Executive Summary

Erdos Institute, Spring 2022

Jack Carlisle
Mohammed Karaki
Cristian Rodriguez

Sentiment analysis refers to the extraction of sentiment (positive or negative) from a given collection of text files (tweets, comments, reviews, etc.). Machine learning techniques can be applied to analyze the sentiment present in large data sets. In this project, we have applied Machine Learning techniques to perform a sentiment analysis on a dataset of over 180,000 tweets related to the 2019 Australian election.

Our process is as follows. First, we trained a variety of ML models on a separate dataset of 0.2 million tweets. Each of these tweets was pre-assigned a value of positive or negative, so we can test the accuracy of our models on this data. After training five models (SVC, Decision Tree, Random Forest, Logistic Regression and Naive Bayes Classifier), and comparing their performance, we found Linear SVC to be the most accurate. More specifically, we chose “Linear Support Vector Machine Classifier using 2-grams without inverse document frequency reweighting”. This model achieved an 80.75% accuracy on the training dataset.

After having selected our model, we implemented our model on our dataset of tweets about the 2019 Australian election. We found that the distribution of sentiment was 61.3% positive, and 38.7% negative. Moreover, we found that the most common phrases present in positively valenced tweets included “win“, “hope”, and “love”, while the most common phrases present in negatively valenced tweets included “racist“, “lost”, and “lie“. This information, as well as its temporally classified version, could be used by political parties or activists in order to influence the opinions and stances of Australian voters.

The techniques we employed to analyze the 2019 Australian election can be leveraged in a variety of ways. For instance, we employed our ML model to predict the party affiliation of twitter users whose tweets appeared in our dataset, by measuring the shift of their sentiment just after the election results became available. Since the party affiliation of twitter users is not known for most of our sample, we were limited in measuring the efficacy of our predictive model. However, by testing our predictive model on some (10) cases for which we provide the labeling ourselves, we find that our model is rather effective (80%) at predicting a twitter user’s party affiliation based on the sentiment analysis of their tweets.