Cury *J Big Data* (2019) 6:42
https://doi.org/10.1186/s40537-019-0208-1

Journal of Big Data

RESEARCH Open Access

Oscillation of tweet sentiments in the election of João Doria Jr. for Mayor

Rubens Mussi Cury[1,2*]

*Correspondence:
rubensmussicury@gmail.com
[1] Fundação Getúlio Vargas, Edifício John F. Kennedy, Av. Nove de Julho, 2029-2º andar, Bela Vista, São Paulo, SP 01313-902, Brazil
Full list of author information is available at the end of the article

Abstract

The purpose of this work is to identify and analyze the oscillation of sentiments expressed by users of the Twitter social media through their direct replies to posts by user @jdoriajr that took place before, during and after the elections for mayor of the city of São Paulo in the year 2016. In order to make this research possible, we used Python 3.6.4 and the Searchtweets 1.6.1 library for consumption of the API Search Twitter, from which it was possible to extract 76,690 tweets. Text sentiment analysis was carried out through the Lexicon-Based Approach method and the Laplacian Smoothing calculation algorithm-which generated a rate that would represent a negative and a positive sentiment ranging from −0.1306 (minimum) to 0.1489 (maximum) respectively, throughout the observed period. As additional tools, WordCloud and t-SNE (t-Distributed Stochastic Neighbor Embedding) Corpus Visualization were used for visualization of the word cloud and cluster, respectively, with both functionalities available at the Yellowbrick 0.8 package also for Python.

Keywords: Sentiment analysis, Twitter, João Doria Jr., Data mining, Big data in politics

Introduction

In this article, we will address the representativeness of democracy in social media, its importance in politics and the ways data can be extracted, from Twitter in this case, so that it can be interpreted through data mining methods and sentiment analysis algorithms. The purpose of this study is to demonstrate and analyze the sentiment expressed in Twitter users posts about São Paulo City former mayor, whose username in this social media is @jdoriajr on specific dates corresponding to important events in this user's trajectory.

Social media representativeness in democracy

A difficult problem known as the "bubble effect" is a reality for social media users, as well as for internet search engines, which usually direct users to find facts representing the preferences of a dominant group.

It is known that in many places in the world the influence of social media in the political scenario is a major factor regarding the way a country expresses its votes in the ballots; however, experts have not been able to establish precisely how this factor can be analyzed. One example partially evidenced through a study carried out in 2015 by students from three highly-regarded US universities, among them the University of

Michigan, suggests that over 60% of Facebook users completely disregard the information directed to them through the social media's feeds, but, on the contrary, they choose to believe in feeds coming from friends and people they follow [1].

An article published in The Guardian and The Computational Propaganda Project of the Oxford Internet Institute maintains that Facebook, for example, does not inform how potential automated accounts—the notorious bots—may act to change an online debate in certain posts, which prevents the study by specialists in this field. In addition, it has been noted that most part of the content shown to Facebook users, instead of originating primarily from what they liked, commented or shared and are interested in, are actually posts promoted monetarily by other users [2].

Hern [2] further suggests that the inexistence of any type of regulation for automations by these algorithms favoring a specific post over another may make a post with 25,000 shares, whether by real users or not, reach a much higher circulation in social media—operating similarly to other search engines, such as Google, and guaranteeing that a page will be displayed at the top of the results. In the political context, a definition by Halavais [3] states that competition between discourses may be restricted with the use of search classification algorithms that favor the access to certain information, for example.

In view of the questions previously addressed in terms of the manner algorithms seem to work, we come up with several indicators that clearly suggest the need for caution when proclaiming the democratic potential of social media, particularly if we consider that the existence of dominant content both in the social media and the search engines disproportionally influence the views of users in an authoritarian way [4].

Considering this issue further, would it be possible to conclude that the daily use of social media could be linked to a limited potential of democratic innovation? For Amartya Sen, professor of Economics and Philosophy at Harvard University, a democracy relies on free information flow. Nonetheless, wherever you go, there will always be conflicts between those who want to freely share their ideas, music, films, etc., and those in search of finding a way to control these activities—whether through private institutions seeking profit or government agencies fearing the debate and democracy [5].

The text mining process and the discovery of information

In addition to the great technological advance in the hardware and software sectors, the importance of text mining has drastically increased in the past years also as a result of the connectivity of the various applications we use, which tend to generate a massive amount of data [6]. Differently from an ordinary research process through an internet search engine mechanism—which returns information from the content generated by thousands of users—the purpose of the text mining process is to find out through data standards—which may come from various sources—information unknown until then, and, thus unavailable on an information basis, as in an ordinary search process.

In fact, the technique used to discover these data patterns depends on the way they are organized and on what one wants to find out. Even though there are several data analysis techniques, the most renowned ones in the corporate field are data mining and text mining. In spite of their similar techniques, whereas Data Mining does not interpret texts in linguistically, the text mining technique does it through natural language processing (NLP) [7].

It is important to highlight that data mining technique focuses on the analysis of structured data—usually originating from databases—whereas text mining involves the analysis of unstructured or semistructured data—for example, an internet page, a document, an email, etc. According to Herschel [8], Merrill Lynch estimates that over 85% of corporate information is not structured.

In view of these facts, it is possible to merge to the conclusion that this massive set of unstructured data has significant commercial potential. If we analyze spoken or written natural language as unstructured information developed to assist communication between us, it is clear we will have better understanding of slang, orthographic variations and contexts. However, in spite of our ability to natively understand natural language, we are at a disadvantage compared to supercomputers in terms massive data volumes, analysis speed and processing. This is the goal of the text mining technique [9].

The text mining technique consists basically of the structuring of a certain input parameter—whether unstructured or semi-structured—through several analysis in order to enable the identification of patterns to determine an expected result. This process usually comprises various stages, which can be combined by following the execution flow below:

- Information retrieval (IR)
- Natural language processing (NLP)
- Data mining (DM)
- Information extraction (IE)

Information retrieval (IR)

This process identifies items in a collection that may somehow correspond to what is being searched. A typical instance of this process in action may be found in library research systems, which basically aim to locate records of books, articles, journals and newspapers based on keywords entered by the user that contain the word sought [10].

When the collection analyzed is not too big, the application of this process becomes quite simple—requiring only the identification and listing of the existing words—but, when dealing with an extensive search universe, exhibiting items with a certain word could become completely impractical. That explains why this process should be combined with the others in order to identify not only the words, but also their relevance in the information sought about the user.

Natural language processing (NLP)

Most internet users that adopt Google as their search engine were probably caught by surprise on the day the incorrect typing of a word, as well of its context, was immediately corrected and shown at the top of the results preceded by the text "Did you mean…?". This a clear application of NLP.

Understanding the real meaning—particularly the context of what we say—may be easy for us humans; however, this has always been a challenging question for computers. Today, with the advances in technology and artificial intelligence, human language comprehension by computers has become a reality. One example may be found in the

various "Home Assistants" commercially available, such as Alexa from Amazon, Siri by Apple and, more recently, the amazing Google Duplex, which can converse with humans and understand them through voice [11].

Data mining (DM)

Once completed the Information retrieval (IR) stage of an unstructured or semi-structured input parameter and the value standards in a set of data are detected, the results obtained become structured in a way to be directly consumed from a database.

From this point on, the Data Mining procedure is applied to make the managing of these data possible in such a way to allow the application of the most diverse analyses.

Information extraction (IE)

The IE process is entirely connected to the NLP explained in Section "Natural language processing (NLP)". First, the linguistic interpretation process needs to be applied in a way that the resulting structured data can be extracted according with the desired purpose.

Analysis components

This chapter aims to explain the main analysis components to be carried out in this paper. This section briefly presents the Twitter social media platform, which will be used as an unstructured data source—presenting a general view of the Twitter target user and explaining the way the sentiment analysis algorithm is used.

The Twitter social media platform

Used by celebrities, politicians and companies form almost all over the world, Twitter has turned into an excellent communication opportunity, considerably more active among these entities and their followers, be they admirers, fans, critics, and many others. There are countless conveniences identified through the reactions of thousands of users of this social media platform, since they represent an external feedback from people connected to a specific subject, and thus can provide, on news of any nature, the most diversified opinions, given the number of people and the freedom they have to express themselves.

In a nutshell, for users to receive news of their interest, all they have to do is follow other users generating content through their posts on Twitter. From then on, all news will be shown to all of those following a certain user. That explains why the reaction of the followers of a certain user matters so much—if they are following a specific user, it means that they are interested in knowing what is going on with him or her. Similarly, if the posts of a certain user become uninteresting, his or her followers only have to unfollow that user to stop receiving updates.

The @jdoriajr user in Twitter

The @jdoriajr user in Twitter is the indirect generator of all the content that will be analyzed in this work. We assign him the role of indirect generator of content because only the replies of his followers to his posts will be analyzed.

This social media user was entrepreneur João Doria Jr., who became a public figure in October 2006, when he was elected to the office of mayor of the city of São Paulo. From

that moment on, the Twitter social media platform started to be used by him and his team as one of the main tools to publicize activities of his mandate, as well as interact with followers and voters in a more active way in terms of freedom of expression by the user (@jdoriajr)—through his posts—and more intensely by his followers through their respective replies to his posts or simple mentions to them.

Figure 1 presents a timeline of the trajectory of João Doria Jr. in the past 2 years as a political figure.

Sentiment analysis

The sentiment analysis technique is a very popular method of text classification and it aims to reveal valuable information extracted through unstructured data by means of the natural language processing (NLP) explained in Section 2.2. Differently from a textual analysis to discover gender by classifying a certain word as having been written by a person of the female or male gender, sentiment analysis goes beyond that because the meaning of a certain unstructured piece of data is influenced by its context.

A certain sentiment cannot be configured as a language characteristic as gender is, for example, because in spite of being possible to recognize the sentiment expressed by a certain isolated word as positive, negative or neutral, when we employ it within a context, this sentiment can change completely [12].

If we hear or read the phrase "my beautiful home", we have a positive sentiment without the need to think too much. Similarly, if we read or hear something as "my home is terrible", we likewise have a negative sentiment. In both examples our minds implicitly captured the positive and negative sentiments in the words "beautiful" and "terrible", respectively. On the other hand, if we questioned the meaning of the phrase "my beautiful old home", maybe we would understand a certain neutrality of sentiment. Finally, there are also cases of phrases such as "my home is not old and terrible, but beautiful", which combine negations that can reverse a meaning.

In spite of the several meaning variations considered in the previous examples, it is possible to identify the predominant feelings by mapping word order in phrases through the various sentiment analysis methods that will be discussed in the next section.

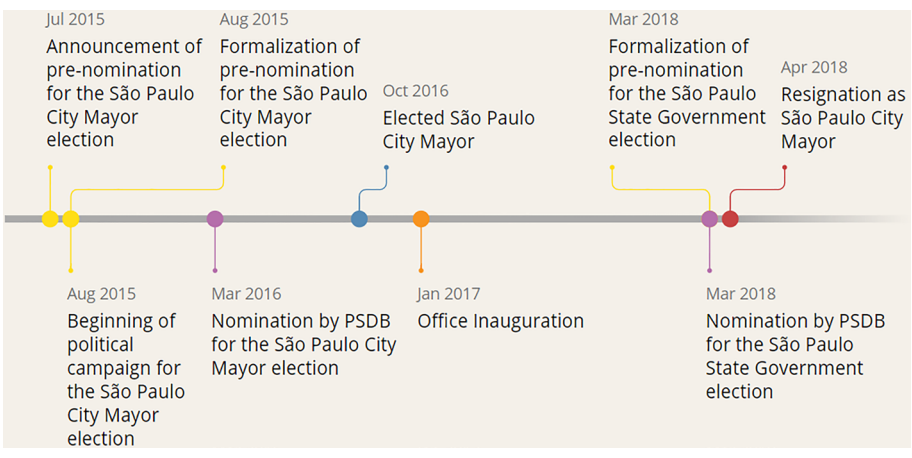

Fig. 1 Time line of the João Dória Jr. public figure (Source: prepared by the author)

Methodology

The data gathered for the sentiment analysis was carried out with Python 3.6.4 language and Searchtweets 1.6.1 library developed especially for Twitter API Search consumption in the Premium mode. A total of 77,179 tweets were collected in Brazilian Portuguese in replies to posts by @jdoriajr user from June 1, 2015 to July 31, 2017. Tweets containing links, media (video, sound, etc.) and images were not considered in the process—remaining only the content in text characters with up to 128 characters (including spaces) and at least one mention to a Twitter user (to @jdoriajr) and hashtags (when available).

The return through Twitter API Search in JSON format contained, in addition to the texts by Twitter users, several other attributes, including post date and time, which was also used in the analysis to identify user sentiment variation along a period of time of analysis of their posts.

Data cleansing

The cleansing process and preparation for analysis of the tweets collected was carried out in stages. The first step consisted of eliminating any links from the tweet content, since tweets exceeding 128 characters are always accompanied by a link at the end to allow its online access, despite the fact that requests made through API filter potentially existing links. The next step involved removing mentions to social media users from the tweet—first because they are of no interest for the analysis proposed in this work, and second, all collected tweets are replies to posts by the @jdoriajr user, who is always mentioned in the posts, making references to other users unusable for this study.

The third stage of the process included correcting possible typographical errors, such as "j o ã o", corrected simply to "joão". Lastly, we applied a filter to allow only Brazilian Portuguese non-numeric characters in the tweet content transformed into capital letters and to discard 489 tweets with duplicate users, creation date and text.

Correction and validation

In addition to the cleansing procedures previously described, some validations, corrections and improvements were required to be made in the content of the collected tweets. Because we are dealing with content from social media, it is very common to find abbreviations, typos and use of words unknown in the formal language universe.

In order to validate the grammar of the collected words, we used as reference the Python nltk.corpus package version 3.3 This is a corpus with words used in the work by world-renowned Brazilian writer Machado de Assis, in addition to over one million words from journalistic texts extracted from ten sections of Brazilian periodical Folha de São Paulo in 1994, and over 10 million words of the "Floresta Sintáctica" project—a set of sentences (corpus) also in Brazilian Portuguese analyzed (morpho)syntactically.

Adding up all words of the sources described, we obtained a reference base in Brazilian Portuguese with over 4.5 billion of nonunique words that, when grouped, were reduced to approximately one million distinct words. From then on it was possible to continue the validation process of words typed in the tweets, checking which ones could not be found in this reference base. The next step was to find out how many times the

words typed in the tweets with no correspondents in the reference base appeared in all tweets collected. Through this verification we were able to find a relation with almost 16 thousand words, slangs or phrases typed one to over 1500 times in the tweets. Later on a validation criterion was defined to evaluate only words typed 50 times or more, reducing this number to 120 words and classifying them as "to correct", for cases of abbreviations, accents, Brazilian Portuguese errors, etc., "to ignore" for expressions such as "kk", "aff", rs, etc. or unknown names; and "to analyze" for occurrences that may represent positive or negative sentiments.

For all cases of words classified as "to correct", certain adaptations were made for typing or abbreviation errors by applying formal Portuguese grammar rules, allowing words that qualified for sentiment analysis to be duly identified when applicable. Words classified as "to ignore" were kept unchanged since they did not have any representation for the analysis method to be used. As for the words classified as "to analyze", we obtained a random set of 20 tweets out of the total amount to check possible sentiments each one of them could represents and, subsequently, we classified them as "positive" or "negative".

Analysis method

Considering the use of the grammar words used in the collected tweets, since they are typically used in social media, in addition to the political content for this case, it was not possible to obtain a reliable database for training in terms of employability of a Machine Learning algorithm in order to enable the automatic extraction of resources corresponding to positive and negative sentiments. Even though the Python packages, such as TextBlob, offer film review database for training of the classification models using Naive Bayes, we ruled out the possibility of using them in view of the issues previously addressed—particularly with regard to the grammar used and the content.

For the sentiment analysis of the collected tweets, we chose to use the Lexicon-based approach, one of the most traditional and efficient analysis methods for this purpose. We obtained a list entitled "A list of English positive and negative opinion words or sentiment words" (Hu and Liu, KDD-2004) with approximately 6800 words, slangs and phrases in English, which were reduced to 4300 (1410 positive words and 2899 negative words) as a result of the lack of correspondent meanings in Brazilian Portuguese. Because the words obtained were in English, the translation of each word was needed using Google Translator and, subsequently, all translations were checked. Since we worked with translations of isolated words, the tool chosen for this task was considered reliable, causing minimum correction events, such as the need to interpret phrases "2-faces" as "duas caras"—and not "2-caras".

In addition to the words obtained, 27 others (2 positive and 25 negative) were added to their respective lists as described in Table 1, since some of them are considered slangs or political words.

The Lexicon-based approach basically generates a score that can be referred to as "positivity degree" for each tweet by associating the words used in the tweet with the respective lists built of positive and negative words. The following tweet examples show the use of the word "congratulations", "well" and "respect", which find correspondents in the list of positive words. In the second example, we find the use of the words "steal" and "poor", which find correspondents in the lists of negative words.

Table 1 Analysis of unknown words in formal language. Source: prepared by the author

Word	Total	Classification	Word	Total	Classification
Prefake	441	Negative	Marqueteiro	80	Negative
Coxinha	292	Negative	Prefeitaço	76	Positive
Bosta	176	Negative	Petralha	72	Negative
Golpista	118	Negative	Periferias	68	Negative
Mortadela	114	Negative	Petezada	60	Negative
Esquerdopatas	104	Negative	Poxa	59	Negative
Decepcione	101	Negative	Ptralhas	58	Negative
Comunas	97	Negative	Apadrinhado	57	Negative
Pichação	97	Negative	Pilantra	55	Negative
Almofadinha	95	Negative	Prefeitão	53	Positive
Mimimi	84	Negative	Piram	52	Negative
Fake	81	Negative	Riquinho	52	Negative
Petralha	81	Negative	Caviar	50	Negative
			Esquerdalha	50	Negative

Positive tweet

"Congratulations Mr. Mayor your work shows that there are good people you have my respect".

Negative tweet

"You have so much money and want to steal more from us who is poor and is taking in".

Once identified the words with correspondents in the sentiment lists developed, the Lexicon-based approach calculates the "positivity degree" of the tweet using the Laplacian Smoothing method through the following equation:

$$\begin{aligned}\text{Positivity\_degree} \;=\; & \text{float}(\text{len}(\text{positive\_words}) + 1) \\ & / \,\text{float}\big(\text{len}(\text{positive\_words}) + \text{len}\big(\text{negative\_words}\big) + 2\big) - 0.5\end{aligned}$$

.

The table below shows the comparison between 20 random tweets classified by the Lexicon-based approach and manually, in order to evaluate the accuracy of the method used (Table 2).

As can be observed, the Lexicon-based approach correctly classified 15 of the 20 tweets analyzed, indicating an accuracy of 75%. Nevertheless, only 2 of the 5 tweets with classification errors were categorized with opposite sentiment (Negative/Positive and Positive/Negative). Like most sentiment classification models, the Lexicon-based approach also has limitations in identifying contexts involving irony and sarcasm, for example.

Table 2 Comparison of sentiment classified by Lexicon and manually

Tweet	Lexicon	Manual
Olá senhor prefeito eu não sabia que ser oportunista era retratar a vdd queria que o senhor apenas respeitasse a arte	Negative	Negative
Mano será que você pode pelo menos mandar aparar a grama do anhangabaú grata	Neutral	Negative
Saia fora deste entre no e leve uma multidão com vc	Neutral	Negative
Aproveita e derruba o minhocão	Negative	Negative
Já que a amazon gosta tanto da arte pichações poderia convidar os meliantes para decorar a sede deles	Negative	Negative
O joao dorio por vc nao fala mais do leite sas criancas das escolas	Negative	Positive
Que tal instalar sensores de presença com monitoramento de empresas de segurança para avisar quando pichadores chegam no local	Neutral	Neutral
Espero que o novo prefeito não acabe com o emprego dos cobradores de ônibus	Neutral	Neutral
Meu respeito e cumprimentos sr prefeito	Positive	Positive
Boa sorte prefeito	Positive	Positive
Sou prefeito de caldas novas goiás e li sua entrevista no correio braziliense prefeito joão dória parabéns	Positive	Positive
Essa vírgula está errada alcaide você não é jornalista	Neutral	Neutral
Gcm podia trabalhar tirando os brandidos da rua esses covardes lixos chutar caxhorro morto é fácil	Neutral	Neutral
Por favor deem uma olhada na av roberto marinho e entorno diversas construções irregulares embaixo do monotrilho	Positive	Positive
Mais importante do que a cor da tinta éq ela não foi de graça público indo p lata de lixo além do mais arte não tem preço	Negative	Negative
Solta os pitbulls nesses petralhas prefeito	Neutral	Neutral
Show prefeito aula de cidadania e marketing parabéns	Positive	Positive
O dória é foda sp acertou nas eleições finalmente	Positive	Positive
Prefeito só não humilha o povo sofrido quer mostrar trabalho então mostre veja se algum médico está indo mesmo atender o povo	Positive	Negative
O senhor inventou o almoço grátis ou tá enganando a população mesmo	Neutral	Negative

Results

The tweets were collected every 2 months considering only the months when the number of tweets exceeded a total of 600, thus representing greater interaction between the followers and posts by the user @jdoriajr. From August 2016 to July 2017, the total of tweets available used in the grouping represented almost 100% of the total collected (75,372 out of 76,690) (Table 3).

The tweets whose positivity degree calculation resulted in zero—either for the inexistence of a positive or negative word with correspondents in the word list, or for the calculation result—were discarded since zero represents a possible neutral sentiment. Thus, the number of tweets submitted to the analyses had a 43% reduction, from 75,372 to 42,466.

For each one of the six bimesters, from August/September 2016 to June/July 2017, all corresponding tweets were divided into hour groups based on their respective creation dates, with up to 100 tweets per hour in order to obtain a better sampling equivalence in a 24-h period. Subsequently, the arithmetic mean of the positivity degree was calculated for every 24 h, beginning at 0 and ending at 23.

In the mean scale of the positivity degree calculated for each hour, with minimum of -0.1306 and maximum of 0.1489, except in the bimesters of the months of Aug.-16/

Table 3 Total tweets collected per month and year. Source: Prepared by the author

Year and month	Total tweets
Aug./16	624
Sept./16	2849
Oct./16	2371
Nov./16	624
Dec./16	692
Jan./17	8216
Feb./17	4752
Mar./17	11,716
Apr./17	13,530
May./17	13,125
Jun./17	6706
Jul./17	10,167
Overall total	75,372

Sept.-16 and Oct.-16/Nov.-16, which could be noted in the negative sentiment peaks between 6 and 7 in the morning, reaching -0.0759 and -0.1306 (the minimum possible) respectively, all other bimesters showed that the tweet sentiment fluctuated in a relatively constant cyclical pattern. It is also possible to note that except in the first (Aug.-16/Sept.-16) and last (Jun.-17/Jul.-17) bimesters studied, all others maintained the fluctuations previously discussed at an average of positive sentiment with minimum of 0.0189 for Apr.-16/May.-16, and maximum of 0.0728 for Dec.-16/Jan.-17 (Fig. 2).

In the analyses aimed at assessing the polarity of the average of positivity degree calculated, we noted maximum peaks of negative sentiment versus positive sentiment at 7 a.m. in Aug.-16/Sep.-16 and Dec.-16/Jan.-17, respectively. In order to reinforce the sentiment polarity verification noted during the described bimesters and times, the cluster analysis confirmed this conclusion by forming two clusters, which concentrate the negative and positive sentiments between 6 and 8 a.m. for each of the bimesters considered (Figs. 3, 4).

Discussion

A study published in Science in 2011 on the isolated variation of positive and negative humor according to the time of the day, sleep cycle and pace of work showed that we basically reach a good humor peak between 8 and 10 a.m. after we awake—which is noted between 12 and 9 p.m., and other two peaks that can occur from noon or afternoon to early evening. As for negative humor, peaks were noted in mid-morning, afternoon and evening, whereas other studies showed that negative sentiments are subject to significant variations throughout the day [13].

In a second analysis, also discussed in the same study published in Science, whose samples could be better individualized using Twitter posts, thus allowing user follow-up in real time—exactly as carried out in this study—they found that the pace of positive and negative sentiments was almost identical in weekdays with two notable peaks, one in early morning and the other around noon. When these sentiments were analyzed on

Fig. 2 Sentiment variation by time/bimester (Source: prepared by the author)

Saturdays and Sundays, higher levels of positive sentiments were registered, which may be due to lack of stress felt at work on weekdays [13].

Recent studies have found that the sentiment expressed by people in social media and related to specific events, for example, an opinion about a political candidate, can vary from day to day as a result of the direct influence of thousands of communication vehicles, constantly affecting the way these people express their opinions and sentiments [14].

Since the object of this study was the evaluation of the sentiments in the selected tweets, it is important to carry out a judgmental analysis of the noticeable polarity of

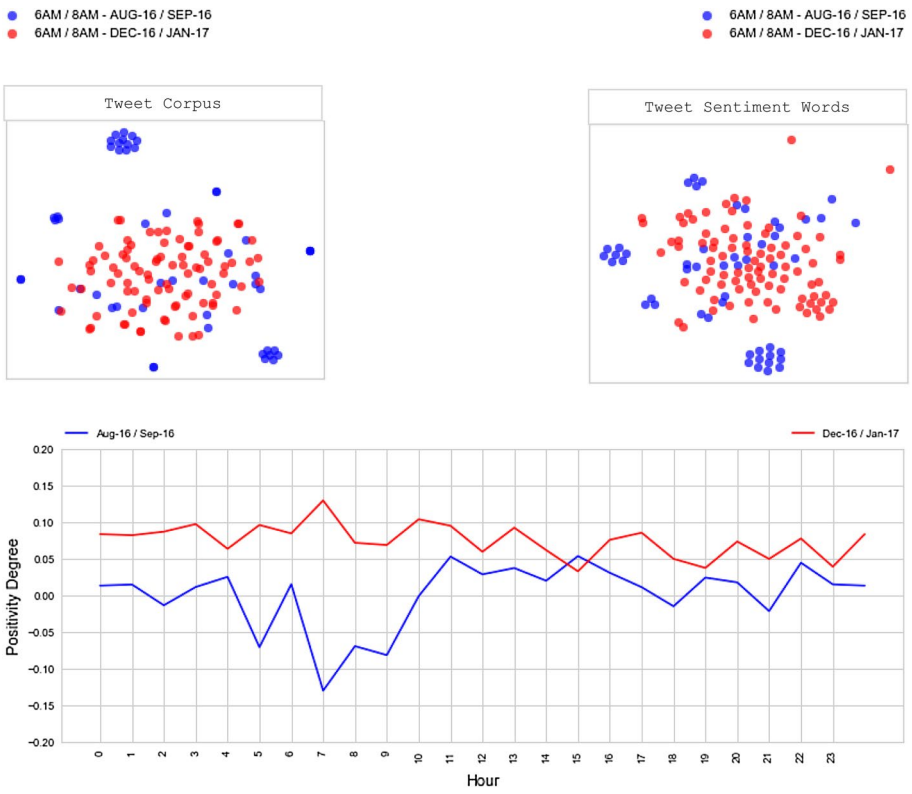

Fig. 3 t-DSNE Cluster analysis of posts between 6 and 8 a.m. in the bimesters of Aug.-16/Sept.-16 and Dec.-16/Jan.-17 (Source: prepared by the author)

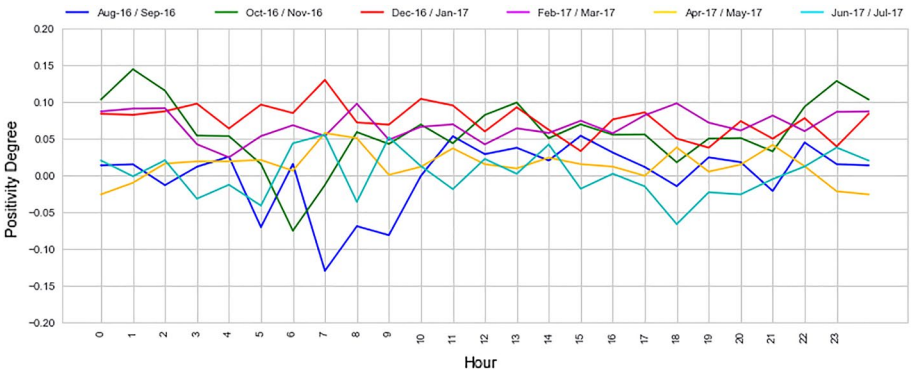

Fig. 4 Sentiment variation by time/bimester (all bimesters) (Source: prepared by the author)

positive and negative sentiment, particularly observed at 7 a.m. in the months of Aug-16/Sept-16 and Dec-16/Jan-16.

Considering that João Doria Jr. was elected mayor of the city of São Paulo on October 2, 2016, but only took office 2 months later, on January 1, 2017, this fact may explain the predominance of positive sentiments in the second bimester (Dec-16/Jan-17) analyzed, if we consider the high expectation by his constituents with the recent inauguration of his government. As for the neutrality mean observed in the positivity degree for

the 24 h of every day of the first bimester (Aug-16/Sept-16), we could relate it to the wish of the opposition to reduce the popularity of the candidate on the eve of election. This assumption is not only due to the positivity degree mean observed, but mainly because of the understanding of the clusters formed when we analyzed the period from 6 to 8 a.m. in these two bimesters, since even analyzing another negative sentiment peak in the bimesters after the election, we do not see the formation of such evident cluster, which leads us to conclude that the use of negative words in the bimester previous to the elections clearly differ from those used after this event. Thus, we may identify those that have always been unfavorable to the candidate before the elections from those that only expressed dissatisfactions after the inauguration of his office.

The consecutive decrease of the positivity degree mean observed from the bimester of the election (Oct-16/Nov-16) until the last one analyzed (Jun-17/Jul-17), could be explained based on the history of the gradual loss of expectation by Brazilian constituents towards their candidates after election with respect to a feeling or confirmation that nothing, or very little, has been accomplished by them. Through the word clouds (Fig. 5), it is possible to clearly note the significant increase of negative words in the first months and after 6 months of the mayor's office.

Conclusion

The main purpose of this study was to demonstrate how the sentiment by the followers of Twitter user @jdoriajr oscillated during a 12-month period which preceded and succeeded the date of election for mayor of the city of São Paulo. It was possible to evaluate the reaction of users in the social media based on important political and economic events in Brazil.

We also showed the efficiency and practice of implementing the traditional sentiment analysis method named Lexicon, which basically generates a score capable of measuring the positivity degree of a sentence based on the words used and their respective positive and negative correspondences. We also think it is important to clarify that, in spite of the practicality of the implementation of this method, its use is not recommended for small samples of sentences, mainly due to their inability to categorize negative sentences, which invert the meaning of a positive word to negative and vice versa.

Cluster exploration was essential to classify a clear difference in profiles, which concerns the words used in the tweets analyzed, suggesting the existence of user groups totally contrary and favorable to the candidate and/or his posts in the period studied. In

Fig. 5 Word clouds in the Oct-16/Nov-16 and Jun-17/Jul-17 bimesters (Source: prepared by the author)

addition to the clusters, word clouds showed the contrast of positive and negative words in the first months and after 6 months of the mayor's office.

In order to achieve greater precision in the analysis of sentiments in the tweets, we employed evaluation methods contemplating machine learning with the use of reliable categorization database of the sentiments expressed in tweets with political content.

Abbreviations
Jan.: January; Feb.: February; Mar.: March; Apr.: April; May: May; Jun.: June; Jul.: July; Aug.: August; Sep. or Sept.: September; Oct.: October; Nov.: November; Dec.: December; API: application programming interface; NLP: natural language processing.

Acknowledgements
Not applicable.

Authors' contributions
The author read approved the final manuscript.

Authors' information
Rubens Mussi Cury is professional with problem-solving skills and talent to develop solutions. Over 15 years' experience in the Data Analysis & Software Development.

MBA in Big Data & Business Analytics. Constantly improving my knowledge in NLP, Machine Learning, Predictive Analytics and Data Engineering.

Dual Citizenship • BRA & ITA.

Funding
Not applicable

Availability of data and materials
All data used in this study are publicly available and accessible in the source Twitter.com.

Competing interests
Data mining, Big data, Machine learning and Natural language processing, Sentiment analysis, Big data in politics.

Author details
[1] Fundação Getúlio Vargas, Edifício John F. Kennedy, Av. Nove de Julho, 2029-2º andar, Bela Vista, São Paulo, SP 01313-902, Brazil. [2] São Paulo, Brazil.

Received: 29 March 2019 Accepted: 24 May 2019
Published online: 01 June 2019

References

1. Eslami M, et al. I always assumed that I wasn't really that close to [her]: reasoning about invisible algorithms in the news feed. 2015. http://www-personal.umich.edu/~csandvig/research/Eslami_Algorithms_CHI15.pdf. Accessed 3 July 2018.
2. Hern A. How social media filter bubbles and algorithms influence the election. 2017. https://www.theguardian.com/technology/2017/may/22/social-media-election-facebook-filter-bubbles. Accessed 4 July 2018.
3. Halavais A. Search Engine Society. 1st Edition. ed. [S.l.]: Polity Press, 2009. 196 p. https://www.amazon.com/Search-Engine-Society-Alexander-Halavais/dp/0745642152. Accessed 4 July 2018.
4. Loader BD, Mercea D. Networking democracy? Social media innovations in participatory politics. 2011. http://openaccess.city.ac.uk/5528/. Accessed 1 July 2018.
5. Sen AK. Development as freedom. 1st ed. United States: Anchor Books; 2000.
6. Aggarwal CC, Zhai C. Mining text data. New York: Springer; 2012.
7. Hearst MA. Untangling text data mining. 1999. http://citation.cfm. Accessed 12 July 2018.
8. Herschel RT. Organizational applications of business intelligence management: intelligence management: emerging trends. Un: IGI Global, 2012. https://www.amazon.com/Organizational-Applications-Business-Intelligence-Management/dp/1466602791. Accessed 12 July 2018.
9. Gupta V, Lehal GS. A survey of text mining techniques and applications. 2009. http://learnpunjabi.org/pdf/gslehal-pap18.pdf. Accessed 12 July 2018.
10. Ghosh S, Roy S, Bandyopadhyay SK. A tutorial review on text mining algorithms. 2012. https://pdfs.semanticscholar.org/5fc6/b674cde1f39847b8783349af200eb68c9d48.pdf. Accessed 12 July 2018.
11. Pace-Sigge M. Spreading activation, lexical priming and the semantic web: early psycholinguistic theories, corpus linguistics and AI applications. 1st ed. Finland: Palgrave Pivot; 2018. p. 135.
12. Bengfort B, Bilbro R, Ojeda T. Applied text analysis with python (Locais do Kindle 2). O'Reilly Media. Kindle edition: enabling language-aware data products with machine learning. 1. ed. United States: O'Reilly Media, 2018. 332 p.

https://www.amazon.com.br/Applied-Text-Analysis-Python-Language-Aware-ebook/dp/B07DNKHJL8. Accessed 12 July 2018.
13. Golder SA, Macy MW. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Department of Sociology, Cornell University, Ithaca, NY 14853, USA.: American Association For The Advancement Of Science, 2011. 5 p. v. 333. http://science.sciencemag.org/content/333/6051/1878/. Accessed 30 Aug 2018.
14. O'CONNOR, Brendan et al. From tweets to polls: linking text sentiment to public opinion time series. 2010. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842. Accessed 6 Sept 2018.

Publisher's Note