Opinion Analysis Applied to Politics: A case study based on *Twitter*

Gilberto Nunes
Federal Institute of Education, Science and Technology of Piauí - IFPI / Picos, Piauí, Brazil
gilberto.nunes@ifpi.edu.br

Denivaldo Lopes and Zair Abdelouahab
Federal University of Maranhão - UFMA / São Luís, Maranhão, Brazil
denivaldo.lopes@ufma.br, zair@dee.ufma.br

Abstract

Nowadays, social networks such as *Facebook* and *Twitter* are openly available for everyone around the world over the Internet. These websites provide some functionality without costs, such as: creation/edition of communities and social networks; it provides support to a large variety of multimedia contents (e.g. audio and video) and support to interactive communications (e.g. chats and post). *Twitter*'s users post comments about a range of subjects, such as, products, famous persons and politics. The dissemination of the information in these social networks should be considered due to their global coverage. An important functionality of *Twitter* is the support to georeferenced posts making the localization of posts possible. In this paper, we propose an approach to make the Sentiment Analysis or Opinion Mining. Our approach is based on Mining Web and of Opinion, Geographic Information System (GIS) and Machine Learning in order to recover relevant information from *tweets*. The information recovered follows our approach is essential to provide support to the verification of population trends, e.g. in politics domain. We propose a prototype that makes the analysis of population trends, in special, Brazil's politic context and the *impeachment* process in course.

Keywords **—** Web Mining; Opinion Mining; Machine Learning; Opinion Analysis; Twitter; Geographic Information System.

1 Introduction

Prasetyo and Hauff (2015), Jungherr (2013) and Lampos (2012) propose approaches to determine the voter intention polls based on information recovered from *Twitter*.

In this paper, we propose another approach based on opinion analysis applied to politics in order to colect information from *Twitter* and determine the public opinion about the current *impeachment* process in Brazil that is submitted the elected president in October 2014.

During the process of *impeachment*, as well as the electoral process, opinion surveys are applied such as presented by Rothschild[1]. He says that this opinion survey is generally based on data obtained from printed forms filled by the population. Our approach is based on opinion analysis to analyze messages obtained from *Twitter* to determine the Brazilian population's opinion about the *impeachment* of the Brazil's president. According to Currie (1998), *impeachment* is considered a process that can result in the removal of a person from public office after this person has violated the Constitution of her country.

In this paper, we show an approach based on knowledge discovery of textual sources, data from social networks (e.g. *Twitter*), Mining Web and of Opinion, Geographic Information System and Machine Learning. Applying our proposed approach, opinion trends about *impeachment* can be identified in the Brazilian population.

This paper is presented as follows. Section 3 presents some fundamental concepts to this research work. Section 4 presents our approach for performing opinion analysis from data obtained in *Twitter*, with Web Mining support, about *impeach-*

[1] **Forecasting Elections: Voter Intentions versus Expectations — Brookings Institution** - Link for ebook: http://www.brookings.edu/research/papers/2012/11/01-voter-expectations-wolfers.

ment process development in Brazil. Section 5 presents some results about the *impeachment* process development in Brazil. Section 6 shows a case study according to the *impeachment* process. Section 7 presents some conclusions and future directions.

2 Related Works

Most the related works use sentiment analysis and opinion mining for evaluate voting intentions, taking into consideration only the content the posts. For instance, we use for illustration purpose the following approaches: Prasetyo and Hauff (Dwi Prasetyo and Hauff, 2015), Jungherr (Jungherr, 2013) and Lampos (Lampos, 2012), as will be described below.

Prasetyo and Hauff (Dwi Prasetyo and Hauff, 2015), propose the use Twitter-based election forecasting, sentiment analysis and machine learning techniques to determine voting intention. For Indonesia's presidential elections 2014.

Jungherr (Jungherr, 2013), shows a work using four metrics to determine voting intention, likewise: the total number hashtags mentioning a given political party; the dynamics between mentions positive or negative a given political party; the total number hashtags mentioning one the candidates; and the total number users who used hashtags mentioning a given party or candidate. For Germany presidential elections 2009.

Lampos (Lampos, 2012), shows a study techniques and patterns for extracting positive or negative sentiment from tweets, which build on each other, through a supervised approach for turning sentiment into voting intention percentages. For United Kingdom presidential elections 2010.

Differently from approaches mentioned above, our work uses georeferenced data, addition to the textual content. Thus we can easily perform a spatial analysis, as shown the proposed case study (vide section 6). In the next section, we described the technological used in our case study.

3 Overview

In this section, we present the subjects Web Mining, Opinion Mining, Geographic Information System (GIS), Machine Learning and *Twitter*.

3.1 Web Mining

Web Mining is a process extracting data or information from web sources, as described by Zhang (2011). Second author (Zhang, 2011), Web Mining aims to find useful knowledge from Web and on the basis data mining, text mining, and multimedia to combine the traditional data mining techniques with Web. This mining type can be subdivided in: Web Content Mining, Web Usage Mining, Web Structure Mining. Mining Web Content refers to the extraction Web page content, the text contained on those pages is a good example contents to be extracted. Web Usage Mining is the automated recognition user utilization patterns based on the Web site. Web Structure Mining is based on interconnection between data or information in documents or sites by Web. Figure 1 illustrates the subdivision.

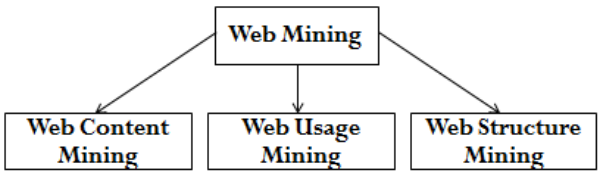

Figure 1: Basic Taxonomy for Web Mining. Font the image: (Zhang, 2011).

3.2 Opinion Mining

Opinion mining can be defined as a computational technique that takes care opinion in textual sources (Pang and Lee, 2008). It aims to extract information based on sentiment analysis (e.g. positive, negative and neutral) expressed by one or more writers and their texts (Pang and Lee, 2008). Opinion mining has a process that analyzes a large volume textual documents that contains a range subjects, such as, entertainment, politics, education and marketing. The social networks like *Twitter* have supported their users to express and share opinions and points view. Thus, social networks can be seen as a large documents volume in textual source and digital format.

3.3 Geographic Information System (GIS)

According to *Nuhcan* (2014), Geographic Information System (GIS) can be understood as a computational information system like any other, but the differential is the database that stores georeferenced data, i.e. the database includes latitude/longitude information linked to the data. Initially, GIS applications were restricted to desktop computers, but nowadays they are present the *Web*

(Servers to maps) and the *Smartphones* (map applications).

3.4 Machine Learning

Machine Learning is a subarea of Artificial Intelligence where the focus is to develop computational methods order to provide intelligent behavior to computers (Arel et al., 2010). Examples of Machine Learning are Support Vector Machine (SVM) (1998), Random Forest (2001) and Naive Bayes (2006).

3.5 Twitter

Twitter is a social networking service that enables the users to send and receive messages denominated *tweets* that have 140 character of maximum size for each *post*. *Twitter* has a large number of content such as profiles, general information, *tweets*, *emotions*, *hastags* and other (Tiara et al., 2015). This social network provides basically two API[2] to support the recovery of data: *Search* API and *Streaming* API. In our approach, we apply the *Twitter* API in order to recover the *tweets* and the georeferenced location where they were posted.

3.6 Metrics Evaluation

Once the *Twitter* data have been collected and processed, it needs a mechanism to determine the validity of the classification applied (Sokolova and Lapalme, 2009). Table 1 introduces the confusion matrix that is used to assist the calculation of the evaluation.

Table 1: Confusion Matrix.

		Predict positive	**Predict** negative
Real	positive	**TP** True Positive	**FN** False Negative
	negative	**FP** False Positive	**TN** True Negative

Font of data (Sokolova and Lapalme, 2009).

The metrics used this article are derived from the Confusion Matrix, which are: Accuracy, Sensitivity or Recall, Specificity, F1-Score and Precision (Sokolova and Lapalme, 2009).

[2]**Documentation Twitter Developers** - Link for documentation: https://dev.twitter.com/overview/documentation.

4 Proposed Approach for Opinion Mining Applied to Politics

Our proposed approach for opinion mining applied to politics is based on Knowledge Discovery in Databases (KDD) (Fayyad et al., 1996). To reach the proposed objectives this article, it was developed an approach that consists of five steps. This approach has been implemented by a *Software* Prototype[3] that assists its execution. The prototype composed to two modules, one for recovery (Works interconnected to *Search* API) and another for the analysis (Works interconnected to API WEKA) of data. In the first stage, occurs the acquisition of data (*tweets*). In the second stage, preprocessing data, to remove noisy structures. In the third stage, the feature extraction of the data using TF-IDF (Robertson, 2004). In the fourth step, we have the Text Mining, by means of the applied of the algorithms to Machine Learning presented previously. The fifth and last step, contemplates the evaluation the results obtained through the analysis of the Confusion Matrix (Sokolova and Lapalme, 2009). Finally, we found the positive and negative opinions. Figura 2 introduces the proposed approach which is based on KDD (Fayyad et al., 1996).

It is worth mentioning the importance of using the WEKA[4] tool and its API during execution of the steps present in this approach, with the exception of the data stage acquisition.

4.1 Acquisition of datas

The data (*tweets*) were recovered using the *Search Twitter* API. During the recovery process of *tweets* it is necessary Web Mining, specifically the Web Content Mining, in which recovered the texts contained in the posts by users the *Twitter*. A total of 1,218 georeferenced *tweets* were collected, based on posts related to *impeachment* of the President Dilma, during the March of 2016 which contained the following *hashtags*:

#FicaDilma, #SouMaisDilma, #NaoVaiTerGolpe, #FicaPT, #FicaLula, #NaoAoGolpe, #ForaDilma, #ForaPT, #ForaPTralhas, #ForaLula, #ForaDilmaLulaPT,

[3]**Software Prototype** - It is the result of applying a software process, as defined (Sommerville, 2006).

[4]**Machine Learning Group at the University of Waikato** - Version 3.7.12 and documentation, link: http://www.cs.waikato.ac.nz/ml/weka/documentation.html.

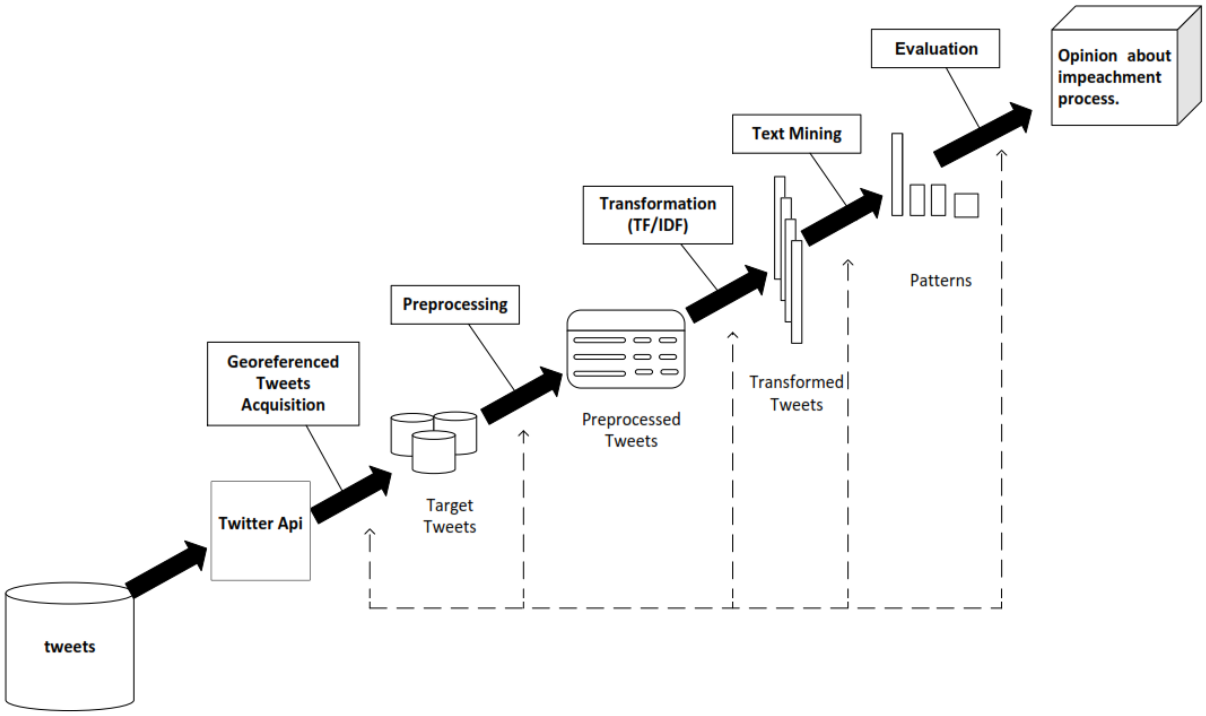

Figure 2: Proposed Approach for Opinion Mining (Based on KDD (Fayyad et al., 1996)).

#DilmaNao, #VaiTerImpeachment, #NaiVaiTerGolpeVaiTerImpeachment.

The georeferencing of *tweets* corresponds to the 26 Brazilian state capitals and the Brazilian Federal District. This step it is performed by recovery module the prototype made during the search.

4.2 Preprocessing

Before feature extraction of the *tweets*, it is important to remove the unwanted structures, such like: *hyperlinks* irrelevant words, special characters, and other references. After removing those, it is necessary the stemming and normalization applied on *tweets*. It is important to emphasize that the preprocessing occurs in copies of *tweets* collected (*corpus* (Khairnar and Kinikar, 2013)). Such device, it seeks to maintain the original *tweets* intact for avoid any inconsistencies. As the previous step, this step it is also performed by the recovery module contained in the prototype.

4.3 Feature Extraction

After the preprocessing stage, *tweets* were submitted to the feature extraction process, through the TF-IDF (Robertson, 2004) method. Once you have applied the method of TF-IDF (Robertson, 2004), the *tweets* are represented by the matrix of numeric values (*bag-of-words model*) as the mathematical definition of the TF-IDF (Robertson, 2004) model. The Text Mining has a model the representation using often as feature set, known as "*bag-of-words model*", with the help of the WEKA[4] tool and using your StringToWordVector method one created the model used this paper. In this model, documents are represented as a word vector. Thus, all documents are represented as a giant document/term matrix. In this paper, TF/IDF (Robertson, 2004) was used as the cell value to dampen the importance of those terms if it appears in many documents. This step it is performed by the analysis module assisted by the prototype.

4.4 Text Mining

Once generated the numeric matrix values, these values are used as inputs to the classification algorithms presented previously. These algorithms are seeking patterns of data interpretable within the matrix of values for determinate the classes of the *tweets* in positive or negative for Dilma's *impeachment*. This step it is also performed by the analysis module.

4.5 Evaluation

Lastly, we have the evaluation of the classification of data the confusion matrix and its metrics. Providing the obtaining of information, which will

provide the acquisition of knowledge at the end of the process of KDD (Fayyad et al., 1996).

5 Results

The Results Section of this research is divided into three subsections. Subsection 5.1 is responsible for describing the database that contains the samples used for training and testing. Subsection 5.2 includes the training models and test. Subsection 5.3 shows the results for the classification of *tweets*.

5.1 Data Base

This research, the database has 500 positive samples and 500 negative of *tweets* to posts related to the *impeachment* of the president Dilma. Totaling 1,000 samples in the database. Is worth emphasizing that the samples were divided only into positives and negatives, because the neutral samples have no representativity, as seen during the experiments. Samples were collected an automatic manner by Search API, but the labeling process was performed manually. During manual labeling it was aimed the selection of samples which had good representativity for the classification process, that is the most variable possible. Recalling that the *tweets* used this subsection are different from those used in subsection regarding the Case Study. These are geo-referenced to the capital and federal district that make up Brazil and a period of posts different from the month of March 2016. Thus, we seek to avoid potential problems in the *tweets* classification.

5.2 Training and Test Models

The generation of training models and test took place with the help of the WEKA[4] tool. Through this, we used the implementations of algorithms (SVM, *Naive Bayes* and *Random* Forest) classification, necessary for the creation of models. Scenarios were generated, respecting the training models and test as:

- **80% of the samples for training and 20% of test samples;**
- **60% of the samples for training and 40% of test samples;**
- **40% of the samples for training and 60% of test samples;**
- **20% of the samples for training and 80% of test samples.**

The algorithm that showed the best model was used in the case study this paper. The results for the proposed scenario and the best designs for each algorithm can be viewed in subsection (5.3) next.

5.3 Training and Cross-Validation results

Table 2: Results obtained with the application of metrics for each of the proportions using the classification algorithms.

Algorithms	Training/ Test	**Metrics Evaluation**		
		AC	SE	ES
SVM[1](Linear Kernel)	80%-20%	96.7%	98.1%	95.7%
	60%-40%	**96.9**%	**97.1**%	**96.7**%
	40%-60%	96.5%	98.0%	95.5%
	20%-80%	95.5%	97.4%	94.2%
Random Forest[2]	80%-20%	95.1%	94.5%	95.5%
	60%-40%	94.5%	93.0%	95.6%
	40%-60%	94.9%	91.6%	97.6%
	20%-80%	**95.8**%	**96.4**%	**97.7**%
Naive Bayes[3]	80%-20%	77.5%	76.2%	78.4%
	60%-40%	84.8%	84.4%	85.1%
	40%-60%	85.0%	83.8%	85.8%
	20%-80%	**89.3**%	**94.0**%	**86.7**%
Subtitle: AC: Accuracy; SE: Sensibility; ES: Specificity.				

[1] **Parameters WEKA** - type kernel: *linear;* SVM type: *C-SVC (classification);* gama: *0.5* and *other paratemtros with default values*;
[2] **Parameters WEKA** - Number of trees: *10;* and *other paratemtros with default values*;
[3] **Parameters WEKA** - *All paratemtros with default values.*

According to Table 5.3, can be checked that the greatest amount of accuracy was found for the proportion of 60% - 40%, using the SVM, with a hit rate 96.9%. While the lowest value was recorded by the accuracy *Naive Bayes* with a hit rate of 89.3% for the proportion of 20% - 80%.

According to the analysis results for Sensitivity in Table 5.3, we can conclude that the SVM has the highest rate in relation to the number of true positive feedback. With a Sensitivity rate of 98.1 % for the proportion of 80% - 20%.

Analyzing the data in Table 5.3 concerning Specificity, one can infer that the *Random Forest* presents the best result for true negative reviews, with a Specificity rate of 97.7% for the proportion of 80% - 20%.

Table 3: Results obtained with the application of metrics for each of the proportions using the classification algorithms for Cross-validation.

Algorithms	Quantities of folds	Metrics Evaluation		
		PR	RE	F1
SVM[1](Linear Kernel)	10	98.5%	97.8%	98.4%
Random Forest[2]	10	98.2%	97.5%	98.1%
Naive Bayes[3]	10	76.5%	82.9%	76.4%
Subtitle: PR: Precision; RE: Recall; F1: F1-Score.				

[1] **Parameters WEKA** - type kernel: *linear;* SVM type: *C-SVC (classification);* gama: *0.5* and *other paratemtros with default values*;
[2] **Parameters WEKA** - Number of trees: *10;* and *other paratemtros with default values*;
[3] **Parameters WEKA** - *All paratemtros with default values.*

According to the analysis results for Precision, Recall and F1-Score in Table 5.3, we can conclude that the SVM has the highest rate for the metrics used in cross-validation with 10 folds. With a Precision rate of 98.5 %, Recall rate of 97.8% and F1-Score rate of 98.4%.

6 Case study

In this case study were analyzed a total of 1,218 *tweets* georeferenced, highlighting that the *tweets* not georeferenced were discarded. These posts are referring to the period of March 2016, linked to the process of *impeachment* the president of the country. This period was selected based on two large manifestation schedules for the month. The first manifestation favorable[5] to *impeachment*, occurred on day 13 and the second contrary[6] on day 31.

Figure 3 presents the results to *tweets* collected and analyzed in the form of map for the regions of Brazil, using the approach proposed. Reminding that for plotting of the map used the *GeoServer* (*Web Map*), as shown in (Huang and Xu, 2011) and based on *shapefiles*[7] to the five regions of Brazil. These occurrences are posts containing *hashtags* cited previously in subsection 4.1.

Analyzing Figure 3, one can see that in the Midwest, Southeast and South map there most records in favor of *impeachment*. Assuming the map of the Northeast region, little more of most records are of opposed to *impeachment*. The map of the northern region is the only one of the five regions presenting the same results for the reviews.

It is important to note that other research related to the *impeachment* process have already been carried out since 2015 in Brazil, when the first evidences to the process. One of those researches are very similar to the one presented in the research in this study, being presented in the Veja[8] magazine. In it the magazine exposes results of a research on social networks by the company Torabit[9], in which 49.3 % of posts on social networks are favorable to *impeachment* and only 31.7 % contrary. Considering the results of the report and the proposed approach, it can be seen that the present work presents valid trends in relation to the *impeachment* process. It is remarkable that the proposed work informs trends by region, which does not happen with the work done by Torabit[9].

Seeking to standardize the presentation of data in the map plotted by the proposed approach (see Figure 3) we used a graphic seeking to make it understandable, as shown in Figure 4. Analyzing Figure 4, it can be seen, in simplified way, the percentages by region for each of the opinions, whether favorable or contrary to the *impeachment*.

It can be said that the proposed paper presents information by regions, which can be proven through traditional research survey. It happens because these studies uses past data, while the proposed work can use past or current data. Monthly data was used in the study of proposed case in March 2016. This collection and analysis of daily data can identify possible trends and allows targeting of strategic actions in general. These actions carried out by favorable movements or contrary to *impeachment*.

It is important to note that the case study could be carried out in relation to other periods for the *Twitter* posts. In this new study you can be dispensed the phases by training and testing, since the

[5] **Check the location and time of the demonstrations of March 13 — Congress in focus** - Link for news: http://congressoemfoco.uol.com.br/noticias/confira-o-horario-e-o-local-das-manifestacoes-de-13-de-marco/.

[6] **Manifestations against the coup are scheduled for this Thursday (31/03)** - Link for news: http://www.pragmatismopolitico.com.br/2016/03/manifestacoes-contra-o-golpe-estao-agendadas-para-esta-quinta-feira-3103.html.

[7] **Shapefiles** - It is a well-known format for storing geospatial resources in files, site: http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[8] **49% of mentions on social networks are pro-impeachment, study shows — Radar Online — VEJA.com** - Link for news: http://veja.abril.com.br/blog/radar-on-line/sem-categoria/49-das-mencoes-em-redes-sociais-sao-pro-impeachment-mostra-estudo/.

[9] **Page Home - Torabit** - Link for site: http://www.torabit.com.br/.

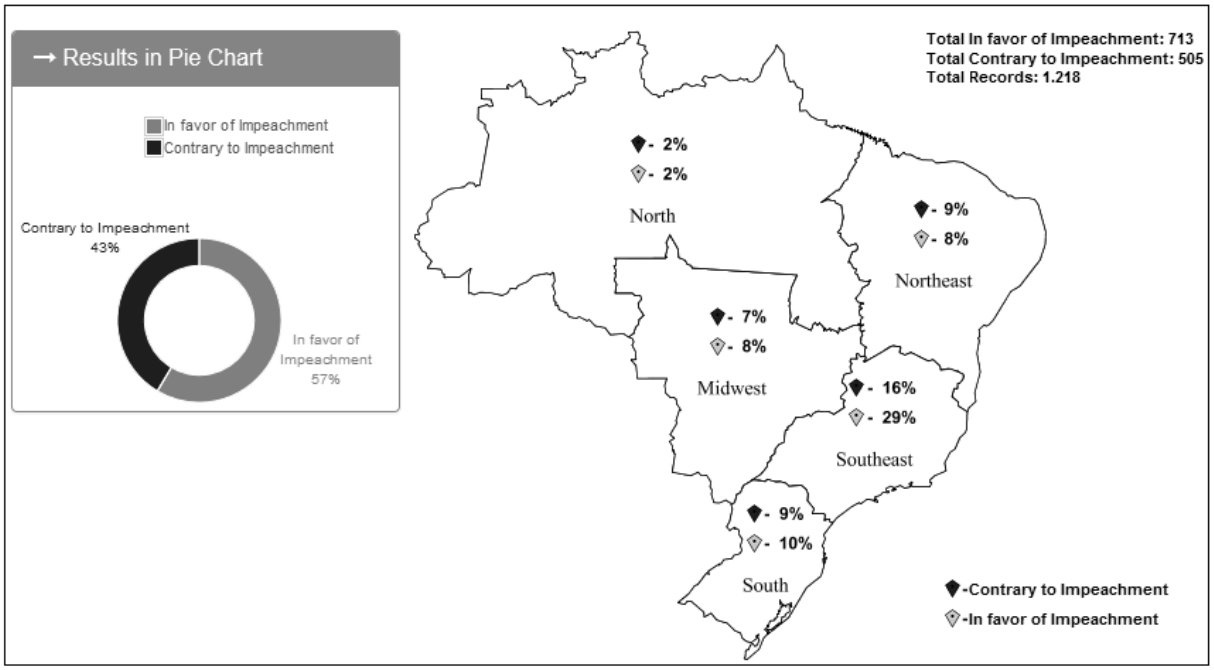

Figure 3: Approach to implementing proposed policy opinion analysis: map with trends *impeachment* for regions of Brazil.

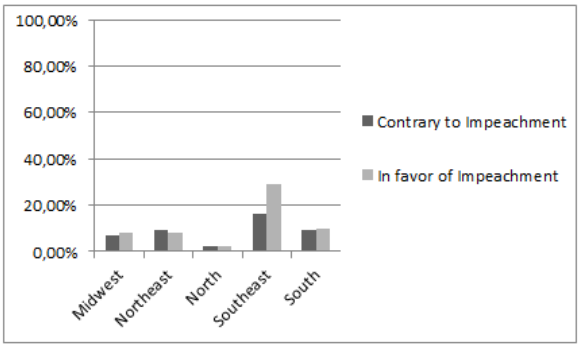

Figure 4: Graphic trends for the *impeachment* for regions of Brazil, based on the proposed approach.

models were obtained in the previous study and the same could be reused for other periods. With the application of this new study results should verify possible trends for the process of *impeachment* for the selected period.

7 Conclusion

It is concluded the proposed approach achieved the goal of providing a solution based on opinion mining to identify policy trends according to public opinion. The result obtained with the proposed work to collect and process data from the *Twitter* is valid and resembles with the other work.

Probably, the results of this study conclude that the data of social networks, such as the *Twitter* (available through its API), can be used for public opinion research purposes that go beyond a simple mechanism for broadcast content. Remember that these networks provide a range of opportunities to detect where and when a topic of interest is being discussed. Monitoring on a particular topic and location, allows researchers to compare it with other collected data using different means. As it was shown in Case Study proposed.

Possibly, the results can be improved through the use of other methods for feature extraction or combination of these, such as: *Latent Semantic Indexing Principal Component Analysis and others*. These improvements can come with implementations of these methods in future work.

References

I. Arel, D.C. Rose, and T.P. Karnowski. 2010. Deep machine learning - a new frontier in artificial intelligence research [research frontier]. *Computational Intelligence Magazine, IEEE*, 5(4):13–18, Nov.

Leo Breiman. 2001. Random forests. *Mach. Learn.*, 45(1):5–32, October.

David P. Currie. 1998. The first impeachment: The constitution's framers and the case of senator william blount. *American Journal of Legal History*, 42(4):427–429.

Nugroho Dwi Prasetyo and Claudia Hauff. 2015. Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pages 149–158, New York, NY, USA. ACM.

Usama Fayyad, Gregory Piatetsky-shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54.

Eibe Frank and Remco R. Bouckaert. 2006. Naive bayes for text classification with unbalanced classes. In *Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases*, PKDD'06, pages 503–510, Berlin, Heidelberg. Springer-Verlag.

Z. Huang and Z. Xu. 2011. A method of using geoserver to publish economy geographical information. In *Control, Automation and Systems Engineering (CASE), 2011 International Conference on*, pages 1–4, July.

Thorsten Joachims. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK. Springer-Verlag.

Andreas Jungherr. 2013. Tweets and votes, a special relationship: The 2009 federal election in germany. In *Proceedings of the 2Nd Workshop on Politics, Elections and Data*, PLEAD '13, pages 5–14, New York, NY, USA. ACM.

Jayashri Khairnar and Mayura Kinikar. 2013. Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*, 3(6):1 – 6.

Vasileios Lampos. 2012. On voting intentions inference from Twitter content: a case study on UK 2010 General Election. *arXiv preprint arXiv:1204.0423*.

Mahmut Onur Karslıoğlu Nuhcan Akçit, Emrah Tomur. 2014. Geographical information systems participating into the pervasive computing. In *GEOProcessing 2014, The Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services*, pages 129–137. ThinkMind, March.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Stephen Robertson. 2004. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, July.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427 – 437.

Ian Sommerville. 2006. *Software Engineering: (Update) (8th Edition) (International Computer Science)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Tiara, M.K. Sabariah, and V. Effendy. 2015. Sentiment analysis on twitter using the combination of lexicon-based and support vector machine for assessing the performance of a television program. pages 386–390, May.

H. Zhang. 2011. The research of web mining in e-commerce. In *Management and Service Science (MASS), 2011 International Conference on*, pages 1–4, Aug.