Portuguese Dialect Classification
(CSE 5525)

Group: João Magalhães
Arjun Bakshi

Introduction

Portuguese is the sixth most spoken language in the world, being an official or co-official language in Brazil, Portugal and other 7 countries [1]. The distance between these territories and their different cultures have created variances — mostly informal — inside the language, meaning that each region has its own dialect of Portuguese. The following table provides some examples of variances between Portuguese used in Portugal and Brazil:

BR	PT	EN
Correto, Caminhão	Correcto, Camião	Correct, Truck
Você faz	Tu fazes	You do
Dublar	Dobrar	Dub
Me diz	Diz-me	Tell me

Even so, people from these countries are still able to communicate with each other, but they know each one were born in different countries or parts of a country. Based on that, this project aims to study cultural differences imparted to a language by making a Portuguese Dialect Classifier, which is intended to conclude if a given text in Portuguese language was written by a Portuguese or Brazilian person. This means that it is tried to check if some model can see differences in the usage of words or expressions of a single language and foresee the person nationality and consequently the dialect used.

The motivation of this project idea is it can contribute to build colloquial dictionaries for a dialect or a unique dictionary for the language, with words or expressions tagged by dialect. It could also help make better translations, including some within the language, or translations appropriate to the user nationality. Finally, it could help identify the person's nationality or which part of a country the person comes from.

Data

Getting data was challenging. Creativity in thinking of different ways and sources to accomplish the objective of the project played a big role. It was the longest part of it because of the several ideas that had to be experimented. Because of it, classes were limited to only Portugal and Brazil since it would take a lot of time to collect data from other countries

The original idea was to use text from interviews, news articles, and speech transcripts to build the model. However, this approach presented some issues. Firstly, colloquialisms tend to be used more often in informal settings, while most interviews, news articles, and speeches follow formal language structure. Secondly, it is difficult to find such a corpora where the range of topics is very diverse or very controlled. This is important as a model should learn to distinguish text based on colloquialisms, not the difference in topics of the training data.

Other two ideas were tried and did not work. One was searching for ready datasets online, but no dataset found had the same goal of this project or could meet its needs. This was because topics explored in the Brazilian dataset weren't explored in the Portuguese dataset and vice-versa. And another idea was to get posts from internet blogs since there could be a better topic and diversity control. About 100 posts equally divided in the two classes were manually collected. Most posts were long enough, but the final number of sentences was too small to be able to explore differences in the language. These blog posts ended up not being used in the final dataset.

To overcome these issues, data was collected from Twitter using two libraries: Python Twitter Tools [2] and Python Twitter [3]. They are both wrappers for the official Twitter API [4]. After finding online a list of the most followed users from each country and ensuring each of them were from Brazil or Portugal, their tweets were collected using [2]. Additionally, using [2], real-time tweets were collected by placing filters on the language and geographic tags of the tweets. This also helped ensuring that collected tweets were in Portuguese and originated in either Portugal or Brazil. In the end, around 21,000 tweets, 13,000 from Brazilian users, and 8,000 from Portuguese users were collected. After collection, these were cleaned by removing emojis, hashtags, user mentions and hyperlinks using regex in Python. Cleaned tweets that became empty were removed. Hereby, the final dataset built is composed of two files, one for each country, where each line of a file is a cleaned tweet of a user, and a tweet is formed by one or more sentences in Portuguese language.

Method

The intuition is that people from different parts of a country or the world, even when speaking the same language, will have slightly different vocabularies and use significantly different colloquial phrases in their informal written or spoken communication. In order to capture these differences, the N-gram models was used.

The explanation is that frequent N-grams will capture the frequently used colloquialisms in the data, and those can be used to identify the source/dialect of the tweet. Frequent N-grams for N set to 1, 2, and 3 were extracted from the dataset for this purpose. They constitute the feature space in which the tweets are represented or projected into. And to train these featurized tweets to predict the dialect of the tweet, a logistic regression classifier was used.

Experiment Approach

In order to evaluate the model, the classification accuracy of the logistic regression classifier is recorded. Multiple runs of training and testing are performed and the average accuracy is reported. Tweets not trained on do not contribute to the construction of the feature vector. To filter the training data, minimum frequency/support of N-gram was set to 5. Feature vector contains N-grams of length 1, 2 and 3. The NLTK Library [5] was used to generate the N-grams and for vectorizing the tweets.

Logistic regression classifier was chosen because, by looking at the weights it assigns to each N-gram, one can get some insight into which words are more predominantly used by one group. Ideally, one would see words used by one group getting very positive weights, while the words used by the other group get very negative weights. The scikit-learn library [6] was used for the logistic regression classifier.

The baseline that is used to be compared against the logistic regression is a majority classifier. It classifies every new tweet as the majority class from the training set.

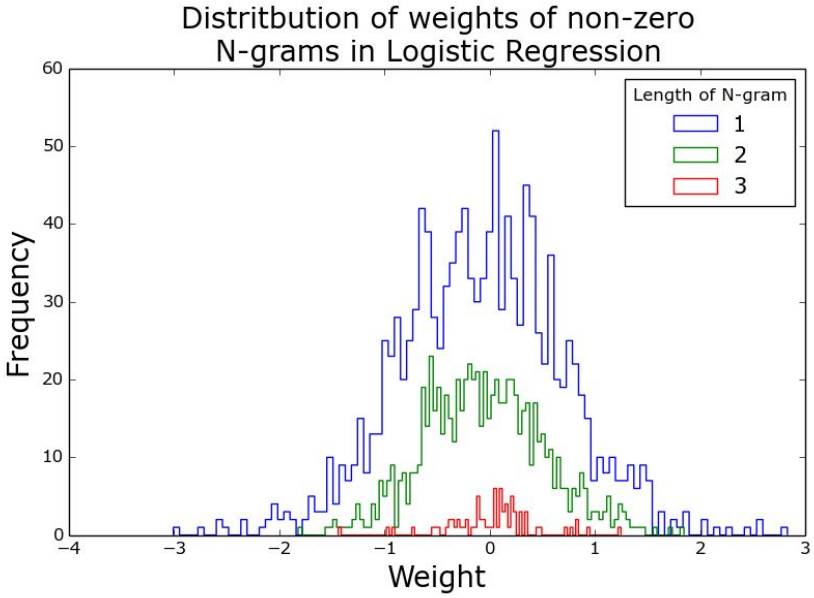
Results

The accuracy of such baseline for this dataset is around 60.5% as about 60% of the tweets are from Brazil. The performance of the logistic regression classifier is shown in the table below:

Training %	# Features	Classification accuracy (%)
10	522	69.59
20	1250	72.88
30	2123	74.56
40	2888	75.78
50	3760	76.70
60	4562	78.03
70	5456	77.77
80	6363	78.92

It can be seen that even when training on only 40% of the dataset the model is able to significantly outperform the baseline.

Looking closer at the weight vector from the classifier, it is seen that about 65% of the weights are 0. For the remaining features, the figure below shows the distribution of weights for N-grams of different lengths. Although the relative frequency of N-grams of different lengths varies, a similar trend in terms of weight distribution within the N-gram category is seen across all categories. Unigrams tend to have more extreme scores, hinting at the fact that they are the most polarizing category terms between the two dialects, followed by bigrams, and trigrams. The frequent bigrams and trigrams should capture common phrases in each dialect, and may show a trend more like the unigrams if the dataset is larger.



Once again, looking at the weight vector and the words associated with them, it is observed that words from different dialects get very different weights. Words predominantly used in Brazil have negative weights and those used in Portugal have positive weights usually:

BR Word	Weight	PT Word	Weight	EN Word
equipe	- 3.01	equipa	2.56	team
vc (você)	-2.31	tu	0.91	you
demais	-1.71	demasiado	1.44	too much
isso	-0.54	isto	2.83	this / it

Conclusion and Future Work

The goal of this project was to distinguish between two dialects of Portuguese using simple techniques in language modeling and text processing. Through the use of the N-gram model and logistic regression, we were able to model the differences in the two dialects studied, Portuguese from Brazil and Portugal, with significant success by achieving an almost 20% absolute gain over the baseline model.

Ultimately, this project serves as an introductory study for Portuguese dialect classification, and it could possibly be enhanced by collecting more data and making better feature selection, since there are so many n-grams with weight 0. In addition, if there are enough experts, this project can be extended to include the remaining 7 countries and specify regions within each.

References:

[1] Community of Portuguese Language Countries: http://www.cplp.org/id-2595.aspx
[2] Python Twitter Tools: https://pypi.python.org/pypi/twitter
[3] Python Twitter: https://python-twitter.readthedocs.io/en/latest/
[4] Twitter API: https://dev.twitter.com/overview/api
[5] NLTK: www.nltk.org/
[6] scikit-learn: scikit-learn.org/