# Investigating Opinion Mining through Language Varieties: a Case Study of Brazilian and European Portuguese *tweets*

**Douglas Vitório[1], Ellen Souza[1,2], Ingryd Teles[1,3], Adriano L. I. Oliveira[2]**

[1]MiningBR Research Group, Federal Rural University of Pernambuco (UFRPE)
Serra Talhada – PE – Brazil

[2]Center of Informatics, Federal University of Pernambuco, (CIn-UFPE)
Recife – PE – Brazil

[3]University of Pernambuco (UPE)
Recife – PE – Brazil

douglas.alisson@ufrpe.br, ellen.ramos@ufrpe.br, ivstp@ecomp.poli.br,

alio@cin.ufpe.br

***Abstract.*** *Portuguese is a pluricentric language comprising variants that differ from each other in different linguistic levels. It is generally agreed that applying text mining resources developed for one specific variant may produce a different result in another variant, but very little research has been done to measure this difference. This study presents an analysis of opinion mining application when dealing with the two main Portuguese language variants: Brazilian and European. According to the experiments, it was observed that the differences between the Portuguese variants reflect on the application results. The use of a variant for training and another for testing brings a substantial performance drop, but the separation of the variants may not be recommended.*

## 1. Introduction

The recent and exponential growth of social media and user-generated content (UGC) on the Internet provides a huge quantity of information that allows discovering the experiences, opinions, and feelings of users or customers. The volume of this kind of data has grown from terabytes to petabytes [Marine-Roig and Clavé 2015].

Understanding what people are thinking or their opinions is fundamental for decision making, mainly in the context where people express their comments voluntarily [Firmino Alves et al. 2014]. However, it is impossible for humans to fully understand UGC in a reasonable amount of time, which is why there has been a growing interest in the scientific community to create systems capable of extracting information from it [Balazs and Velásquez 2016].

According to [Liu and Zhang 2012], opinion mining (OM), also known in the literature as sentiment analysis, is the field of study that analyzes people's sentiments, opinions, evaluations, attitudes, and emotions about entities, such as products, services, organizations, individuals, issues, events, topics, and their attributes, expressed in textual input. This is accomplished through the opinion classification of a document, sentence or feature into categories, e.g. 'positive', 'negative', and 'neutral'. This kind of classification is referred to in the literature as sentiment polarity or polarity classification.

Portuguese is one of the most spoken languages in the world, with almost 270 million speakers in ten countries[1], and it is also the fifth most used language on Twitter [Statista 2013]. Portuguese is a pluricentric language that presents variants, also known in the literature as varieties, that differ subtly from each other in different linguistic levels, such as lexical, syntactic, and orthographic [Castro et al. 2016]. These variants, especially the Brazilian and European ones, have specific Natural Language Processing (NLP) resources and tools for many tasks and it is generally agreed that applying text mining resources developed for one specific variant may produce a different result in another variant, but very little research has been done to measured this difference [Fonseca and Aluísio 2016].

There are several OM applications using Twitter data and many others that deal with multilingual scenarios, using *tweets* or not [Ravi and Ravi 2015, Balahur and Perea-Ortega 2015]. Also, there are a great amount of language identification studies [Castro et al. 2017] that focus on language varieties, including the Portuguese ones. However, although we have performed an extensive search, no studies analyzing language varieties and its differences when applied to opinion mining were found.

In this sense, this study presents an analysis of OM when dealing with the two main Portuguese language variants: Brazilian and European. The objective is to investigate whether the language variant influences the application performance. Therefore, two annotated corpora for OM are provided: one containing *tweets* written in Brazilian Portuguese and another containing *tweets* written in European Portuguese. The research was done by crossing the language variants during the classifiers' training and testing steps resulting in nine different configurations. Furthermore, three supervised machine learning classifiers were evaluated together with a smoothed pre-processing technique.

The rest of this paper is structured as follows: Section 2 discusses the related work. Section 3 describes the method used and Section 4 presents the experimental setup. In Section 5, the findings are reported and discussed. Finally, Section 6 draws the conclusions.

## 2. Related Work

Two studies [Fonseca and Aluísio 2016, Garcia et al. 2014] focused on Part-of-Speech (PoS) tagging applications through the two main Portuguese language varieties: Brazilian and European. In [Fonseca and Aluísio 2016], the authors used corpora containing news from Brazil and Portugal to evaluate a PoS tagger in cross-variant settings. They used word embeddings, learned from texts in either variant, resulting in twenty configurations, which differ in three variables: the variant used for training, the variant used for testing, and the origin of the embedding model. The best result (accuracy of 96.85%) was achieved using the Brazilian variant for training and testing with embedding models from both variants.

[Garcia et al. 2014] evaluated a PoS tagger trained with several combinations of Brazilian and European corpora and tested in the two main Portuguese variants, besides the African variant (from Angola and Mozambique), with samples before and after the

[1]Brazil (202,656,788), Mozambique (24,692,144), Angola (24,300,000), Portugal (10,813,834), Guinea-Bissau (1,693,398), East Timor (1,201,542), Equatorial Guinea (722,254), Macau (587,914), Cabo Verde (538,535) and São Tomé e Princípe (190,428).

Portuguese Language Orthographic Agreement of 1990, which unified the spelling system from the Portuguese-language countries. They built the train models combining one variant or both with dictionaries, where the best model (EPtag) used the European variant and achieved a micro-average accuracy of 96.85%. The best result was also achieved by EPtag: when tested in the Angola dataset, it reached an accuracy of 98.18%.

A study of OM, considering Arabic Language colloquial varieties [Al-Obaidi and Samawi 2016], performed sentiment analysis evaluating three classifiers using as corpus containing online reviews from five different Arabian cities, where each one has a different dialect. The best result (F-measure of 86.75%) was achieved using a Maximum Entropy classifier with N-gram models. However, although they have considered that five different Arabic dialects were present in the dataset, the impact of their differences in OM was not measured, since they did not perform experiments with those dialects separately.

## 3. Method

Figure 1 presents the method adopted in this study. It has four steps which are explained in the following subsections.

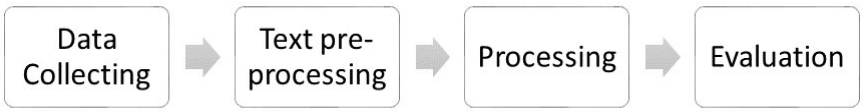


**Figure 1. Proposed method.**

### 3.1. Data collecting

The creation of corpora was performed in three steps which are detailed bellow. Both corpora are publicly available[2].

#### 3.1.1. *Tweets* extraction

We collected 5,424 *tweets* published by 2,691 different users on May 13, 2016, from which 2,544 were written in Brazil and the 2,880 others were written in Portugal. The extraction was performed using *Tweepy*, a Python library for accessing the Twitter API. This API searches samples of the data published in the past 7 days [Twitter 2017].

For the *tweets* collection, we did not look specifically for *hashtags*, *users*, or keywords. Instead, we selected *tweets*, ignoring the *retweets*, i.e., *tweets* that were posted by a user and reposted by other users, published in Brazil and in Portugal using the geography search from the Twitter API, which filters *tweets* by country. Therefore, there were two extraction streams: one for Brazil and another for Portugal. To build the corpora, we also considered that *tweets* published in Brazil were written in Brazilian Portuguese, as well as *tweets* published in Portugal were written in European Portuguese. In the subsection 3.1.2 we discuss the treatment used in cases of *tweets* written in a language other

[2]http://miningbrgroup.com.br/index.php/resources/

than Portuguese. In addition, as we did not look for specific domains, such as Political or Business, the collected *tweets* included different subjects, which may have influenced the classifiers' accuracies [Pang and Lee 2008].

### 3.1.2. Tweets filtering

After the extraction, *tweets* containing solely *hashtags*, URLs, and/or *emoji*, i.e., *tweets* that do not have words in their messages, were manually excluded. Moreover, even collecting only *tweets* published in Brazil and in Portugal, we found several ones written in a language other than Portuguese, such as English, Spanish, Italian, and French. Those *tweets* were also excluded. This filtering process excluded 1,344 *tweets*, resulting in a total of 4,000 *tweets*: 2,000 for each Portuguese variant.

### 3.1.3. Manual Annotation

We manually annotated all the remaining *tweets* as:

- Positive: *tweets* containing positive sentiments or opinions.
- Negative: *tweets* containing negative sentiments or opinions.
- Mixed: *tweets* presenting both positive and negative opinions.
- Neutral: *tweets* which do not present sentiments or opinions, i.e. objective text.

Five researchers participated in the process. The first annotator classified all the 4,000 *tweets*; after it, the four remaining annotators were divided into two pairs, where the members of each pair classified sets of 2,000 *tweets*. So, each *tweet* was classified by three different annotators.

The final polarity of each *tweet* was defined as the polarity assigned to it by the majority of the three annotators. In the cases where the three annotators disagreed, i.e., each one classified the *tweet* as a different polarity, the first annotator decided the *tweet* polarity.

We computed the Fleiss' Kappa coefficient [Fleiss 1971] in order to discover the agreement between the three annotators. A total of 52.34% was achieved for the Brazilian corpus and 54.25% for the European corpus. But it is worth mentioning that, although the use of more than two annotators is advisable [Artstein and Poesio 2005], the inter-annotator agreement drops as the number of annotators increases [Das and Bandyopadhyay 2010].

Table 1 displays the distribution of *tweets* in each corpus according to their polarities, where these unbalanced datasets represent the *real feed* from Brazilian and Portuguese Twitter users.

**Table 1. Quantity of *tweets* of each polarity in the corpora.**

| Corpus | #positive | #negative | #mixed | #neutral | #total |
|---|---|---|---|---|---|
| Brazil | 390 | 509 | 61 | 1,040 | 2,000 |
| Portugal | 388 | 415 | 25 | 1,172 | 2,000 |

### 3.2. Text pre-processing

Textual information is often unstructured and without standardization rules. To prepare the text information in a way that classifiers can understand and work with, we use some pre-processing methods, such as: Tokenization, Filtering and Smoothing. The terms were structured using a Vector Space Model (VSM).

For this work, we developed a Python application, which is publicly available[3], using the Python NLTK library (Natural Language Toolkit) for the text pre-processing step. This library was provides many features to text processing.

#### 3.2.1. Tokenization

The first method used to treat a text is the Tokenization, which is the splitting of each document into words named tokens [Weiss et al. 2004]. For this application, we used the TweetTokenizer from the NLTK library, which performs the separation of specific pieces of the *tweet*, such as *hashtags*, *users*, punctuation, emoticons, among others.

#### 3.2.2. Filtering

Filtering is the process of removing some tokens of the feature vector that are considered irrelevant for the application. In this step, we removed the following ones:

- all the *users*, i.e., tokens initiated by '@';
- all the *hashtags*, i.e., tokens initiated by '#';
- and all the URLs.

#### 3.2.3. Smoothing

The Python NLTK library provides several smoothing techniques and, among them, there is the Lidstone smoothing technique. This technique makes the terms frequency distribution more uniform, ignoring very low probabilities, such as zero, or very high ones. It is not only a *leveling* method that usually prevents zero probability, but also tries to improve the accuracy of the model [Chen and Goodman 1999]. Lidstone smoothing is parameterized by a $\lambda$ value, which varies between 0 and 1. In our application, we used $\lambda = 0.1$. The studies of [Teles et al. 2016], [Castro et al. 2016], and [Castro et al. 2017] achieved the best results using this smoothing technique, justifying our choice.

### 3.3. Processing

In the processing step, which is the effective realization of the polarity classification of *tweets*, we used three machine learning algorithms: Multinomial Naïve Bayes (MNB), a Suport Vector Machine (SVM) classifier called Linear SVC, and the Logistic Regression (LR) algorithm. According to [Souza et al. 2016b] and [Souza et al. 2016a], Bayesian and SVM classifiers are the most used processing techniques for OM and for text mining with user-generated content, respectively, in the Portuguese language. And Logistic Regression also proved to be an efficient algorithm, as observed in [Teles et al. 2016].

[3]http://miningbrgroup.com.br/index.php/resources/

To implement these classifiers, we used the Scikit-learn library, which is an open source library of machine learning from Python programming language.

### 3.4. Evaluation

For the evaluation step of the configurations, a 10-fold cross-validation technique was adopted. This method divides the dataset into 10 similar parts of approximately equal size, which requires 10 rounds. In each round, nine blocks of the dataset are used for training the classifier and the remaining block is used for testing, at the end of each round, the accuracy (A) is measured. When all the rounds are completed, it is computed the average accuracy as the final result of the configuration. Due to the paper size, we only presented the results in terms of accuracy.

## 4. Experimental Setup

To run our experiments, first we normalized the annotated corpora by removing the 'mixed' class and 254 randomly selected *tweets* from the three remaining classes. The purpose of this normalization was to make the size of each class equal in both corpora and the removal of the 'mixed' class occurred due to the fact that there were only a few *tweets* annotated for this class, which could hinder the classification. Table 2 shows the final corpora.

**Table 2. Corpora used in the experiments with three classes.**

| Corpus | #positive | #negative | #neutral | #total |
|---|---|---|---|---|
| Brazil | 387 | 414 | 1,029 | 1,830 |
| Portugal | 387 | 414 | 1,029 | 1,830 |
| Datasets built for the experiments | | | | |
| BR | 258 | 276 | 686 | 1,220 |
| PT | 258 | 276 | 686 | 1,220 |
| MIX | 258 | 276 | 686 | 1,220 |

To perform the experiments, we divided the two corpora into three datasets of 1,220 *tweets* each, which were built as follows:

- BR: 1,220 *tweets* randomly selected from the Brazil corpus, i.e., this dataset contained only *tweets* written in Brazilian Portuguese;
- PT: 1,220 *tweets* randomly selected from the Portugal corpus, i.e., this dataset contained only *tweets* written in European Portuguese;
- MIX: composed by the remaining 1,220 *tweets*: 610 from the Brazil corpus and 610 from the Portugal corpus, i.e., this dataset contained *tweets* from both variants.

Table 2 also shows the class distribution of *tweets* in the three datasets.

Based on these datasets, we obtained nine configurations divided into two categories: 'same-variant', where both training and testing were performed using the same dataset; and 'cross-variant', where the training was performed using a dataset from one variant and the testing was done using a dataset from the other. The nine configurations are explained in Table 3.

**Table 3. Configurations built and executed in this study.**

| # | Configuration | Category | Variant for training | Variant for testing |
|---|---|---|---|---|
| 1. | BR-BR | same-variant | Brazilian | Brazilian |
| 2. | PT-BR | cross-variant | European | Brazilian |
| 3. | MIX-BR | cross-variant | Both | Brazilian |
| 4. | PT-PT | same-variant | European | European |
| 5. | BR-PT | cross-variant | Brazilian | European |
| 6. | MIX-PT | cross-variant | Both | European |
| 7. | MIX-MIX | same-variant | Both | Both |
| 8. | PT-MIX | cross-variant | European | Both |
| 9. | BR-MIX | cross-variant | Brazilian | Both |

## 5. Results and Discussion

Table 4 reports the accuracies reached by the classifiers (Multinomial Naïve Bayes (MNB), Linear Regression (LR), and SVC Linear) for each configuration. The configuration MIX-MIX (#7) achieved the best results.

**Table 4. Accuracies reached by each configuration with three classes.**

| # | Configuration | Training | Testing | MNB | LR | SVC |
|---|---|---|---|---|---|---|
| 1. | BR-BR | Brazilian | Brazilian | 61.72% | **64.02%** | **64.51%** |
| 2. | PT-BR | European | Brazilian | 59.92% | 60.98% | 58.93% |
| 3. | MIX-BR | Both | Brazilian | **61.56%** | 63.03% | 62.46% |
| 4. | PT-PT | European | European | 63.20% | 65.08% | **65.57%** |
| 5. | BR-PT | Brazilian | European | 60.66% | 63.11% | 62.62% |
| 6. | MIX-PT | Both | European | **63.44%** | **66.15%** | 64.92% |
| 7. | MIX-MIX | Both | Both | **65.33%** | **67.46%** | **67.46%** |
| 8. | PT-MIX | European | Both | 63.52% | 64.84% | 67.38% |
| 9. | BR-MIX | Brazilian | Both | 62.30% | 63.20% | 61.89% |

### 5.1. Discussion

According to the experiments, the European variant was ”easier” to classify than the Brazilian one. The three configurations that uses ‘PT’ as testing dataset (PT-PT, BR-PT, and MIX-PT) presented better results than the correspondent ‘BR’ configurations (in order: BR-BR, PT-BR, and MIX-BR). The European dataset also proved to be better when used to train the classifiers, as we could observe by analyzing the accuracies of the two configurations that use only one variant for training and both variants for testing (PT-MIX and BR-MIX). This may be justified by the fact of the Fleiss' Kappa coefficient is higher to our European corpus.

We could also notice that differences between the Portuguese variants reflect on the opinion mining results. This could be observed comparing the ‘same-variant’ configurations that use only one variant for training and testing (PT-PT and BR-BR) with their respective ‘cross-variant’ configurations (BR-PT and PT-BR): the ‘same-variant’ ones always achieved better results, with improvements reaching 5%.

Thus, the language variant identification is important for OM since the use of a variant for training and another for testing brings a substantial performance drop. However, the separation of the variants from mixed corpora may not be recommended, as the best results of all configurations have been achieved using the 'MIX' dataset, i.e., both variants together, and this separation is often expensive.

Although [Fonseca and Aluísio 2016] and [Garcia et al. 2014] performed studies evaluating Part-of-Speech (PoS) tagging, which has several differences from opinion mining, and used corpora containing a type of text different from the type used in our study, we can superficially compare their results with our findings, since PoS-tagging is also a classification task. Just as in our study, the European variant showed better results in [Garcia et al. 2014], while, in [Fonseca and Aluísio 2016], the Brazilian variant was easier to classify. In [Castro et al. 2016], which analyzed language identification with the Portuguese variants using *tweets*, the European variant also performed better.

In [Fonseca and Aluísio 2016] and [Garcia et al. 2014], the European datasets contain sentences longer than the Brazilian ones, which did not happen in our corpora and nor in the dataset used by [Castro et al. 2016]. So, we can not accurately determine if one or another Portuguese variant is actually easier to classify. As we did not find other studies analyzing Portuguese language varieties, when applied to opinion mining, it was not possible to point out which variables may affect the OM results, such as: language variant, documents size, numbers of unique tokens, sentences length, text domain or others.

## 6. Conclusion

In this study, a single-label and document level sentiment analysis has been performed in order to investigate whether a language variant influences the opinion mining application performance. A corpus containing *tweets* from the two main Portuguese language variants, the Brazilian from Brazil and the European from Portugal, was built for the experiments.

The research was done by crossing the language variants during the classifiers' training and testing steps resulting in nine different configurations divided into two main categories: 'same-variant', where both training and testing were performed using the same dataset; and 'cross-variant', where the training was performed using a dataset from one variant and the testing was done using a dataset from another variant. Furthermore, three supervised machine learning classifiers were evaluated together with a *smoothed* preprocessing technique.

According to the experiments, the configuration MIX-MIX, which was trained and tested using a dataset containing both Portuguese variants, achieved the best results. Thus, it was observed that differences between the Portuguese variants reflect on the application results. The use of a variant for training and another for testing brings a substantial performance drop. However, the separation of the variants may not be recommended, as the best results were achieved using a mixed dataset containing opinions from both variants together.

As no studies analyzing Portuguese language varieties applied to opinion mining were found, further studies need to be made. The text domain, for example, is a variable

which may affect the results, thus researches using another Portuguese dataset must be carried out as a way to investigate the impact of the variants in specific domains. In the same way, a dataset containing texts written in African Portuguese, i.e. the Portuguese variety spoken in African countries, such as Angola and Mozambique, should also be built and analyzed similarly as we did for the Brazilian and European Portuguese varieties.

Furthermore, the results achieved for the Portuguese language may not be the same for other languages, so it is necessary to perform investigations with datasets containing variants of different languages. And pre-processing techniques are also an important variable for the results. In this study, we only used language independent techniques and more research using pre-processing techniques specific for Portuguese should be analyzed. Experiments should also be performed for other languages variants.

## References


Al-Obaidi, A. Y. and Samawi, V. W. (2016). Opinion mining: Analysis of comments written in arabic colloquial. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1.

Artstein, R. and Poesio, M. (2005). Bias decreases in proportion to the number of annotators. In *Proceedings of the 10th conference on Formal Grammar and the 9th Meeting on Mathematics of Language*, FG-MoL '05, pages 141–150. CSLI Publications.

Balahur, A. and Perea-Ortega, J. M. (2015). Sentiment analysis system adaptation for multilingual processing: The case of tweets. *Information Processing and Management*, 51(4):547 – 556.

Balazs, J. A. and Velásquez, J. D. (2016). Opinion Mining and Information Fusion: A survey. *Information Fusion*, 27:95–110.

Castro, D., Souza, E., and Oliveira, A. L. I. (2016). Discriminating between brazilian and european portuguese national varieties on twitter texts. *Proceedings of 5th Brazilian Conference on Intelligent Systems (BRACIS'2016)*, pages 265–270.

Castro, D. W., Souza, E., Vitório, D., Santos, D., and Oliveira, A. L. I. (2017). Smoothed n-gram based models for tweet language identification: A case study of the brazilian and european portuguese national varieties. *Applied Soft Computing*.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.*, 13(4):359–394.

Das, A. and Bandyopadhyay, S. (2010). Topic-based bengali opinion summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 232—240. ACM.

Firmino Alves, A. L., Baptista, C. d. S., Firmino, A. A., Oliveira, M. G. a. d., and Paiva, A. C. d. (2014). A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 fifa confederations cup. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pages 123—130. ACM.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378—-382.

Fonseca, E. R. and Aluísio, S. M. (2016). *Improving POS Tagging Across Portuguese Variants with Word Embeddings*, pages 227–232. Springer International Publishing, Cham.

Garcia, M., Gamallo, P., Gayo, I., and Cruz, M. A. P. (2014). Pos-tagging the web in portuguese. national varieties, text typologies and spelling systems. *Procesamiento del Lenguaje Natural*, 53(0):95–101.

Liu, B. and Zhang, L. (2012). *A Survey of Opinion Mining and Sentiment Analysis*, pages 415–463. Springer US, Boston, MA.

Marine-Roig, E. and Clavé, S. A. (2015). Tourism analytics with massive user-generated content: A case study of barcelona. *Journal of Destination Marketing and Management*, 4(3):162—-172.

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

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Souza, E., Castro, D., Vitório, D., Teles, I., Oliveira, A. L. I., and Gusmão, C. (2016a). *Characterizing User-Generated Text Content Mining: A Systematic Mapping Study of the Portuguese Language*, pages 1015–1024. Springer International Publishing, Cham.

Souza, E., Vitório, D., Castro, D., Oliveira, A. L. I., and Gusmão, C. (2016b). *Characterizing Opinion Mining: A Systematic Mapping Study of the Portuguese Language*, pages 122–127. Springer International Publishing, Cham.

Statista (2013). Most-used languages on twitter as of september 2013. https://www.statista.com/statistics/267129/most-used-languages-on-twitter/. February, 2017.

Teles, V., Santos, D., and Souza, E. (2016). Uma análise comparativa de técnicas supervisionadas para mineração de opinião de consumidores brasileiros no twitter. In *Proceedings of the XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2016)*, pages 217–228. BDBComp.

Twitter (2017). The search api. https://dev.twitter.com/rest/public/search/. February, 2017.

Weiss, S., Indurkhya, N., Zhang, T., and Damerau, F. (2004). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. SpringerVerlag.