Sentiment Classification of Tweets in Spanish

Capstone Project 2 Milestone Report

Springboard Data Science Career Track
June 12^{th}, 2019

Logan Rudd

1 Introduction

Sentiment analysis, also know as opinion mining, uses a mix of Natural Language Processing (NLP), machine learning, and statistics to extract and classify attitudes, opinions, and emotions (known as sentiments) from text to identify the underlying tone of it, whether it be positive, negative, or neutral. From reputation management to customer support the applications of sentiment analysis in business cannot be overlooked. And with today's social media coverage through platforms such as Twitter and Facebook, there are a plethora of data available to mine. These data can be used to analyze a person's feelings, thoughts, or judgements about a particular product, event, or company, and can be used to help inform business decisions, or product development.

Many machine learning models have already been analyzed and tuned for sentiment analysis of Tweets in English, but far fewer for sentiment analysis for Tweets in Spanish. This particular project focuses on classifying positive and negative Tweets in Spanish as well as highlights common NLP processes to consider specific to NLP in the Spanish language and analyzing Tweets.

1.1 About the Dataset

This project will utilize two datasets found on GitHub, and scrap a third dataset from a Wikcionario page to improve the lemmatization of the Spanish vocab. The first dataset is the TwitterSentimentDataset which contains approximately 250k tweets in Spanish, of which, approximately 70k are unlabeled, 55k are labeled with happy emojis, and 122k are labeled with sad emojis.[1] A second dataset provides us with a list of 497,560 Spanish lemmatization pairs to reduce Spanish words to their root.[2] The third dataset will be scrapped from the Wikcionario page 'Frecuentes-(1-1000)-Subttulos de peculas', which provides us a list of the 1000 most frequent used words in Spanish from subtitles of movies and television shows containing 22 million words.[3]

2 Data Preprocessing and Initial Findings

Before performing any analysis, the TwitterSentimentDataset and lemmatization pairs list were first downloaded from their respective GitHub repositories and read into a Pandas dataframe, while the 500 most frequent Spanish words list was scrapped from a Wikcionario page and saved as a text file. The labeled and unlabeled tweets from the TwitterSentimentDataset were then standardized by removing URL's, hashtags, mentions, and emojis, as they don't provide any useful information about the sentiments of words and would likely serve as noise for the model.

After standardizing both labeled and unlabeled Tweets, we first created our dataset for training and testing the sentiment classifiers, and then we created the dataset to be used with Word2Vec to create word embeddings for the CNN+LSTM sentiment classifier. The was process was nearly identical for processing both datasets. First the lengths of Tweets in number of words were were visualized and outliers were explored showing the existence of a total of 69 empty Tweets as well as thousands of duplicates. Empty Tweets were removed and duplicates were left until after lemmatization as this process would likely produce more duplicates.

The remaining Tweets were then lemmatized, removing inflectional endings and reducing the word to its base or dictionary form, known as the lemma. To lemmatize the Tweets, first a dictionary was created from the dataset list of lemmatization pairs to replace words with their lemma. A list of the top 500 most common words in Spanish was then used in conjunction with the dictionary of lemmatization pairs when lemmatizing the Tweets to prevent changing common words that have more than one different meaning to the lesser used word and therefore distorting the context and meaning of that word. For example the word 'como' is often used in Spanish to mean 'like/as' but can also be the first person present tenser of the verb 'comer' which means to eat. Lemmatization would normally convert 'como' to 'comer' when in the vast majority of cases it should be left as is to prevent the incorrect association of the the verb 'comer'. After lemmatizing the Tweets, a total of 24,274 duplicate Tweets were removed.

Appendix, Figure 1 and Figure 2 show histograms of the lengths of labeled and unlabeled Tweets respectively (in number of words) after all preprocessing was completed. The resulting datasets were then written to a text file to be used for feature selection and predictive modeling in the notebook '2.0-lwr-predictive-modeling'. For the dataset

created for Word2Vec algorithm used to extract the feature vectors for sentiment classification in the Convoluted Neural Network (CNN) + Long Short Term Memory (LSTM) network, all of the unlabeled Tweets were concatenated with only the training set of the labeled Tweets before being written to a text file. In that way, no information is leaked from the hold out/test set into the CNN+LSTM model.

References

[1] GitHub.
TwitterSentimentDataset, 2015
`https://github.com/garnachod/TwitterSentimentDataset`

[2] GitHub.
Lemmatization Lists, 2018
`https://github.com/michmech/lemmatization-lists`

[3] Wikcionario.
Frecuentes-(1-1000)-Subttulos de pelculas, 2019
`https://es.wiktionary.org/wiki/Wikcionario:Frecuentes-(1-1000)-Subttulos_de_pelculas`

Appendix

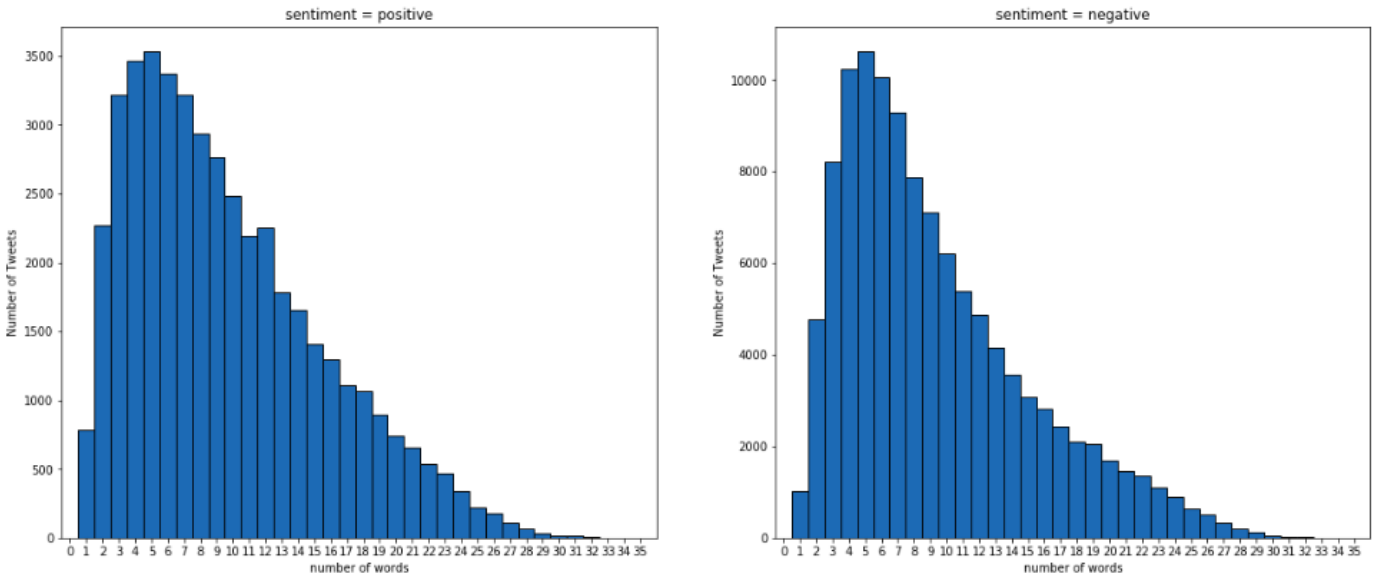

Figure 1: Histograms of lengths (in # of words) of labeled Tweets by class

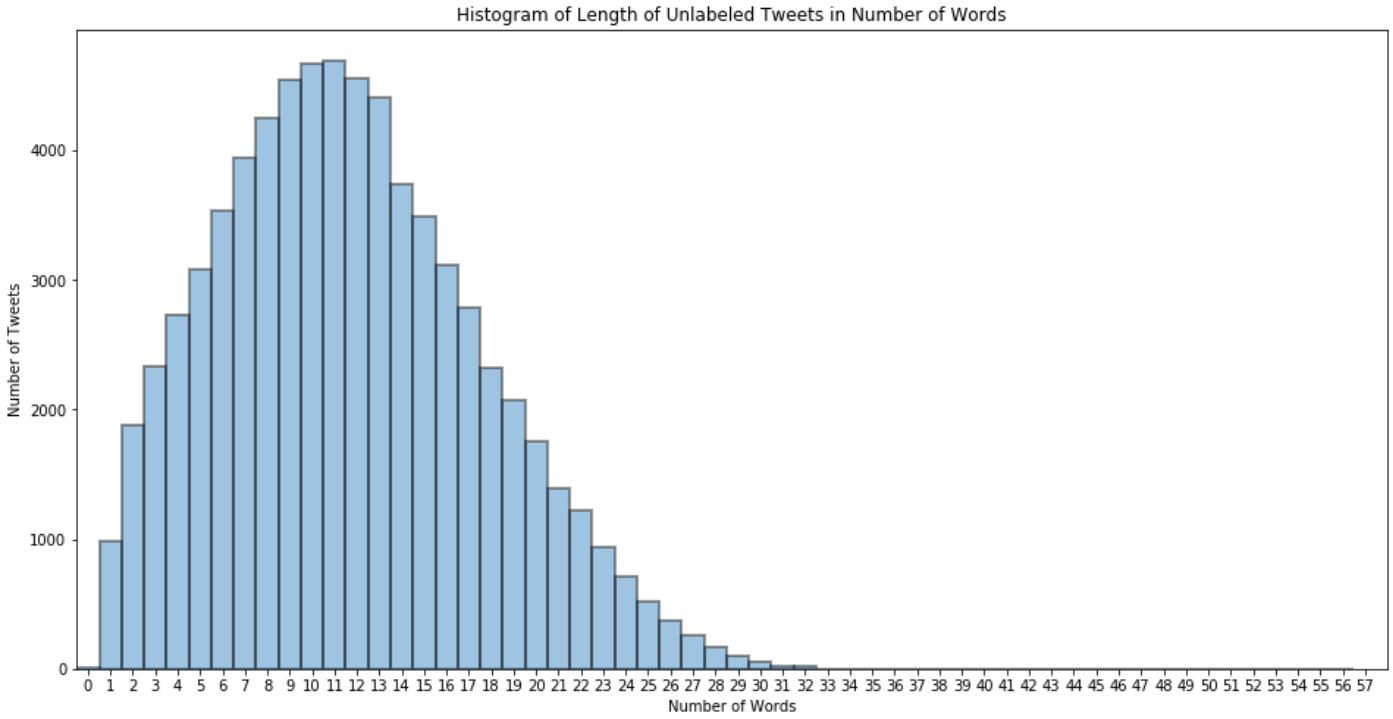

Figure 2: Histogram of lengths (in # of words) of unlabeled Tweets