

Recognition of Brazilian Baiano and Gaucho Regional Dialects on Twitter Using Text Mining

Maria Fernanda Alvares Travassos de Avelino Novaes¹




¹Master in Information Science, Faculty of Engineering, University of Porto, Porto, Portugal (up201809035@fe.up.pt) ORCID [0000-0001-5957-7216](https://orcid.org/0000-0001-5957-7216)

Abstract

The internet has broken geographical barriers and brought people and cultures closer independent to their physical location. However, the language, idiom, dialects and accents continue to characterize individuals in their origins. The Brazilian regional dialect is the object of study of this research, which deals with linguistic corpora analyzed from a volume of data extracted on Twitter. This paper presents the results of the mining phase that makes up a first stage of the project to create a technique for recognizing the Brazilian Portuguese regional dialects. Analysis and conclusions were made only for the baiano and gaucho dialects, considering the significant size of the samples and the need to reach a diagnosis of the collect data set.

Author Keywords. Dialect, Linguistics, Text Mining, Information Retrieval, Classification.

Type: Research article

 Open Access  Peer Reviewed  CC BY

1. Introduction

At the time of Brazil's discovery, studies estimated that more than 1.000 languages were spoken in the country. In the early years of colonization, the indigenous languages were used even by the Portuguese colonists and, being spoken by almost all inhabitants of Brazil, became known as a common language disappearing almost entirely in the eighteenth century, when the Portuguese language became the official language.

With the escape of the Portuguese Court to Brazil in 1808, the Portuguese linguistic norm, which took on the dialects spoken between Coimbra and Lisbon, evolved into the dialect of Rio de Janeiro.

Language is a system formed by rules and values assimilated and stored by the speakers of a particular language community, learned from the experience and shared experience with other speakers. However, it differs from the language that identifies the idiom of one nation in relation to others and is related to the existence of a political state.

A dialect is a linguistic variety with grammatical, phonological, morphological, syntactic, semantic and lexical rules, duly known, even implicitly, by their speakers, and therefore there is no mechanism that determines their inferiority before a language. They vary according to their geographical location and, in some conditions, by their socioeconomic status.

Recognition of these linguistic variations when observing a speaker is a daily action of the human being. However, identifying them automatically using a text analysis tool becomes a complex task that requires specific information retrieval, mining and classification techniques.

For this, a variant of data mining called Text Mining emerges, defined as the process of extracting interesting and non-trivial natural language patterns or knowledge from a set of textual documents (Tan 1999), making it possible to transform this volume of unstructured data, in useful and often innovative knowledge.

The dependence of digital technologies on modern life is creating opportunities for the study of human behavior, as well as their social trends, based on what can be mined from networks. Although studies at Northeastern University (Mocanu et al. 2013) have already investigated the dynamics of languages on Twitter, for example the analysis does not evaluate or cross-check aspects such as the internet dialect, its origin and its geographical location, but specifically does not address issues involving the recognition of Brazilian regional dialects.

For this work, in order to ensure assertive results, given that the data collection base has a significant volume and the number of words to be analyzed is equally vast and distributed, only two dialects were chosen: baiano and gaucho (Leite and Callou 2010), to intensify the research in these groups, reducing the scope and ensuring an accurate outcome when exposed to the method and tools used.

The purpose of this article is to present the results obtained in the text mining process in a sample of Twitter posts on October 23rd and 24th, 2014, explaining the inferences found and the degrees of confidence and accuracy from the application of discovery of structured data knowledge (KDD) and discovery of knowledge in unstructured data (KDT). This is one of the stages of the project that proposes the creation of a technique for recognizing the Bahian and Gaucho dialects in any data collected from this social network.

This document begins with a literature review and a brief approach to the concepts that underlie this proposal and the execution of the activities contemplated in the project. The following section presents the methodology and tools used, followed by the development and results sections.

2. Literature Review

For the development of this work, some fundamental concepts are needed as language, idiom, dialect, and accent, computational linguistics, information retrieval, knowledge discovery and text mining.

2.1. Language, idiom, dialect and accent

Brazilian Portuguese, regulated by the Brazilian Academy of Letters (ABL) is what is called the variety of Portuguese language spoken by over 200 million people in Brazil, ratifying it as the most widely spoken, read and written variant in the world. Throughout the history of this country, Brazilian Portuguese has incorporated terms from the Native American and African languages, French, Castilian, Italian, German and English, which together give rise to numerous regional variations.

The language is a system whose structure is studied from a corpus also considered as a set of necessary conversions, adopted by the social body, to allow the exercise of language (Rabaça and Barbosa 1987; Rodrigues 2008).

Thus, according to Travaglia (1997, 22):

“[...] language is seen as a code, that is, as a set of signs that combine according to rules, and that is capable of transmitting a message, information from a sender to a receiver. This code

must therefore be mastered by speakers for communication to be effective. As the use of the code that is the language a social act, consequently involving two people, it is necessary that the code be used in a similar, pre-established, agreed-upon manner for communication to take place”.

Idiom is any form of expression particular from a people. It refers to the language that identifies one nation in relation to others and is related to the existence of a political state and linked to the official language of a country. References can be found that treat language and idiom as equivalent terms (Almeida 1998).

Dialet (from Greek διάλεκτος, translit. diálektos: talk, conversation, discussion by questions and answers; way of speaking, a country's own language (Houaiss, Villar, and Franco 2001) is a subset of idioms, defined by the way a language is spoken and understood in a given geographical region, determined by its own phonological, syntactic, semantic and morphological characteristics.

For a dialect to be considered an autonomous language, linguistics considers that there must be mutual understanding between at least two communities and the existence of a common linguistic corpus, in other words, literary works considered inheritance of both without the need for translation.

Note that there is an explicit difference between dialect and accent. While for Houaiss, Villar, and Franco (2001), the first is any regional variation of an idiom that does not compromise the mutual intelligibility between the main language speaker and the dialect speaker, the accent is a concept of popular use that in scientific terms do not exist, which usually designates only a change in the intonation of the word or the imperfect pronunciation of some phonemes performed by a foreigner.

2.2. Computational linguistics

Linguistics, as a science of language, deals, among other aspects, with texts, speech and dialogue. Words, which form structured sentences, are embedded in a situation, have independent predictable speakers and structures that can be formally described. Computational linguistics is the area of knowledge that explores the relationship between linguistics and informatics dealing with the computational treatment of language and natural languages for various practical purposes. It embraces a set of activities aimed at enabling communication with machines using the natural skills of human communication (Vieira and Lima 2001).

It is divided into two subareas: corpus linguistics and natural language processing (NLP), which is concerned with the study of language aimed at the construction of specific software like automatic translators, chatterbots, parsers, automatic speech recognizers, among others that interprets or generates information provided in natural language.

Corpus Linguistics, in turn, is the area of linguistics that deals with the collection and exploration of corpora, carefully collected, for the purpose of researching a language or linguistic variety through computer-extracted empirical evidence (Sardinha 2000).

According to Sanchez (1995, 8-9 cited by Vale and Tagnin 2008), corpus is:

“A set of linguistic data [...] systematized according to certain criteria, sufficiently extensive in breadth and depth to be representative of all linguistic use or any of its scopes, arranged

in such a way that it can be processed by computer, in order to provide various and useful results for description and analysis”.

2.3. Information retrieval

The predominant language on the Internet is still English, but not as early as in the 1960s, when it was born from American researches to build a robust and flawless network communication (Leiner et al. 1997). It is estimated that today there are more than 5 billion words in Portuguese on the network (Sardinha and Almeida 2008). This number can be collected from information retrieval (IR) techniques, methods and models that ensure the highest quality of information returned.

IR is one of the areas of knowledge that make up and contribute to text mining. It is basically about extracting a certain unstructured volume of data that satisfies a specific need (Shiri 2004) that is typically stored on computers. With the effective use of the internet, search engines are daily examples of these tools either for web searches or for email box searches.

However, linguistic resources are still scarce in these tools and little is observed regarding morphological and syntactic analysis, and the semantic analysis is implicitly done.

Information retrieval is the area involved in obtaining relevant documents from a particular topic. It works with a set of techniques such as indexing, searching, filtering, organizing and handling multiple languages, which serve the purpose of finding relevant data according to a specific search (Vieira and Lima 2001).

IR systems have traditionally been based on keyword or similarity search, but the most complex ones work with lexicons, knowledge bases, and ontology networks. However, they have limitations when they do not model relationships, dependencies, actions or events. Importantly, the words are vague, ambiguous and can have several meanings and express the same object through several words. Therefore, no study can disregard the use of elaborate semantic models that model the language in use, especially considering the relations and grammatical elements.

2.4. Knowledge discovery

Knowledge discovery is an automated, computer-supported process for analyzing data or information from its collection and processing with the primary purpose of enabling the acquisition of new knowledge by manipulating large databases.

However, the discovery process is strongly related to the way information is processed. In the automation of this process, there are two approaches: KDD characterized by data mining, and KDT, the basis of this project, based on text mining (Fayyad, Piatetsky-Shapiro, and Smyth 1996).

KDT has well-defined steps. Constituting its base is the collection stage: the process of searching and retrieving texts in order to form the textual database from which some kind of knowledge is to be extracted. However, the first challenge is the location of data sources. There are three main environments: the file folders found on users' disks, the tables of the various databases and the internet (Aranha and Passos 2007). The latter, with all its heterogeneity, is the data source of the proposed project, with Twitter as the source of the volume of texts collected.

stopwords or irrelevant words, those that make no difference when indexed, reduces the size of the terms and consequently increase the performance of system as a whole. All of these actions were performed from the command lines below with the result shown in [Figure 2](#).

```
install.packages("wordcloud")
install.packages("tm")
library(tm)
library(wordcloud)
library(RColorBrewer)
tweets <- Corpus(DirSource("C:/DADOS/BASE_TWITTER"))
tweets <- tm_map(tweets, content_transformer ( tolower ))
tweets <- tm_map(tweets,removePunctuation)
tweets <- tm_map(tweets,removeNumbers)
tweets <- tm_map(tweets, function(x) removeWords(x, stopwords("portuguese")))
TDM <- DocumentTermMatrix(tweets)
b <- as.matrix(TDM)
a <- sort(colSums(b),decreasing=TRUE)
d <- data.frame(word = names(a),freq=a)
pal <- brewer.pal(9, "BuGn")
pal <- pal[-(1:2)]
png("nuvem.png", width=1280,height=800)
wordcloud(d$word,d$freq, scale=c(8,.3),min.freq=2,max.words=100, random.order=F,
rot.per=.15, colors=brewer.pal(8, "Dark2"), vfont=c("sans serif","plain"))
dev.off()
```

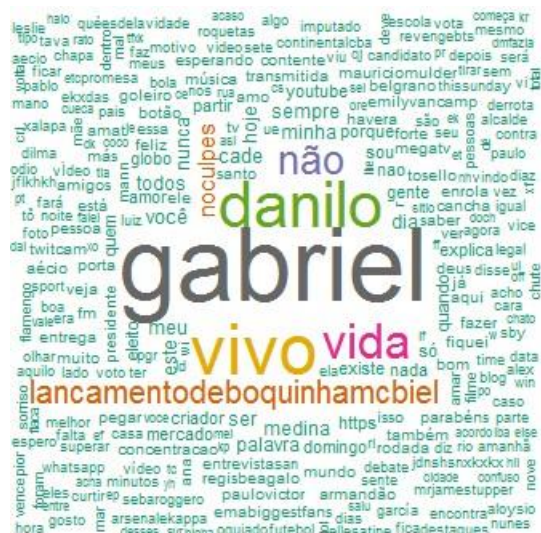


Figure 2: Word cloud based on term frequency

However, it was observed that, in the format in which the database was located, the desired classification would not be possible from R, requiring new post manipulation to perform the next steps in the Weka tool.

Thus, the consolidated data volume after the first preprocessing was re-exported, this time considering 8 attributes: text, source, place, user_name, user_screen_name, user_location, user_description, and user_lang, unlike the first time with only the text column. The choice of these attributes was due to the need for new filtering to reduce

the number of records, given the hardware limitations presented when processing the base with 13.081 found.

The user_lang = "pt" and non-empty place attributes were then filtered, reaching a sample of 435 records. Although insignificant in its quantity, the study continued as proposed for the classification to be completed and a result to be found as expected.

The next step was the creation of a specific text file for each post, that is, 435 files, manually tagged in the provided classes: Bahia, Gaucho, none (Figure 1), premise for using the SCA-Classifer tool (Matos 2010), user-friendly interface of Weka algorithms and features.

5. Results

The use of the SCA-Classifer tool allowed the execution of classification algorithms such as Naive Bayes and ID3, as well as the training of the obtained base, reaching the results shown in Figure 3 and Figure 4.

The Naive Bayes classifier is the most used for machine learning. Called naive, it assumes that attributes are conditionally independent, in other words, the information of an event is not informative about any other (Fayyad et al. 1998) and does not require a large amount of data to reach a reliable and meaningful results.

In the proposed sample, approximately 92% of instances were considered to be correctly classified, while only 8% were considered to be incorrectly classified.

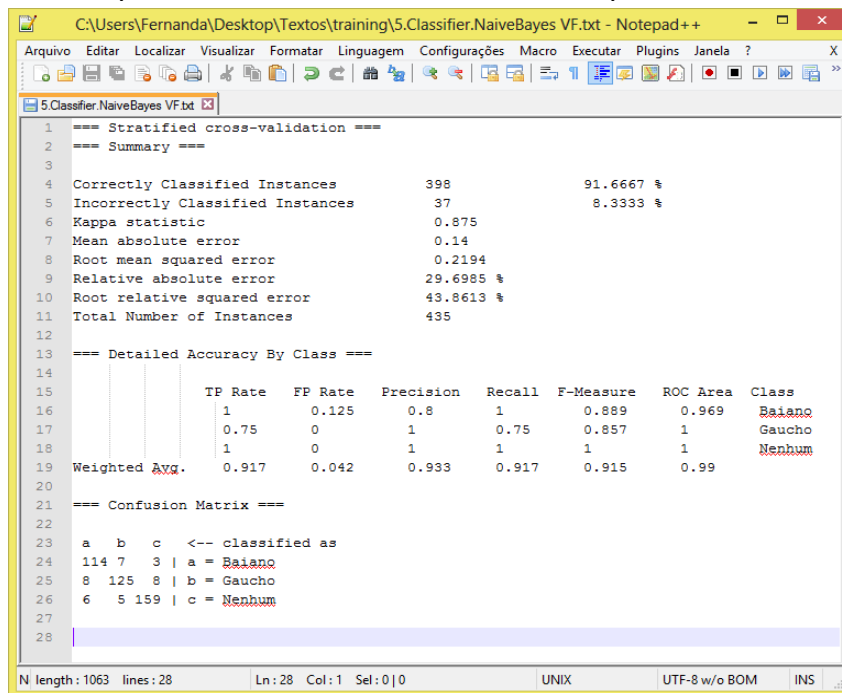


Figure 3: Classification with Naive Bayes algorithm

ID3 also applied to this data sample uses decision trees and is among the most popular inductive inference algorithms. The main choice is to select which test attribute will be used in each node of the tree. Thus, a statistical property called "information gain" is defined, which measures how a given attribute separates training examples according to their classification. ID3 uses the "information gain" to select among the candidates the attributes that will be used at each step while building the tree.

Although distributed differently, the percentage of instance classification in this algorithm was similar to that of Naive Bayes classification, as well as its precision

measurements: 0.8, 1 and 1 for baiano, gaucho and none, and confidence of 1, 0.75 and 1. for the dialects in this same order.

```

1  === Stratified cross-validation ===
2  === Summary ===
3
4  Correctly Classified Instances      393          90.3448 %
5  Incorrectly Classified Instances    42           9.6552 %
6  Kappa statistic                    0.875
7  Mean absolute error                0.0556
8  Root mean squared error            0.2357
9  Relative absolute error             11.7876 %
10 Root relative squared error        47.1142 %
11 Total Number of Instances          435
12
13 === Detailed Accuracy By Class ===
14
15                                     TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
16                                     1        0.125   0.8        1        0.889     0.938    Baiano
17                                     0.75    0        1         0.75    0.857     0.875    Gaucho
18                                     1        0        1         1        1         1        Nenhum
19 Weighted Avg. 0.917 0.042 0.933 0.917 0.915 0.938
20
21 === Confusion Matrix ===
22
23 a  b  c  <-- classified as
24 100 15  1 | a = Baiano
25  6 129  8 | b = Gaucho
26  5  7 164 | c = Nenhum
27
28
    
```

Figure 4: Classification with ID3 algorithm

6. Conclusion

The information society is the main beneficiary of computational linguistics, as studies in speech, text and image processing are advancing to make the vast amount of information currently available on the world wide web, more accessible.

The use of corpus for analysis and validation of a data sample has been used for centuries, but the use of computers to treat these samples is still incipient. Whether due to the lack of tools or the lack of professional technical knowledge in this field, the role of technology in this study has been relevant in recent years.

The larger a corpus is, the greater its representativeness. Thus, for the study of the lexicon, considering that there are rarely used words, such as the dialects presented here, the larger the corpus, the greater the possibility of the appearance of relevant terms in the sample.

For the work mentioned, which is an integral part of the Baiano and Gaucho Brazilian Dialect Recognition project on Twitter, the sample size and the format in which the texts are posted on this social network were limiting factors in the analysis. With a maximum length of 140 characters and the removal of symbols, punctuation, stopwords, and numerals, terms that are irrelevant to the proposed mining, the post is reduced to a size that can compromise the end result.

However, the application of the KDD and KDT techniques proved that it is possible to achieve a reliable end result, although there is still significant manual work in preprocessing.

The algorithms used were selected for their ability to work effectively even in small data volumes, presenting similar results in the classification performed, either by machine learning or by decision tree.

As a future proposal, it is first suggested to automate the categorization process which will allow the analysis and mining of a more significant data volume, as well as an increase in the capacity of the hardware used so as not to compromise the execution of the techniques and algorithms, enabling future real-time analysis of messages posted on both Twitter and any other social network.

References

- Almeida, N. M. de. 1998. *Gramática metódica da língua Portuguesa*. 42nd ed. Saraiva.
- Aranha, C., and E. Passos. 2007. "Automatic NLP for competitive intelligence". In *Emerging technologies of text mining: Techniques and applications*, 54-76. <https://doi.org/10.4018/978-1-59904-373-9.ch003>.
- Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth. 1996. "Knowledge discovery and data mining: Towards a unifying framework". In *KDD-96: Proceedings of the second international Conference on Knowledge Discovery and Data Mining*, 82-88. <https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf>.
- Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds. 1998. *Advances in knowledge discovery and data mining*. Menlo Park, California: AAAI.
- Houaiss, A., M. de S. Villar, and F. M. de M. Franco. 2001. "Dialect". In *Dicionário Houaiss da língua portuguesa*. Rio de Janeiro: Objetiva.
- Insite Linguistics Group. n.d. "Grupo de Linguística e Computação Cognitiva da Insite". Accessed November 25, 2019. <http://linguistica.insite.com.br>.
- Leiner, B. M., V. Cerf, D. Clark, R. Kahn, L. Kleinrock, D. Lynch, J. Postel, L. Roberts, and S. Wolff. 1997. "A brief history of the internet". <https://arxiv.org/html/cs/9901011?>
- Leite, Y., and D. Callou. 2010. *Como falam os brasileiros*. 4th ed. Rio de Janeiro: Zahar.
- Matos, P. 2010. "Metodologia de pré-processamento textual para extração de informação sobre efeitos de doenças em artigos científicos do domínio biomédico". Master's thesis, Centro de Ciências Exatas e de Tecnologia, Universidade Federal de São Carlos. <https://repositorio.ufscar.br/handle/ufscar/448>.
- Mocanu, D., A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. 2013. "The twitter of Babel: Mapping world languages through microblogging platforms". *PLoS ONE* 8, no. 4 (april): Article number e61981. <https://doi.org/10.1371/journal.pone.0061981>.
- Mugridge, Rebecca L. 2011. "Introduction to Modern Information Retrieval, 3rd ed". *Library Collections, Acquisitions, & Technical Services* 35, no. 4: 136-37. <https://doi.org/10.1080/14649055.2011.10766318>.
- Rabaça, C., and G. Barbosa. 1987. *Dicionário de comunicação*. 3rd ed. São Paulo: Ática.
- Rodrigues, Rômulo da Silva Vargas. 2008. "Saussure e a definição da língua como objeto de estudos". *Revista Virtual de Estudos da Linguagem - ReVEL* 6, no. 2 (november). http://www.revel.inf.br/files/artigos/revel_esp_2_saussure_e_a_definicao_de_lingua.pdf.
- Sardinha, T. B. 2000. "O que é um corpus representativo?" *Direct Papers*, 44.
- Sardinha, T. B., and G. M. Almeida. 2008. "A linguística de corpus no Brasil". In *Avanços da linguística de corpus no Brasil*, 17-40. São Paulo: Humanitas.
- Shiri, A. 2004. "Introduction to Modern Information Retrieval (2nd edition)". *Library Review* 53, no. 9: 462-63. <https://doi.org/10.1108/00242530410565256>.

- Tan, A. 1999. "Text Mining: The state of the art and the challenges". In *Proceedings of the PAKDD 1999 workshop on Knowledge Discovery from Advanced Databases*, 65-70.
- Travaglia, L. 1997. *Gramática e interação: uma proposta para o ensino de gramática no 1º e 2º graus*. São Paulo: Cortez.
- Vale, O., and S. Tagnin. 2008. *Avanços da linguística de Corpus no Brasil*. São Paulo: Ed. Humanitas.
- Vieira, R., and V. L. S. Lima. 2001. "Linguística computacional: Princípios e aplicações". In *Anais da IX Escola de Informática da SBC-Sul*, edited by L. Nedel, 27-61. Passo Fundo, Maringá, São José. SBC-Sul.