# Sentiment Classification of Tweets in Spanish

## Capstone Project 2: Proposal

Springboard Data Science Career Track
May 23$^{rd}$, 2019

**Logan Rudd**

# 1 Background and Justification

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 will focus on classifying positive and negative Tweets in Spanish as well as highlight common NLP processes to consider specific to NLP in the Spanish language and analyzing Tweets.

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 pelculas', 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 Project Plan

This project will focus on comparing three different algorithms using two methods. The first method will use Term Frequency-Inverse Document Frequency (TF-IDF) of uni-grams and bi-grams in the tweets to classify sentiments using the two algorithms Multinomial Naive Bayes and Logistic Regression, and the second method will use word embeddings to predict sentiment class using a combination of a Convolutional Neural Network (CNN) and Bidirectional Long Short Term Memory (LSTM) Network.

The project will begin by downloading and importing the tweets into a Jupyter notebook, and then cleaning and preparing them by making any necessary transformations such as removing links, emojis, hashtags, and mentions. Missing values will be handled accordingly, and the data will then be processed even further, using the lemmatization list to reduce conjugated verbs into their infinitive form.

A list of the top 500 most common words in Spanish will be used in conjunction with the lemmatization list 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 preprocessing the Tweets, various features (such as frequency and distribution of uni-grams and bi-grams, stopwords, length of tweets, vocab size, etc.) will explored to reveal insights about how to best select features for predicting positive and negative sentiments. For the CNN-LSTM algorithm a Word2Vec model will first be produced from the training set of labeled tweets, as well as the set of unlabeled tweets, and the word vectors for these words will be used as features for the CNN-LSTM algorithm.

Since negative sentiments (about a business product or service) are usually more helpful to improve the product or service that a business offers and since the dataset of labeled Tweets is imbalanced, with approximately 72% belonging to the negative class, the models will be tuned and cross-validated using the metric f1-score (on the negative class), as this metric has been shown to be a good un-biased metric for imbalanced datasets.[4,5] Final model selection will be made based off a combination of the highest f1-score on negative sentiments as well as the highest f1-score on positive Tweets. This project will conclude with a Jupyter notebook in GitHub, milestone report, final report, and slide deck.

## 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

[4] Hewlett-Packard Development Company, Technical Reports. *Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement: United States, 2009* (Report No. HPL-2009-359). Retrieved from https://www.hpl.hp.com/techreports/2009/HPL-2009-359.pdf

[5] Swalin, Alvira. "Choosing the Right Metric for Evaluating Machine Learning ModelsPart 2" Medium, 2 May 2018, https://medium.com/usf-msds/choosing-the-right-metric-for-evaluating-machine-learning-models-part-2-86d5649a5428