as an important complementary source of analysis to the electoral surveys is verified, presenting as main advantages the less granularity of time and possibility of interpretation of results almost instantaneously. However, the proportion of posts can not and should not be used to estimate the proportion of voters of a given candidate. For this, the surveys present much more coherent results, because in them the respondent is exposed to objective questions in which he chooses the candidate in which he has more affinity considering a certain established scenario. Such procedure does not exist in social networks, in which the user has no limits to the exposure of his ideas and opinions.

Another relevant aspect is that the distribution of posts and voters in the Brazilian regions was close to similar. This fact, in no way, can indicate that there is a similarity in relation to the other sociodemographic aspects. However, it is an interesting indication that there is a need for further in-depth investigation.

In relation to the profile of the users who made postings about the candidates in the considered period, it can be said that, each one has, on average, 1,382 followers, being this distribution quite asymmetric, since the average number is of 183 followers. Medias are important disseminators of the political information, since they were among the most retwitted users.

As for the Sentiment analysis, it is verified that the developed Sentiment model manages to capture in a coherent way the polarity of the posts. However, the information generated by him is not related to the variations in the voting intentions captured in the IBOPE surveys, as the music of the group "Charlie Brown Jr." would say: "Say good things, say bad things, but talk about me".

Regarding the analysis of topics (LDA model), it is possible to identify pertinent topics capable of offering a quick interpretation of the information presented. Solutions based on this type of modeling may provide a faster assessment of political news in the upcoming elections.

9 LIMITATIONS AND SUGGESTION FOR NEW RESEARCH

It's considerable the fact that correlations are made considering a very small number of points, due to the limitation of the number of Brazil's polls conducted and reported. For this reason, joint evaluations of the candidates were carried out in order to increase the robustness of the conclusions presented.

It is strongly recommended that studies like this be conducted in future elections, in order to verify if the conclusions obtained will be maintained. The expectation is that the correlations become larger in quantity, due to the increased access of the Brazilian population to the Internet and, consequently, the social network Twitter. However, this is a hypothesis that must be checked.

The use of voting intention information from other research institutes may also be a promising option for future analysis.

REFERENCES

Araujo, M., Gonçalves, P., & Benevenuto, F. (2013). *Métodos para análise de sentimentos no Twitter*. In Proceedings of the Simpósio Brasileiro de Sistemas Multimídia e Web (Web media).

Barion, E. C. N., & Lago, D. (2008). Mineração de textos. *Revista de Ciências Exatas e Tecnologia*.

Bourdieu, P. 2000 [1973]. La opinión pública no existe. *Cuestiones de Sociología*, 220-232. Madrid: Istmo.

Boyte, H. C. (1995). Public opinion as public judgement. In T. L. Glasser, & C. T. Salmon, (Eds.). *Public Opinion and the Communication of Consent*. Nueva York: The Guilford Press, 417-436.

Breiman, L. (2001). *Random Forests. Machine Learning*. *45*(1), 5-32. doi*:* 10.1023/A:1010933404324

Cavallari, M. (2016). *Congresso em Foco*. Entrevista. Recuperado de http://congressoemfoco.uol.com.br/noticias/marcia-cavallari-%E2%80%9Cpesquisa-nao-e-infalivel%E2%80%9D/

Chein, E. (2016). *Introduction to latent Dirichlet allocation*. Retrieved from http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

Converse, J. M. (1987). *Survey research in the United States: Roots and emergence, 1890-1960*. University of California Press, Berkeley, California.

Gosh, R. A. (2016). *Social media for giant instant opinion polls: Twitter political index.* Retrieved from http://sentimentsymposium.com/SS2012w/presentations/SAS12w-RishabGhosh.pdf

Gramacho, W. G. (2015). Surveys pré-eleitorais nas eleições brasileiras de 2014: Erros, acertos e polêmicas. *REB - Revista de Estudios Brasileños*, Primer semestre, *2*(2), 115-113, Madrid: Universia.

Habermas J. (1998). *Facticidad y validez: Sobre el derecho y el estado democrático de derecho en términos de teoría del discurso*. Madrid: Editorial Trotta.

IBOPE Inteligência. (2016). *Avaliar a relação existente entre o índice de sentimento de candidatos no Twitter, durante a campanha eleitoral para presidência de 2014 e a intenção de voto dos principais candidatos*. Pesquisa. Recuperado de http://www.eleicoes.ibopeinteligencia.com.br

Lane, R. E., & Sears D. O. (1964). *Public opinion*. Englewood Cliffs: Prentice Hall, 13.

Lunden, I. (2012). *Analyst: Twitter passed 500m users*. In June 2012, 140M of them in US; Jakarta 'Biggest Tweeting' City. Retrieved from https://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/

O'Connor B., Balasubramanyan, R., Routledge B. R., & Smith, N. A. (2010). *From tweets to polls: Linking text sentiment to public opinion time series.* Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

Price, V. (1992). *Communication concepts 4: Public opinion*. Newbury Park: Sage Publications.

Ribeiro, R. O. A., Tavares, T. G. B., & Cohen, D. O. (2014). Análise de usuários que conversam sobre cerveja no Twitter. *PMKT - Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia*, *14*, 174-195.

Salton, G., & Mcgill, M. J. (1983). *Introduction to modern information retrieval*. Computer Science Series, USA: McGraw-Hill.

SamPedro, V. (2000). *Opinión pública y democracia: Medios, sondeos y urnas*. Madrid: Istmo.

Sartori, G. (2002). *Elementos de teoría política*. Madrid: Alianza Editorial.

Sivic, J. (2009). *Efficient visual search of videos cast as text retrieval*. Transactions on pattern analysis and machine intelligence, *31*(4), IEEE.

Speier, H. (1950). Historical development of public opinion. *American Journal of Sociology*, *55*, 376-388.

Strachan, D. (2009). *Twitter: How to set up your account.* Retrieved from http://www.telegraph.co.uk/travel/4698589/Twitter-how-to-set-up-your-account.html

Tribunal Superior Eleitoral. (2016). Número de candidatos para o cargo de Presidente da República em 2014. Recuperado de http://www.tse.jus.br/

Worcester, R. (1997). Public opinion and the environment. In M. Jacobs (Comp.). *Greening the Millennium? The new politics of the environment*. Oxford: Blackwell Publishers.