Twitter Users: What Do They Know? Do They Know Things? Let's Find Out!

Analyzing the accuracy of predictions of public approval of the government using sentiment analysis of Twitter data

Marco Antônio Ribeiro de Toledo
RA: 11796419
B.Sc. in Computer Science
Instituto de Ciências Matemáticas e de Computação
University of São Paulo (USP)
mardt@usp.br
ORCiD: orcid.org/0000-0002-0484-8450

***Abstract*—Social media has become more and more the unofficial "town's square" of our society, allowing for the sharing of opinions and discussions in this marketplace of attention. With that in mind, this paper aims to show how mining these discussions may allow for statistical inferences of its population through processes much easier than previously held standards of in person surveying, specifically in the context of political opinions, although any statistical data collected must be taken with a grain of salt due to how sensitive the accuracy of these results are to initial assumptions and data modeling.**

***Index Terms*—data mining, social media, political analysis, prediction accuracy**

I. Introduction

The current paradigm for web content generation, the so called Web 2.0, was defined in 2005 by Tim O'Reilly as [1]

> *[...] the network as platform, spanning all connected devices; Web 2.0 applications [...] [get] better the more people use it, consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others, creating network effects through an "architecture of participation," and going beyond the page metaphor of Web 1.0 to deliver rich user experiences.*

This idea of federated content creation allows the ideas expressed in these online spaces to more closely resemble the actual opinions of its constituents, encouraging users to express their personal opinions and relate to ideas expressed by peers.

The decentralized nature of this content and its ample access also allows for easy mining of huge sets of data, so considering the online content as a truthful analog for real world ideas, mining of online data asymptotically matches the real world data for the population that the users are a sample of.

With that in mind, the goal of this paper is to analyze the accuracy of online mined sentiment data, specifically in the context of government approval, compared to traditional polling, testing the hypothesis that this easily implemented method is a good real world approximation for traditionally obtained data.

II. Related work

With the widespread usage of social media, like the microblogging platform *Twitter*, by the general population, its usage as a barometer for public opinion has been discussed more and more in recent years, even more so in politics, a field where accurate predictions of public opinion are essential. This resulted in ample research being done in the area, from initial work like Tumasjan et al. [16] in 2010 using Linguistic Inquiry and Word Count to analyze favorability rates of parties and politicians and, therefore, predict election results, to more recent and complex models like Nasrul et al. [4] in 2018 using Support Vector Machines to identify public satisfaction with government services, resulting in fairly accurate predictive models.

To support this kind of research, much has been analyzed on the ability to accurately extract sentiment data of *tweets* at random, with different methods of analysis constantly being tested for their accuracy, from simple semantic scoring of corpuses as in [10] to more complex lexicon-based sentence analysis like the work done by Meduru et al. [11], with recent research showing that the best performing classifier considering accuracy, precision, recall and F1-score is a combination of the more complex and computationally intensive Logistic Regression and Stochastic Gradient Descent as seen in [19].

In the field of brazilian politics, studies have showed the accuracy of sentiment analysis models when working with *tweets* from brazilian users on a portuguese corpus, as seen in [18], and others have been conducted in the past on the accuracy of twitter content as a predictor of public opinion as

[1] http://radar.oreilly.com/2005/10/web-20-compact-definition.html

seen in [12] and [13] giving accurate results, although running into limitations.

This approach has also been used by private companies like *Arquimedes* [3], with results comparable to traditional surveys conducted by companies like *Vox Populi*, *IBOPE* and *Datafolha*, resulting in those analyses now being regularly used by many reputable sources (as seen in [2]).

All code used for this paper, as well as this document itself, will be available in the accompanying *Github* repository.

A. Hypothesis

Taking into account previous research done in the field, our main hypothesis is that opinion data mined straight from a given social media platform (*Twitter* in our case) is within margin of error of traditional polling, allowing its results to be taken as statically equivalent to the much more costly and time consuming traditional, in person, polling techniques.

This hypothesis is supported by previous work in other contexts that show the reliability of such methods in politics, as those done in Indonesia in 2018 [6], in the context of the french elections in 2017 [17] and many others [16][9], showing results that closely correspond to traditionally collected data. Also, other studies have already theorized on the positive impact on the popular participation in politics that could be brought to the brazilian political landscape by such data mining [13] [12].

III. Methodology

The main objective of this study is to compare the accuracy of data mining methods to the traditional polling methods on the public's opinion of the current brazilian administration, so we try to abstract the qualitative data mining classification into a quantitative identification of the public's opinion in a 3-way classification akin to the one done by IPEC in its opinion polling on the approval of the current government (as seen on its monthly press report [14]).

Here we outline the methodological procedure for each step of this study: the data collection, extracting the relevant information from *Twitter* for analysis; data preprocessing, discarding ambiguous, duplicate or otherwise irrelevant texts; sentiment analysis, extracting the sentiment data of each extracted *tweet*; statistical analysis, comparing the results with ones obtained by traditional polling.

A. Data collection

Twitter itself provides its own API with support for keyword based searches over any time spam with the *Full-Archive search* [2], so the information was collected running searches over relevant keywords for our context: *'Bolsonaro', 'governo', 'presidente', 'país'* and *'mito'* within the analyzed period for the analysis. The time spent in this step is directly related to the usage of the Twitter API and its search quota for the given access level, being the bulk of the processing time.

B. Data preprocessing

For a more representative data set, focusing on reducing inconsistencies, redundancies and misleading information in the data, we had to, before starting the sentiment analysis, clear the data set removing:

- *Tweets* containing URLs, which may indicate an ambiguous text (differentiating if the sentiment expressed refers to the contents of the URL or the subject would require further investigating), detected with regular expressions
- Repeated, non retweeted text, which may indicate content by spam bots, detected by keeping a set of unique tweets
- User handles, anonymizing the data for publishing, replaced using regular expressions

C. Sentiment analysis

Once we had the striped text for each *tweet* and its related keyword from whose search the data was extracted, we could analyze the sentiment of each instance and tally them to the overall sentiment of each keyword. Due to the limitations in the scope of this paper, the sentiment mining was done using LeIA [1] a brazilian-portuguese *fork* of the lexicon-based sentiment analysis tool VADER [7]. The original tool has great accuracy for this kind of analysis considering its 0.96 F1 score on 3-class accuracy for a corpus of *tweets* while also maintaining good performance due to its lexical nature.

D. Statistical analysis

We propose then taking the harmonic mean of these values as an abstraction of the overall opinion on the current brazilian administration, which can be compared to IPEC's public opinion poll on administrative/political subjects (*PESQUISA DE OPINIÃO PÚBLICA SOBRE ASSUNTOS POLÍTICOS/ADMINISTRATIVOS*) [15] in the respective time period, ideally being within its margin of error of 2 percentage points.

[2] https://developer.twitter.com/en/docs/twitter-api/premium/search-api/quick-start/premium-full-archive

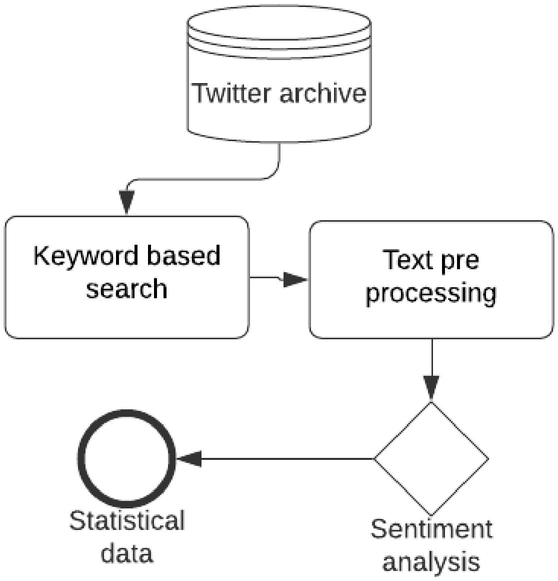

Fig. 1. Data analysis flowchart.

IV. Results

The latest compiled public opinion poll on the brazilian government done by IPEC dated June 2021 (from 06-17-2021 to 06-21-2021) [8] was used as reference for the analysis, with the *Twitter* data being collected from the same span of time. The underlying reference question for the classification was *Do you personally approve or disapprove of the way* President Jair Bolsonaro *is governing Brazil?*, with the following data (0.95 confidence with a margin of error of 0.02):

	Approves	Disapproves	Other*	Total
Count	601	1321	80	2002
Percentage	30%	66%	4%	100%

** either 'doesn't know' or 'didn't answer'.*

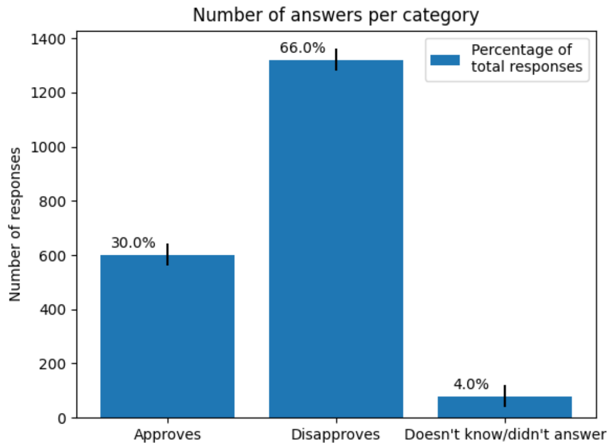

Fig. 2. Polling results.

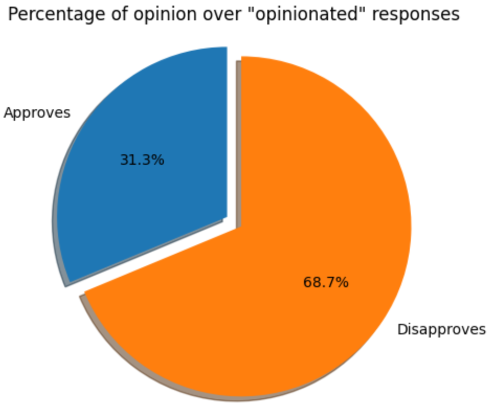

Fig. 3. Overral classification of binary responses.

Due to limitations of the free tier *Twitter* API, only 1750 *tweets* were collected using the archive search over the corresponding time span, which resulted in a data set of $n = 1524$ usable *tweets* after the post processing that were later classified using a custom fork of LeIA [3], adapting some features for this specific analysis (this process is also explained in the documentation for the accompanying code).

The classification of the overall sentiment of a *tweet* was done using the compound score for the analyzed text. As per the original documentation for the tool, *tweets* with a compound score of over 0.05 were classified as positive, with $m_{positive}$ total tweets, those with a scores lower than -0.05 were classified as negative, with $m_{negative}$ total tweets, and the rest were included as "neutral", totaling $m_{neutral}$ tweets.

[3] https://github.com/Ocramoi/LeIA

The percentage of *tweets* in each category was taken as an estimator for the overall opinion of the population as such:

$$\hat{p}_{approves} = \frac{m_{positive}}{n}$$

$$\hat{p}_{disapproves} = \frac{m_{negative}}{n}$$

$$\hat{p}_{other} = \frac{m_{neutral}}{n}$$

Also, due to the lack of precision in keyword based mining, resulting in an elevated number of neutral tweets, the ratio of approval within those with a binary answer (approves/disapproves) was also estimated, as such:

$$\hat{p}_{ratio} = \frac{m_{positive}}{n - m_{neutral}}$$

The raw data collected was as follows:

	Positive	Negative	Neutral	Total
Count	388	836	300	1524
Percentage	25%	55%	20%	100%

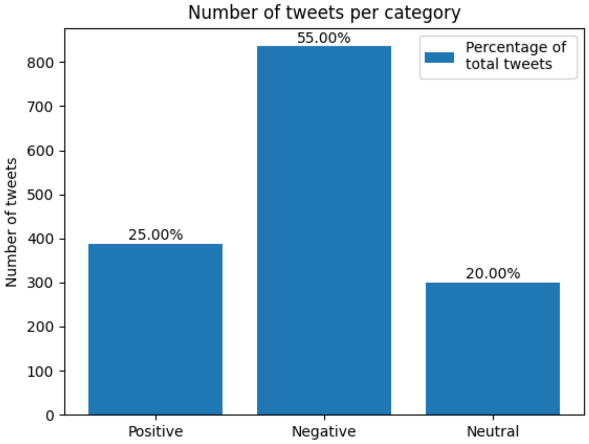

Fig. 4. Mining results.

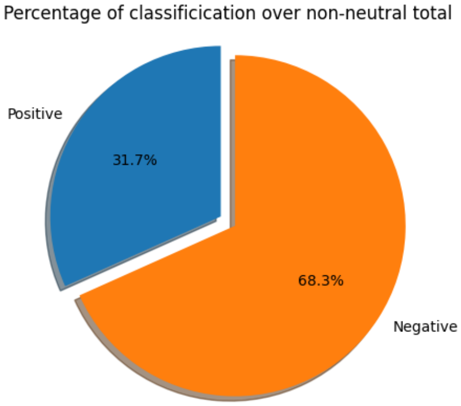

Fig. 5. Overral classification of non-neutral tweets.

From which we can calculate our biases as follows

$$V(\hat{p}_{approves}) = \frac{388}{1524} - 0.30 \approx -0.045$$

$$V(\hat{p}_{disapproves}) = \frac{836}{1524} - 0.66 \approx -0.111$$

$$V(\hat{p}_{other}) = \frac{300}{1524} - 0.04 \approx 0.157$$

$$V(\hat{p}_{ratio}) = \frac{388}{1524 - 300} - \frac{601}{2002 - 80} \approx 0.004$$

This shows a strong bias in favor of neutral responses, with the total number of *tweets* in each category being a weak indicator for the overall population taking the polling as a standard, with values well over the margin of error. However, the estimated ratio of positive responses over the total number of “opinionated” ones does describe the sample fairly well, reinforcing the idea that the naive data mining over the *tweets* with only keywords tends to over represent neutral responses.

V. Discussion and Future Work

The generated estimates show that when well modeled, estimates over social media mined data may give us true-to-life results, allowing for an easier, cheaper and faster way of assessing sentiment data over a online-represented population, like the Brazilian demographic in our example. However, it also shows how sensitive the accuracy of this predictions is to the definition of initial assumptions and statistical models.

In the field of political analysis, this kind of data is of utmost importance for everyday analysis involving public assessments, for example, political campaigns, policy satisfaction surveying, voting intention, etc. Considering the price of traditional in person polling, running anywhere from R$18.95 to R$115.50 per person [4], the near gratuity of a data mining method like the one presented in this paper (safe some

[4] https://www.moneytimes.com.br/quanto-custam-as-pesquisas-eleitorais-veja-as-mais-caras/

men/hours of initial setup and statistical work) expands greatly the accessibility of such data, showing the importance and utility of more research in the area.

Future work in this question could tackle some of the short comings of the methods used, from the small sample size, inflexibility of the lexicon matching (not accepting acronym expansion, misspellings, and other important features, for example) or lackluster statistical analysis. Other methods of sentiment analysis should surely be tested, from expanded lexicon based ones to models using supervised machine learning like a Support Vector Machine, or, as suggested by Yousaf, et. al [19], Stochastic Gradient Descent/Linear Regression, over pre classified *Twitter* data, like the reliable *TweetSentBR* [5]. The manual classification of *tweets* for the classifiers could yield even better results for more recent analyses, making sure to base estimations on up to date models, although this would grow considerably the scope and costs of such analysis.

References

[1] Rafael J. A. Almeida. *LeIA - Léxico para Inferência Adaptada*. 2018. URL: https://github.com/rafjaa/LeIA.

[2] Arquimedes. *Mídia - Arquimedes*. Oct. 2021. URL: https://arquimedes.social/news/.

[3] Arquimedes. *Sobre - Arquimedes*. Oct. 2021. URL: https://arquimedes.social/sobre/.

[4] Moh. Nasru Aziz et al. *Sentiment Analysis and Topic Modelling for Identification of Government Service Satisfaction*. Department of Informatics, Institut Teknologi Spuluh Nopember, 2018. DOI: 10.1109/ICITACEE.2018.8576974. URL: https://api.twitter.com..

[5] Henrico Bertini Brum and Maria Das Graças Volpe Nunes. *Building a Sentiment Corpus of Tweets in Brazilian Portuguese*. Institute of Mathematical and Computer Sciences, University of São Paulo, 2017. URL: https://bitbucket.org/HBrum/tweetsentbr.

[6] Widodo Budiharto and Meiliana Meiliana. "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis". In: *Journal of Big Data* 5 (1 Dec. 2018). ISSN: 21961115. DOI: 10.1186/s40537-018-0164-1.

[7] C J Hutto and Eric Gilbert. *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. 2014. URL: http://sentic.net/.

[8] IPEC. *PESQUISA DE OPINIÃO PÚBLICA SOBRE AVALIAÇÃO DO GOVERNO FEDERAL*. URL: https://www.ipec-inteligencia.com.br/Repository/Files/26/04_13_Ipec_JOB_21_0046-7_Avaliacao_do_Governo_Relatorio_de_tabelas.pdf.

[9] Parnian Kassraie, Alireza Modirshanechi, and Hamid K. Aghajan. "Election vote share prediction using a sentiment-based fusion of Twitter data with Google trends and online polls". In: SciTePress, 2017, pp. 363–370. ISBN: 9789897582554. DOI: 10.5220/0006484303630370.

[10] Akshi Kumar, Prakhar Dogra, and Vikrant Dabas. "Emotion Analysis of Twitter using Opinion Mining". In: *International Conference on Contemporary Computing* (8 2015).

[11] Manogna Meduru et al. "Opinion Mining Using Twitter Feeds for Political Analysis". In: *International Journal of Computer* (). ISSN: 2307-4523. URL: http://ijcjournal.org/.

[12] Daniel José Silva Oliveira and Paulo Henrique de Souza Bermejo. "Mídias sociais e administração pública: análise do sentimento social perante a atuação do Governo Federal brasileiro". In: *Organizações & Sociedade* 24 (82 Sept. 2017), pp. 491–508. ISSN: 1984-9230. DOI: 10.1590/1984-9240827.

[13] Daniel José Silva Oliveira et al. "The application of the sentiment analysis technique in social media as a tool for social management practices at the governmental level". In: *Revista de Administracao Publica* 53 (1 Jan. 2019), pp. 235–251. ISSN: 19823134. DOI: 10.1590/0034-7612174204.

[14] *PESQUISA DE OPINIÃO PÚBLICA SOBRE ASSUNTOS POLÍTICOS/ADMINISTRATIVOS*. IPEC, Sept. 2021.

[15] *Pesquisas - IPEC*. 2021. URL: https://www.ipec-inteligencia.com.br/pesquisas/.

[16] Andranik Tumasjan et al. *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment*. 2010. URL: www.aaai.org.

[17] Lei Wang and John Gan. "Prediction of the 2017 French Election Based on Twitter Data Analysis". In: 2017. ISBN: 9781538630075.

[18] Augusto Weiand, Fernanda Rodrigues, and Ribeiro Weiand. "Análise de sentimentos do Twitter com Naïve Bayes e NLTK Augusto We". In: *ScientiaTec: Revista de Educação, Ciência e Tecnologia do IFRS* 4 (2017).

[19] Anam Yousaf et al. "Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD)". In: *IEEE Access* 9 (2021), pp. 6286–6295. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3047831.