Executive Summary

Erdos Institute, Spring 2022

Jack Carlisle
Mohammed Karaki
Cristian Rodriguez

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

Our process is as follows. First, we prepare our data for training. This involves “cleaning” each tweet (removing special characters, making lower case, stemming, and removing stopwords). What is left is the essential data of the tweet that our model will be able to analyze effectively. After this, we vectorize our data, so that each cleaned tweet is converted into a vector of length 5,000. These vectors will serve as the input for our models.

Having prepared our data, we construct our various models. These models include a Naive Bayesian classifier, a Decision Tree Classifier, a Random Forest Classifier, a Logistic Regression Classifier, and finally a Support Vector Classifier. We train each model on a subset of 10,000 of our labeled tweets, and compare their performance on a testing set. Moreover, on several of the models, we perform a hyperparameter analysis to improve the performance of our model.

In the end, we find that the Support Vector Classifier is most effective, achieving an accuracy of about 80% on our testing set. While this model proved to be most effective, it is also the most computationally intensive to train, and so our other models provide effective, yet computationally feasible alternatives to the SVC model.

The techniques we employed in our sentiment analysis could be leveraged in a variety of ways. For instance, one could predict the party affiliation of a twitter user by analyzing the shift in sentiment before and after the results of the election are posted. Moreover, the sentiment of twitter users’ tweets could be used to inform the nature of political ads shown to such users.