CS 7311 Project Report
Texas State University
Fall 2018

Lia Nogueira deMoura
l_n63
l_n63@txstate.edu

Abstract

This project investigates the use of English loan words in Brazilian Portuguese informal writing using a dataset of ~100,000 tweets from 2006 to 2018. This data sample provides an initial analysis of the use of these words. Logic created for this project can be reused for the analysis of a more comprehensive dataset. The results showed that 13% of the tweets had a least one English word and the total number of distinct words found was ~1,500. It also showed that new words start being used throughout the years, so this cannot be a static search and needs to continue as the time passes. These words need to be taken into account when doing content or sentiment analysis of texts.

1. Introduction

Many English words have been added to Brazil's vocabulary throughout the years. This influence from the English language results in different types of borrowings such as loan words, extensions and loan translations. *[1]* This study is focused on loan words that have been borrowed directly from English and are used in their original English spelling.

The evidence and trend of this influence will be presented in this report by analyzing ~100,000 Brazilian Portuguese Tweets from 2006 to 2018. This report will also show some of the challenges of classifying the words in Portuguese text as English or not. The hope is that this investigation can continuously help us better understand the impact of this influence on the Brazilian Portuguese NLP research.

2. Related Word

Several studies have been conducted on the social and linguistic aspects of the insertion of English words in Brazilian Portuguese. Kennedy *[1]*, for instance, studied examples of loan words, extensions, and loan translations from English by manually gathering data from Brazilian newspapers and magazines, literature and songs, textbooks of Brazilian Portuguese for Americans, conversations and correspondence with Brazilians, etc. Diniz de Figueiredo *[2]* studied the use of English loanwords as slang by manually gathering data from Brazilian websites.

Even though these studies provide great insights, they have limited data and only represent the data as of a specific point in time.

This project intends to automate the search for English words by using Twitter's data. This way it will be possible to continuously analyze the information as time passes.

3. Data

3.1 Twitter

Data between 2006 and 2018 was extracted from Twitter's *fullarchive* and *30Day* APIs.

The APIs requests were done in a way to filter only the appropriate tweets, which were those where the tweet identified language was *Portuguese* and the country in the user profile was *Brazil*. This criteria insured that only Brazilian-Portuguese tweets were selected, excluding Portugal Portuguese tweets and also tweets from Brazilian users writing in English.

The number of tweets found with these criteria was small in the first years of Twitter's existence (2006-2008). The search started to find a greater amount of results from 2009 forward. A total of 104,823 tweets were collected for this project, distributed into almost every year and month since 2006. *(Table 1)*

Since there was a limited number of API requests available, the requests were done in a way so that the available data could be evenly distributed over the months and years. A greater number exists in 2018 because of the extra available requests in the *30Day* API.

The year 2009 was used in the test phase of the code. That is the reason why a larger number of Tweets exist in that year.

Twitter's API returns the result files with all tweets for that particular request in JSON format. Every tweet document contains a number of attributes about the tweet and about the user. For the purpose of this project, the interest lies only on the attributes *tweet text*, *create time* and *user location*. All other attributes were saved in the dataset, but they will not be used in this study. More knowledge can be extracted from these other attributes later on.

Number of Tweets Per Year	
Year	**#Of Tweets**
2006	3
2007	11
2008	1,835
2009	38,074
2010	6,000
2011	6,000
2012	6,000
2013	6,000
2014	6,000
2015	6,000
2016	6,000
2017	6,000
2018	16,900
Total	**104,823**

(Table 1 – Number of tweets in the dataset)

3.2 NLTK corpora

Five word lists from NLTK library were used to help identify the words as English and Portuguese.

The following are the 5 NLTK corpora:

words: The Words Corpus is the /usr/share/dict/words file from Unix, used by some spell checkers. It can be used to find unusual or mis-spelt words in a text corpus.

machado: The Machado corpus includes the complete works of Machado de Assis

morpho: This corpus includes the MAC-MORPHO Brazilian Portuguese POS-tagged news text, with over a million words of journalistic texts extracted from ten sections of the daily newspaper Folha de Sao Paulo, 1994

floresta: This corpus includes 9k sentences, tagged and parsed (Portuguese), available from http://www.linguateca.pt/Floresta/

genesis: This corpus includes misc web sources - 6 texts, 200k words, 6 languages

More information about these corpora is available in the NLTK documentation.

3.3 Other

To complement the previous data from NLTK, another list of words was used from project Unitex-PB. The list was downloaded from http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html

Some additional information was collected on Wikipedia about Brazil's population, cities and states and also an acronym list from sproutsocial.com.

4. Method

4.1 Data cleaning

The first step for analyzing this data was to break down the tweets' text into words. To do that, *word_tokenize* function from NLTK was used. From the 104,823 tweets in the dataset, ~1.8M words were found in this step.

The second step was to remove all symbols that could not be part of a word. The following were the symbols removed: *\@!#$%^&*()_+-={}|[]\:";'<>?,./*. After removing these symbols and excluding any record left with just an empty space, the count of words decreased to ~1.4M.

The third step was to remove any monosyllables. Any records where the word length was 1 was excluded. The word count decreased to ~1.2M after this step, with 149,255 distinct words.

The step after this becomes more challenging as cognates need to be identified to make sure Portuguese words with the same spelling as English words are not included.

4.2 Challenges

Before moving on to the next step and tagging words as English, some challenges in the data had to be taken into account.

There were words in Portuguese that had the exact same spelling in English. Some of those words shared the same meaning, like "perfume", "crime", "cinema", "hospital", etc. and some others did not share same meaning, such as "time", "sons", "age", "grave", etc.

Another similar challenge was with words with close spelling. The word "processo", for example, is a Portuguese word that if spelled incorrectly as "process", becomes an English word. We don't want to tag words as English if that is not really what the user meant.

Another problem was with company names using English words such as "Apple" or "Indeed". These words should not be tagged as English if they are not being used in their original sense of the word. Any other proper name should also not be part of the English words this study is looking for.

4.3 Tagging words

The next step was to tag each of the words as English or not. For this task the library words from NLTK was used. Out of the 149,255 words, 4,431 distinct words were tagged as English. At this stage proper names, misspelled words and Portuguese cognates were still part of the list.

To be able to identify which of those words were also Portuguese, the corpora machado, morpho, floresta and genesis were used. As an additional resource, the list from Unitex-PB project was also used. At the end of this step all 4,431 words were tagged as Y(yes) or N (no) if they were found in each of these corpora.

The hope at this point was that the tags would easily identify which words were English only and not Portuguese as well, but the data showed sparser than anticipated. The same word labeled as Portuguese in one corpus did not receive the same label in a different one.

In an effort to try to identify which Portuguese corpus contained the best performance when tagging words, a manual process of labeling the words as English or Portuguese was done. This process was done by one person only and in a fast paced fashion. Out of the 4,431 words previously tagged as English, only 2,180 were labeled as actual English and 1,366 were tagged as Portuguese. Some of the words were neither identified as English or Portuguese. Based on that manually trained data it was possible to compare how well each of the corpora was performing. (*Table 3)*

A detailed table showing the 4,431 words and their tags at this stage is available in the supplementary material.

Table 2 shows the summary of the word counts throughout the stages of cleaning and tagging the words.

Counts Summary	
# of Tweets	104,823
# of Words in Tweets	1,800,687
# of Words after removing special characters	1,390,445
# of Words after removing words with length 1	1,270,160
# of Distinct of distinct words	149,255
# of Distinct of distinct words Tagged as English using corpus "words"	4,431

(Table 2)

# of words tagged as Portuguese out of the 4,431 words by Corpus		
Corpus	**Count**	**Accuracy compared to manual label**
machado	1,473	75%
mac_morpho	1,972	68%
floresta	806	77%
genesis	975	57%
Unitex-PB	1,441	24%

(Table 3)

4.4 Location

Every tweet JSON file comes with some information about the user, including a free form field indicating the place where the user is from. That field was used to get some insight of the distribution of tweets among the states of Brazil. Since that field is free form, some users just added locations such as "Neverland" or "In my house", instead of an actual location. And even the locations that were real received different formats for different users. For example, "Sao Paulo" or "Sao Paulo, Brasil" or "SP, BR" or "Garulhos/SP" are all the same state. Some cleaning was necessary to get the actual state where the user was from. A list of cities and states of Brazil was used to create the logic to match the states for each Tweet as best as possible. 76,211 out of the 104,823 tweets had a valid location.

There is a geocode location field included in the JSON file, but not all tweets have that field filled. That is why the free form field was chosen instead of the geocode location.

5. Findings

The final number of distinct English words found among all 149,255 distinct words was 1,553. These words were used a total of 19,011 times. Many of the most frequent words are related to social media language and technology. *(Figure 1)*

Out of the 104,823 tweets collected, 13,721 of them had at least one English word.

The number of English words being used across the years did not change much, but the data showed the appearance of new words that had never been used before. For example, there were almost no occurrence of the word "crush" before 2013 but that number started to increase in popularity after that. This is true for many other words.

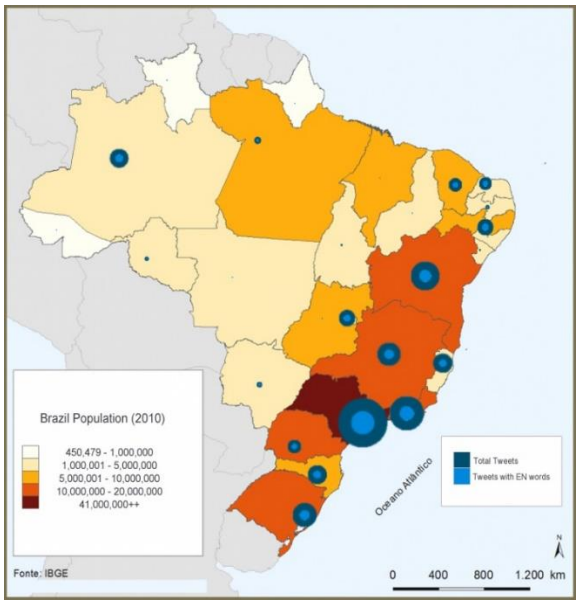
Some words that were expected to appear in this dataset did not show up. To investigate if these words were actually not used, another request to twitter API was done just querying for some of these words.

Thousands of tweets were returned, showing that that current dataset does not represent a big enough sample to be able to find all words being used and the comparison of their frequency.

When plotting the tweet frequency for each state, the number of tweets seems to be proportional to the population of that state. *(Figure 2)*

A detailed table showing the final 1,553 words and their frequency in each year is available in the supplementary material.

(Figure 1 – Word cloud with final list of English words)
*Not all words are included



(Figure 2 – Tweets in each state)

6. Tools and Infrastructure

Because of its simplicity and big range of open source libraries, Python was the language chosen for the execution of this project. Taking this as a starting point, the Natural Language toolkit (NLTK) was chosen as the NLP tool. Python 32bits had to be used since it is more compatible with NLTK.

MongoDB was used to save the data because of the convenience that it provides for manipulating data and the feature of easily making backups and restoring the data.

This project was completely done in a local environment as this was only an initial analysis that can be expanded in the future. Part of the code and datasets are shared in GitHub and can be reused.

Twitter API libraries were used for the twitter requests.

Other tools used for this project were: Jupyter Notebook for coding and troubleshooting; wordclounds.com for word cloud graphs; www.kepler.gl and Photoshop for helping with geocode graphs;

7. Code

Some of the code used for this project is shared on the Texas State github repository. The available code has logic to extract data from Twitter, break tweets into words, tag them using the NLTK corpora and print and export some of the results.

The code includes three classes. Class *TweetExtract* has functionality to request data from Twitter's APIs and save it in MongoDB database. The MongoDB connection string and Tweet Bearer token are required as parameters. Once instantiated the class can be called to do searches on Twitter and accepts the following parameters: ApiName (*30Day* or *fullarchive*), Query, StartDate, EndDate and MaxResults. Class *TagaAndClean* uses the Tweets collection saved on MongoDB to breakdown the words in the tweets text, clean them and tag the words according to the NLTK corpora. The class *Results* has options to print or export some of the results. Also saved in the repository is a main class showing simple examples of how to call these three classes.

The repository also includes some of the source and result datasets used in the project including the complete list of words and their tags, the final list of English words found with their frequency per year and the frequency of tweets per state. It also includes some of the source data such as the Unitex-PB word list.

The repository is available at: https://git.txstate.edu/l-n63/CS7311/tree/master/BR_EN_LoanWords

8. Conclusions

In this study, 104,823 Brazilian Portuguese tweets from 2006 to 2018 were investigated in order to find the frequency of English words being used in Brazilian Portuguese informal writing.

Out of the 104,823 tweets 13,721 of them contained at least one English word which represents around 13% of the total tweets.

Identifying English words in Brazilian Portuguese text has its challenges, since recognizing if a word is truly being used as English or if it is just a cognate is not a simple task.

The final list of words showed that the words being used changed overtime. Some words that had never been used started being used overtime, which means that this should not be a static search since new words can start being used at any moment.

A considerable number of tweets had English words in them and the meaning of those words are important for any content or sentiment analysis. The continuous search for these words should be used in NLP research.

Even though many words were found, the data does not represent a big enough sample to represent all words being used. When searching in Twitter for specific words that did not appear in this dataset, many tweets were found, which shows that a more complete dataset is needed for a more comprehensive result.

9. Next

This experiment was done with a limited amount of Twitter's requests. This study can also be expanded to analyze more data and to continuously extract more information from Twitter. The initial code can be reused.

The manually labeled data can be used to improve the algorithm to classify words. Also, more content analysis is needed to improve the classification logic.

The graph generation can also be automated so that we can continuously get insights about words being used.

Expressions and combined words can also be included in the search so that we have a more comprehensive understanding of the words being used. For example, *"What's Up", "Get it" or "Love You".*

Once this analysis returns solid results without manual interference, some of these results can be used for other NLP research and can eventually be included as part of a NLP library.

The JSON files provide more attributes that could be used to extract more knowledge about the data. For example, user data can be used to get insights about what type of users use what type of words, or what type of users are more likely to use English words.

10. References

[1] Kennedy, James H. "The Influence of English on the Vocabulary of Brazilian Portuguese." Hispania, vol. 54, no. 2, 1971, pp. 327–331. JSTOR, JSTOR, www.jstor.org/stable/337793.

[2] Diniz de Figueiredo, E. (2010). To borrow or not to borrow: The use of English loanwords as slang on websites in Brazilian Portuguese. English Today, 26(4), 5-12. doi:10.1017/S0266078410000301

Other links:

Brazil's population map:
https://atlassocioeconomico.rs.gov.br/populacao-absoluta
(Retrieved December 13, 2018)

Information about NLTK corpora:
http://www.nltk.org/howto/portuguese_en.html
(Retrieved December 13, 2018)

Acronyms list:
https://media.sproutsocial.com/uploads/2015/06/Social-Media-Acronym-Cheatsheet1.png
(Retrieved December 13, 2018)

SUPLEMENTARY MATERIAL

Supplementary material with more details about the data is available at https://git.txstate.edu/l-n63/CS7311/tree/master/BR_EN_LoanWords/Report