Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

A Comparative Study for Sentiment Analysis on Election Brazilian News

Caio Magno Carvalho, Hitoshi Nagano, Allan Kardec Barros

[1]Laboratório de Processamento da Informação Biológica
Universidade Federal do Maranhão
São Luís – Maranhão – Brasil

magno.caio91@gmail.com, hitoshinagano@me.com, allan@dee.ufma.br

Abstract. *Brazilian news media have been accused to be biased over the years, supporting some political parties and its agendas. To judge this statement as truth or lie is a hard task due its subjectivity. In election periods, this controversy become stronger given the influence of the media in the public opinion. Sentiment Analysis could be a useful tool for evaluate political news. Here is proposed a comparative study test between three learning algorithms (Naïve Bayes, SVM and MaxEnt) and three feature selection methods (Chi-Square, CPD and CPPD) for classifying texts related to president and governor of São Paulo 2014 elections in Brazil.*

Resumo. *A midia brasileira tem sido acusada ao longo dos anos de favorecer algumas entidades politicas e suas campanhas. Avaliar a verdade dessa afirmação não é uma tarefa simples dado o grau de subjetividade de quem avalia. Em periodos de eleição, essa controvérsia se torna mais acentuada visto a influência que a mídia exerce na opinião pública. Análise de Sentimento pode ser uma ferramenta útil na avaliação de notícias políticas. Este trabalho propõe um estudo comparativo de desempenho de três algoritmos de aprendizagem (Naïve Bayes, SVM e MaxEnt) e de três métodos de seleção de atributos (Qui Quadrado, CPD e CPPD) para classificação textos relacionados às eleições de 2014 para presidente e governador de São Paulo.*

1. Introduction

There is always a claim among brazilian people about the bias in the news media concerning some political agendas [Soares 2004], [Porto 2007]. Today, with wide Internet access, the population is able to freely expose their opinion related to these subjects as well to obtain more information to understand the current political scenario [Shirky 2011]. Blogs and social media have been the place of an endless controversy between left and right political supporters. Each group accuses the mainstream news agencies of favoring his opposition. This controversy is highlighted especially in elections period [1]. Some people state that mainstream agencies favor some political parties in this period, while other media vehicles, self-called independent, are neutral and free of any political bias.

[1]http://www1.folha.uol.com.br/poder/2014/10/1537985-sede-da-abril-e-pichada-em-protesto-contra-reportagem-da-veja.shtml, acessado em 15/05/2017

In this political context, the computational methods provided by *Natural Language Processing* through *Sentiment Analysis* have building applications addressed to this political issues. An example could be [Carvalho et al. 2011] which built a corpus of political online comments for mining positive opinions. [Park et al. 2011] also targets the online comments for predict political orientation. However, identify an opinion in a newspaper or online news is a complex task for a human due to the inherent bias present in the evaluator personal preferences. Sentiment Analysis could bring a less biased view for evaluate the opinion/sentiment associated to an article.

Under this argument, this work propose to analyze and predict the sentiment of a previously labeled *corpus* found in [de Arruda et al. 2015] which consists in collection of online news about 2014 Brazil elections. We intend to compare the performance of three classification algorithms (Naïve Bayes, SVM and MaxEnt) and evaluate three methods of feature selection (Chi-Square, Categorical Proportional Difference [Simeon and Hilderman 2008], Categorical Probability Proportional Difference[Agarwal and Mittal 2012]) and how it affects the classification task.

This work is organized as follows: the section two expose some works related to sentiment analysis and its applications on online news and politics, section three introduce the methodology used in this work emphasizing the methods for feature selection. In the section four we present the obtained results and its discussions followed by conclusions in section five.

2. Related Work

Sentiment Analysis is the field of NLP that develop algorithms that are capable to classify documents according to sentiment/opinion expressed about some topic [Pang et al. 2008] in them. These algorithms are very often used to analyze product and movie reviews, as showed in [Pang et al. 2002] and [Pang and Lee 2008]. An extensive review about Sentiment Analysis could be found in [Schouten and Frasincar 2016] and [Pang et al. 2008]. There are several works for sentiment analysis using *twitter* and other plataforms of *microblogging* as data source in a wide range of applications. In [Moraes et al. 2015] is built a *corpora* from *twitter* posts about 2014 world cup. It describes how data was collected, cleaned and annotated for posterior applications. The online newspaper comments are also a target for some works of SA. One can be found in [Park et al. 2011] which take the news comments of a online newspaper and predict its political orientation (conservative, liberal or vague) using a TF-IDF feature transformation and Support Vector machine to classify data. In [Tumitan and Becker 2014] the comments of a well-known brazilian newspaper, *Folha de São Paulo* [1], were used to build a sentiment time series about elections for governor, mayor and president in periods from 2010 until 2012. Sentiment classification for each comment was done by Sequential Minimal Optimization (SMO), an algorithm for Support Vector Machines (SVM), using unigrams as features and applying TF-IDF transformation to them. The sentiment classification performance achieved was 81.37% for 2010's comments and 83.24% for 2012's election comments. The work done by [Jose and Chooralil 2015] present a sentiment analysis enhanced tool for also predicting election results. The method used for classify data was SentiWordNet, a word graph which assigns to each english word a mesure for it sentiment (positive, negative or

[1] http://www.folha.uol.com.br/

neutral) between zero or one.

However these works only targets online comments which are highly subjective texts, i.e., very opinionated. On the other hand, articles published by news agencies seek to be more objective than subjective, or less opinionated. Therefore, just a few works analyze journalistic texts. A plausible explanation can be found in [Padmaja et al. 2013] stating that this kind of texts are not simple to classify given its intended neutrality and syntactical similarity with other texts from the same kind.

Though the interest in Sentiment Analysis is growing in the recent years, there is a lack of research, tools and material and human resources addressing applications for brazilian portuguese language [Vieira and Lima 2001] [Pardo et al. 2010]. Related to sentiment analysis in online news processing, there are some works for European Portuguese as [Morgado 2012] and for Brazilian Portuguese as [Dosciatti et al. 2013], [Martinazzo et al. 2011] and [Alvim et al. 2010]. In order to build a annotated *corpora* for brazilian elections, the work done in [de Arruda et al. 2015] collected 131 articles from 5 different sources.

There are still a lack of works for analyze political news in Brazil. Regarding this issue this present work intends to perform a sentiment analysis using the *Corpus Viés*, a corpora built and manually annotated by [de Arruda et al. 2015]. This corpus consists in a set of online news articles splitted by paragraphs where each one is labeled according the expressed sentiment. We use this dataset in order to evaluate machine learning algorithms and feature selection methods to deal with this kind of data in Brazilian Portuguese.

3. Methodology

In this section, we describe our approach to compare and evaluate the feature selection methods and the learning algorithms. We want to classify each labeled paragraph of *Corpus Viés* correctly in one of used classes postive, negative or neutral. We use three methods of feature selection: chi-square [Sharma and Dey 2012], categorical proportional difference [Simeon and Hilderman 2008] and Categorical Probability Proportion Difference [Agarwal and Mittal 2012]. Three learning algorithms were used to classify preprocessed data. The chosen classifiers were Naïve Bayes, Support Vector Machines and Maximum Entropy.These algoritms are very popular in text classification tasks as presented in [Schouten and Frasincar 2016].

We use Python programming language with the SciKit-Learn[2], for the applications of machine learning, and NLTK[3] packages, for the implemented tools of natural language processing.

The dataset used here, *Corpus Viés*, was built by [de Arruda et al. 2015] and it consists in a collection of 131 online news articles about 2014 elections for governor of São Paulo and for president of Brazil. These articles were obtained from five well-know sources in Brazil: *Veja*, *Estadão*, *Folha*, *G1* and *Carta Capital*. The articles are splitted by paragraphs. Each one was manually labeled by four annotators with respect the sentiment orientation presented in the text. The used labels are positive (”PO”), negative (”NE”) and neutral (”NE”). This dataset have 1042 labeled paragraphs which 310 are positive,

[2]scikit-learn.org
[3]www.nltk.org/

391 are negative and 341 are neutral. The *Corpus Viés* is originally stored in XML format and here was converted and stored in a MySQL database in order to facilitate the access and queries to the texts and its labels. Before feature extraction and training algorithms, we remove the stopwords, normalize the text eliminating accents and we also replace all present numbers by a token "NUMBER".

In this work, we use the *bag-of-ngrams* as feature extraction method. The features extracted can be unigrams, bigrams or both. Trigrams and higher order ngrams are not used due to exponential increasing in the number of features. This method is described as follows.

3.1. Feature Extraction with NGrams

All techniques used here to extract features are based in ngrams. A ngram is a sequence of n words extracted from a given text. When just one word is picked at time, this sequence is called *unigram*. We could also pick sequences of two words from text. Each sequence picked in this way is called *bigram*.

Here we used the *bag of words* method, that consist in the frequency that a given ngram appears in a certain document. For a given unigram (one word) set v, also called *vocabulary*, we assign an index number i for each word according to (1), where w_i is the word which has index number i and n is the number of words in vocabulary.

$$v = \{w_1, w_2, ...w_i, ..., w_n\}, i, n \in \mathbb{N} \tag{1}$$

We can represent a given document as vector d with the same dimension n of vocabulary. Each element c_i from vector d stores the counting frequency of w_i from vocabulary v in the document as described in (2).

$$d = [c_1, c_2, c_3, ..., c_i, ..., c_n], i, n \in \mathbb{N} \tag{2}$$

The document vector representation depends directly on the chosen vocabulary. Very often, the vocabulary used consists in a set of all unigrams extracted from *corpora*. Using feature selection methods we can reduce the original vocubuary for smaller set of ngrams which are more informative for the class distinction.

The Chi-Square method measures the level of dependency between ngram and classes. If that ngram is frequent in many classes, the Chi-Square value will be low, if this ngram occurs just in a few classes, then Chi-Square value will be high [Haddi et al. 2013]. Categorical Proportional Difference (CPD) measure how much a term contribute in discriminating the class. It was originally designed for text categorization tasks [Simeon and Hilderman 2008]. The CPD of a term for a class is between -1 and 1, which -1 indicates that term never occur in that class, and 1 indicates that term occurs only in that class. So, we select just the ngrams/terms which CPD is above a established threshold. Categorical Probability Proportional Difference (CPPD) is a improvement of the former method. CPPD measures the level of belongingness for a ngram/term inside a given class [Agarwal and Mittal 2012]. It means while CPD measure just how that ngram/term is spread among the classes, CPPD rank them also by measuring the probability of same ngram/term occur in that class. So, we select just the ngrams/terms which

CPD and the probabiliy is above a established thresholds. In this work we use CPD = 1 as a constraint for CPPD and get n best probability ranked terms, where n is the number of features that we want to select.

We perform two experiments to evaluate both feature extraction and feature selection methods. We do not use any feature selection method in the first experiment in order to evaluate what is the best paragraph representation: unigram, bigram or both. In the second experiment we intend to evaluate the impact of feature selection on the learning algorithms. We apply the feature selection methods varying the number of selected features. The results for both experiments are obtained by applying cross validation with 10 folds, computing the accuracy for each fold and taking the mean at the end of experiment. All feature selection methods here are supervised, then we use 90% of corpora for train each method.

4. Results and Discussion

The obtained results for the first experiment are exposed in Table 1. We note that there is no improvement when using bigrams instead unigrams. From this experiment we see that a feature vector composed by both unigrams and bigram is the best representation for the news paragraphs. This feature combination affects positively all classifiers, but at cost of high dimensionality.

Table 1. Results for news classification without feature selection

Features	Dimensionality	Accuracy		
		Naïve Bayes	SVM	MaxEnt
Unigram	5344	59.1	55.16	56.12
Bigram	18958	55.46	52.77	50.95
Unigram + Bigram	**24302**	**59.3**	**57.18**	**56.99**

In the second experiment, we use just the combination of unigram and bigrams as features and apply the feature selection methods. The impact of each feature selection in the paragraphs classification. Each figure describe the performance for all classifiers varying the number of selected features from 1000 to 5000. Using Chi-Square, we note in the Figure 1 that the best performance is achieved by Naïve Bayes classifier with 71.2% accuracy and 2000 features.

When using CPD (Figure 2), we get worst performances for all classifiers in this range of number of features. The maximum is achieved by Naïve Bayes again, but with 5000 features, more than twice chi-square features. In other hand, CPPD outperformed all methods in all range of number of features, as we can see in Figure 3.

The best performance is achieved by MaxEnt model, which has its maximum using 4000 features. We also note that the accuracy does not vary too much between 2000 and 5000 features. It means that we can use less features without compromising the accuracy.

The observed results show that MaxEnt using the combination of unigrams and bigrams selected by the CPPD method is the most efficient method to classify this kind of texts. However is possible to get some interesting conclusions about the other results found in table 1.

Figure 1. The classifiers performance with Chi-Square feature selection varying the number of selected features.

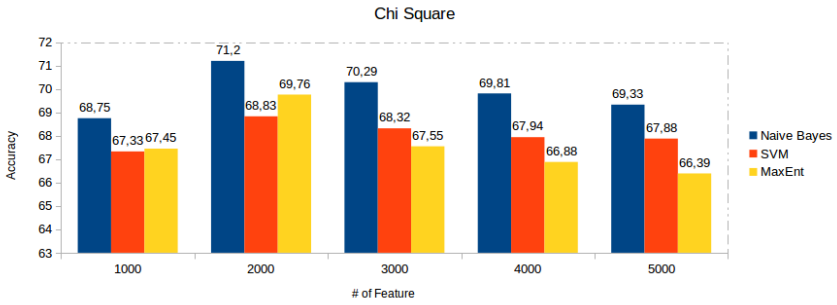
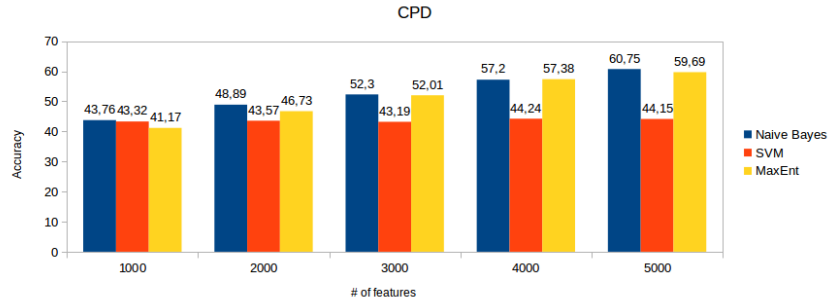


Figure 2. The classifiers performance with CPD feature selection varying the number of selected features.

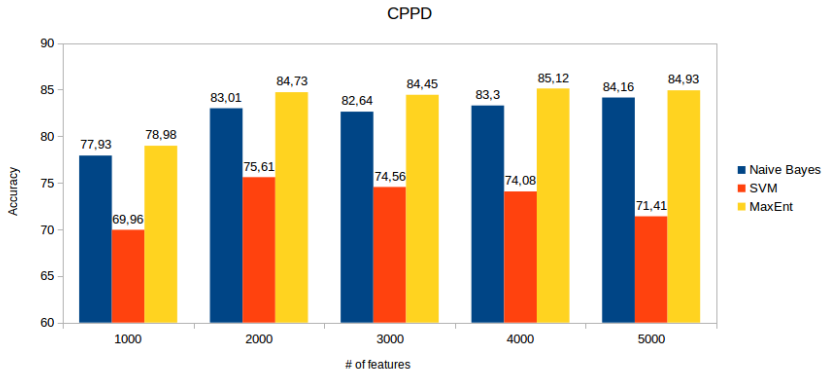


The commom sense states that bigram features bear more information than unigram features. This information is called context for the fact that bigrams (and other ngrams, except unigram) includes the word around the principal feature. It could be a clue about the real meaning of that word [Pedersen 2001]. The result has shown that classifiers which use just unigrams as features has a better perfomance than those which use just bigrams. Hence, if the contextual information is important to distinguish these texts, use only bigrams as feature does not increase the amount of useful information comparing with unigrams. But when this two kind of features are jointly used, the algorithm perform better than using only one of them. It means that for this kind of texts, we get more contextual information using both unigrams and bigrams.

5. Conclusions

In this work we presented a comparative study for learning algorithms and feature selection methods applied to sentiment analysis on elections online news. We conduct two evaluations: the first one aims to know the best paragraph representation; the second one, what is the best classifier and the best feature selection algorithm. We conclude that MaxEnt is the best classifier when is applied to a paragraphs represented by a combination of unigrams and bigrams selected by Categorical Probability Proportional Difference using a CPD threshold equals to 1. As future work, we intend to collect more data not only

Figure 3. The classifiers performance with CPPD feature selection varying the number of selected features.



about elections but other subjects related to politics and from more sources. We also aim to approach the problem of sentiment analysis in a aspect level to make a evalution about the bias in media vehicles.

Though there are few works related to that subject, the results presented here are comparable with the those presented in published works. This reasearch aims colaborate with brazilian sentiment analysis scenario in online news, providing preliminary results through comparison of machine learning tools and feature extraction techiniques also serving as baseline for applications of political opinion mining in online news and bias assesment.

Acknowledgement

We would like to thank those who gathered and annotate the *Corpus Viés* used in this work. We also thank to CAPES and FAPEMA for financial support.

References

Agarwal, B. and Mittal, N. (2012). Categorical probability proportion difference (cppd): a feature selection method for sentiment classification. In *Proceedings of the 2nd workshop on sentiment analysis where AI meets psychology, COLING*, pages 17–26.

Alvim, L., Vilela, P., Motta, E., and Milidiú, R. L. (2010). Sentiment of financial news: a natural language processing approach. In *1st Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology, Buenos Aires*.

Carvalho, P., Sarmento, L., Teixeira, J., and Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 564–568. Association for Computational Linguistics.

de Arruda, G. D., Roman, N. T., and Monteiro, A. M. (2015). An annotated corpus for sentiment analysis in political news.

Dosciatti, M. M., Ferreira, L. P. C., and Paraiso, E. C. (2013). Identificando emoçoes em textos em português do brasil usando máquina de vetores de suporte em soluçao multiclasse. *ENIAC-Encontro Nacional de Inteligência Artificial e Computacional. Fortaleza, Brasil*.

Haddi, E., Liu, X., and Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26–32.

Jose, R. and Chooralil, V. S. (2015). Prediction of election result by enhanced sentiment analysis on twitter data using word sense disambiguation. In *Control Communication & Computing India (ICCC), 2015 International Conference on*, pages 638–641. IEEE.

Martinazzo, B., Dosciatti, M. M., and Paraiso, E. C. (2011). Identifying emotions in short texts for brazilian portuguese. In *IV International Workshop on Web and Text Intelligence (WTI 2012)*.

Moraes, S. M., Manssour, I. H., and Silveira, M. S. (2015). 7x1pt: um corpus extraído do twitter para análise de sentimentos em língua portuguesa.

Morgado, I. C. (2012). Classification of sentiment polarity of portuguese on-line news. In *Proceedings of the 7th Doctoral Symposium in Informatics Engineering*, pages 139–150.

Padmaja, S., Fatima, S. S., and Bandu, S. (2013). Analysis of sentiment on newspaper quotations: A preliminary experiment. In *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, pages 1–5. IEEE.

Pang, B. and Lee, L. (2008). Using very simple statistics for review search: An exploration. In *COLING (Posters)*, pages 75–78.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Pardo, T. A., Gasperin, C. V., Caseli, H. M., and Nunes, M. d. G. V. (2010). Computational linguistics in brazil: an overview. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 1–7. Association for Computational Linguistics.

Park, S., Ko, M., Kim, J., Liu, Y., and Song, J. (2011). The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 113–122. ACM.

Pedersen, T. (2001). A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Porto, M. (2007). Tv news and political change in brazil: The impact of democratization on tv globo's journalism. *Journalism*, 8(4):363–384.

Schouten, K. and Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Sharma, A. and Dey, S. (2012). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, pages 1–7. ACM.

Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign affairs*, pages 28–41.

Simeon, M. and Hilderman, R. (2008). Categorical proportional difference: A feature selection method for text categorization. In *Proceedings of the 7th Australasian Data Mining Conference-Volume 87*, pages 201–208. Australian Computer Society, Inc.

Soares, G. A. D. (2004). A américa latina na imprensa brasileira. *Opinião Pública*, 10(1):63–90.

Tumitan, D. and Becker, K. (2014). Sentiment-based features for predicting election polls: a case study on the brazilian scenario. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 2, pages 126–133. IEEE.

Vieira, R. and Lima, V. L. (2001). Lingüística computacional: princípios e aplicações. In *Anais do XXI Congresso da SBC. I Jornada de Atualização em Inteligência Artificial*, volume 3, pages 47–86. sn.