Identification of circulating information on corruption in Brazil using data mining and machine learning

Douglas Farias Cordeiro [1] **Kátia Kelvis Cassiano** [1] **Núbia Rosa da Silva** [2 3]

Abstract

Corruption is one of the biggest problems in economic, cultural, and social progress and development. In Brazil there are several events reported in relation to corruption. A demand that exists in relation to this phenomenon is the understanding of the information circulating on social networks. This article presents an analytical study using data mining and machine learning techniques to identify informational patterns in a dataset of tweets related to corruption in Brazil. The results obtained show the existence of three main clusters of subjects, as well as the notable confirmation of dissatisfaction on the part of the citizen.

1. Introduction

Corruption is usually related to an illegal act and is commonly associated with power, politics, economic elites and public servants. The notion of legality and illegality is something that involves the idea of corruption. This notion is linked to the history and sets of social values. Therefore, it is possible to consider that different cultures have different conceptions about what is legal or illegal, and consequently differ in perceptions of corruption (Rose-Ackerman & Palifka, 2016).

During elections, corruption is one of the factors that can influence citizens vote. Corruption can be a major obstacle in the guarantee for stability and quality of democracy (De Vries & Solaz, 2017). The growth in news and the visibility of corruption are associated with a better performance of the control mechanisms and public policies implemented (Mancini, 2018).

At the same time, there is a need to raise the citizen's perception of public policies to control and combat corruption in Brazil. The interaction between the State and society is a challenge. Although there are initiatives (Open Data projects, for example) and access channels, communication is not always effective. However, through Web 2.0, there is a potential for accessibility, so that even though there are barriers in terms of establishing direct interactions, in the context of social networks there is a democratic socialization environment, greater participation and freedom of expression in terms of opinions, feelings, debates, or even dissemination of information (Brandtzæg & Heim, 2009).

This paper presents the application of intelligent solutions, based on the use of Doc2Vec and Naive-Bayes methods to understand the circulating information on the social network Twitter related to co-corruption in Brazil, in order to generate strategic inputs that support the implementation of assertive and oriented public policies to the citizen's needs.

2. Metodology

The methodology of the work is based on the KDD process (Knowledge Discovery in Databases) (Fayyad et al., 1996). KDD guides the generation of information by recognizing patterns in a dataset, based on the execution of five steps: selection, pre-processing, transformation, data mining, and interpretation. Our analysis was implemented in Python programming language, with Natural Language Processing (PLN) models gensim.models (Gensim Python Library) and NLTK (Natural Language Toolkit), and data analysis modules (Pandas, Matplotlib , Seaborn).

The analysis was carried out from a dataset of posts extracted from the social network Twitter using web scraping, totaling 102,171 tweets published from July 2018 to June 2019 with the hashtag *#corrupcao* (corruption in Portuguese). In the pre-processing step, stopwords, special characters (@ and #) and emoticons were removed. Doc2Vec (Le & Mikolov, 2014) and Naive Bayes (Goel et al., 2016) were used to extract patterns and classification. NLP techniques were used for semantic analysis of the content in order to better represent the extracted data.

Doc2Vec consists of an unsupervised learning model that

[*]Equal contribution [1]Faculty of Information and Communication (FIC), Federal University of Goiás (UFG), Goiânia, Brazil [2]Institute of Biotechnology (IBiotec), Federal University of Goiás (UFG), Catalão, Brazil [3]Department of Engineering, Federal University of Goiás (UFG), Catalão, Brazil. Correspondence to: Douglas Farias Cordeiro <cordeiro@ufg.br>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020. Copyright 2020 by the author(s).

uses distributed vector representations of the terms or words in a textual document. The model was trained to predict words or terms and thus obtain a distribution based on probabilities of occurrence, and not just frequency of occurrence, in order to consider that words in the same meaning are arranged in the same vector space.

Based on the trained and validated model, an analysis of the semantic similarity of the documents was performed. A weighted graph was generated based on the similarity matrix to illustrate the relationship between the tweets. For each node a weight is associated with the sum of the similarity values with other documents. Through this graph it is possible to explore characteristics and extract patterns from the circulating content, and important information for later stages of knowledge generation in large databases.

The analysis of feelings was performed based on the Naive-Bayes method, as presented in (Goel et al., 2016). The results obtained were labeled in terms of polarity: negative, positive or neutral.

3. Results

Through the application of the Doc2Vec method on the treated dataset, a graph was generated from the similarity matrix between the tweets. Figure 1 shows the result obtained. It is possible to observe the presence of three main clusters, which indicate tweets that have a high similarity with each other, that is, similar content patterns.

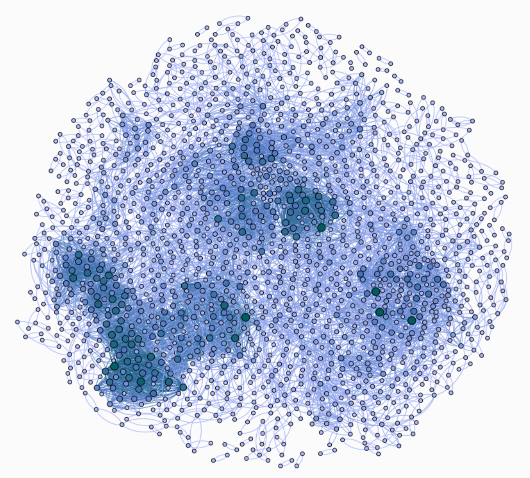

Figure 1. Similarity graph.

The nodes related to the clusters found were analyzed for the most frequent terms. In this sense, the indicator groups the following patterns: 1) posts related to the Brazilian Federal Police's Lava Jato operation; 2) posts related to popular manifestations of dissatisfaction with corruption in Brazil; 3) posts related to actions of the executive and judiciary on corruption. Table 1 presents the most frequent terms for each of the identified clusters.

Cluster	Terms
1	*lava jato*; bribe; company; contractor; snitch; money; scheme; petrobras; phase; operation.
2	attorney; society; democracy; lie; attack; population; change; abuse; persecution; impunity.
3	minister; reporter; inquiry; declaration; senate; presidency; impeachment; obstruction; vote; justice.

Table 1. Ten most frequent terms per class (The terms were translated from Portuguese.).

Using the Naive-Bayes method for sentiment analysis, the sentiment associated with the collected tweets were calculated. Figure 2 shows the distribution of the number of tweets by sentiment. The predominance of tweets classified as negative is evident, and confirms social dissatisfaction with the corruption issue. However, the month of October 2018 has a special emphasis on this value, being associated with the holding of the Brazilian presidential elections.

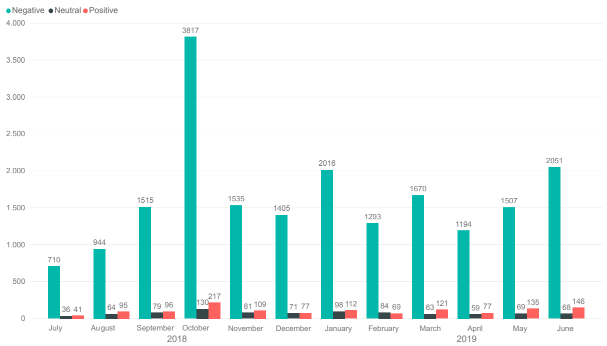

Figure 2. Sentiment analysis.

4. Conclusions

This paper presented the application of Natural Language Processing techniques to knowledge discovery about the perception of citizens in relation to corruption in Brazil, from posts on the social network Twitter. The results showed that the circulating information on the corruption theme in that social network is predominantly negative and can be characterized from three dominant aspects: Lavajato operation, popular dissatisfaction and actions of the executive and judicial powers. Such results generate inputs for the evaluation of State actions to the guarantee of social participation and the exercise of citizenship.

References

Brandtzæg, P. B. and Heim, J. Why people use social networking sites. In Ozok, A. A. and Zaphiris, P. (eds.), *Online Communities and Social Computing*, pp. 143–152, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

De Vries, C. E. and Solaz, H. The electoral consequences of corruption. *Annual Review of Political Science*, 20(1): 391–408, 2017.

Fayyad, U., Piatetsky-Shapiro, G., and P.Smith. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

Goel, A., Gautam, J., and Kumar, S. Real time sentiment analysis of tweets using naive bayes. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pp. 257–261, 2016.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 1–9, Beijing, China, 2014. JMLR.org.

Mancini, P. “assassination campaigns”: Corruption scandals and news media instrumentalization. *International Journal of Communication*, 12(1):3067–3086, 2018.

Rose-Ackerman, S. and Palifka, B. J. *Corruption and Government - Causes, Consquences and Reform.* Cambridge University Press, Cambridge, UK, second edition, 2016.