TWITTER SENTIMENT ANALYSIS OF 2012 US ELECTIONS

Soham Pradhan & Kazi Shahrukh Omar
CS-583 Data Mining and Text Mining, Fall 2021
Department of Computer Science
University of Illinois at Chicago

ABSTRACT

Twitter (formerly known as Twttr) is an American microblogging service and social networking site where users post and share messages known as "tweets". Registered users can post, like, and rewrite tweets, but unregistered users can only read those that are publicly available. Tweets are limited to 280 characters. In this project, we have carried out sentimental analysis for political tweets during 2012 US presidential election between Barack Obama and Mitt Romney. Dataset consisted of columns of tweets and their classification. The tweets are classified as positive, negative or mixed. The first step of the process is cleaning the dataset. After this step, we have created various classifiers in order to classify unlabeled data. The performances of those classifiers are discussed in this report. The performance of the model is determined by various parameters such as precision, recall, f1-score and accuracy.

INTRODUCTION

Twitter (formerly known as Twttr) is an American microblogging service and social networking site which generates more than 500 million tweets per day. This makes twitter, an ideal platform to perform sentiment analysis. Our dataset consisted of twitter data about 2012 US presidential election. The positive opinion is given class of 1, negative opinion is given class of -1, neutral opinion is given class of 0 and mixed opinion is given class of 2. Since tweets on twitter tend to have jargons, spelling mistakes, we first preprocess the dataset in order to make it cleaner. Preprocessing helps us improve our evaluation parameters.

DATA PRE-PROCESSING TECHNIQUES USED

1. CONVERT TO LOWERCASE: All the characters in tweets are converted to lower case, so that same words with different uppercase and lowercase characters are treated as same. Ex: Angry, angry and angRy should be treated the same way.
2. REMOVE URL: URLs do not have any significance on classification of tweets so they are removed.
3. REMOVE USERNAME: While username tends to indicate to us about the person who might have tweeted, in this case there is no contribution done towards influencing the sentiment of the tweet, so we have removed username.
4. REMOVE HTML TAGS: HTML Tags are removed from the tweets because they too do not add any value to the tweet.
5. REMOVE STANDALONE NUMBERS: Since standalone numbers too are not useful in expressing a sentiment, we remove them from the tweets.
6. TOKENIZATION: Tokenization is the process of separating sentences into

words, characters or sub word which form the smaller units and are called as tokens. Space is the most common delimiter.

7. REMOVE STOPWORDS: Stop words are commonly used words in language. We need to eliminate these words since they do not add sentiment to our tweets. We use nltk library to remove stop words of the English language from our tweets.

EXPLORATORY DATA ANALYSIS AFTER PRE-PROCESSING

Exploratory data analysis is done before classification in order to understand the data first and try to gather as many insights from it as possible. Below are some of the charts which help us know the data better. The data results are segregated into two classes, Obama and Romney.

As we can see in Figure 1 and Figure 5, the number of positive and negative tweets in both Obama and Romney datasets are well balanced.

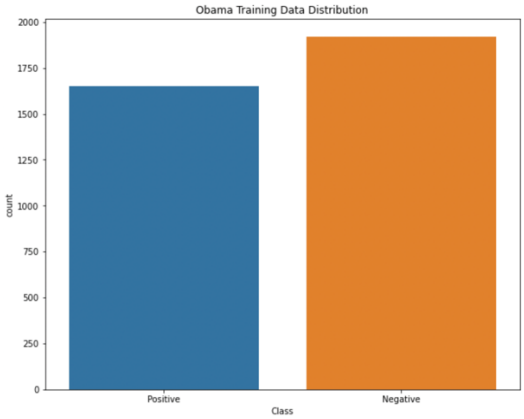

Figure 1: Obama Training Data Distribution

From Figure 2 and Figure 6, we can see that the number of words in positive tweets are less compared to that of negative tweets.

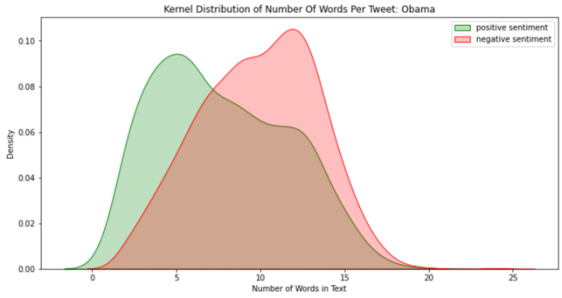

Figure 2: Kernel Distribution of Number of words per tweet

From Figure 3, 4, 7 and 8, the names of two politicians are the most common words in negative as well as positive sentiments.

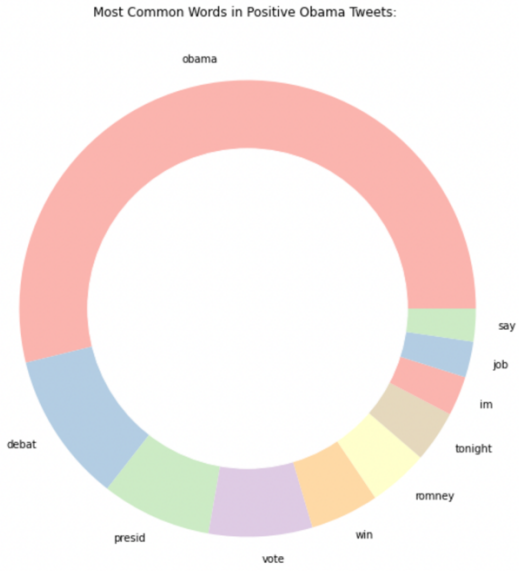

Figure 3: Most common words in positive Obama tweet

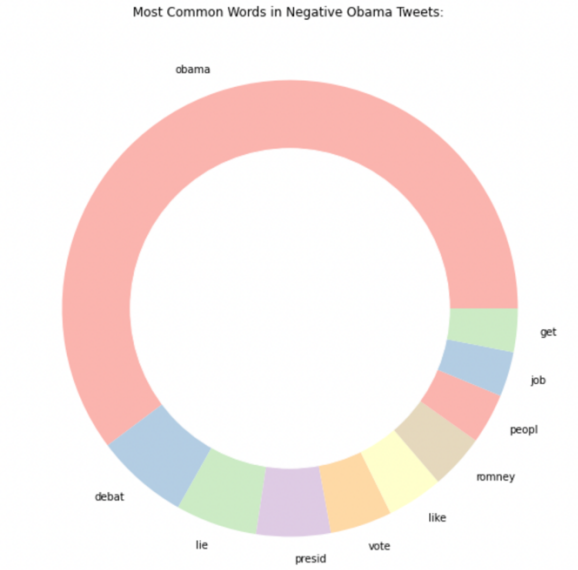

Figure 4: Most common words in Negative Obama Tweets

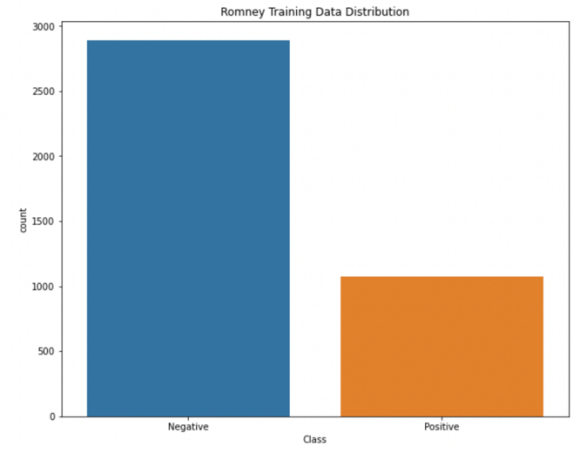

Figure 5: Romney Training Data Distribution

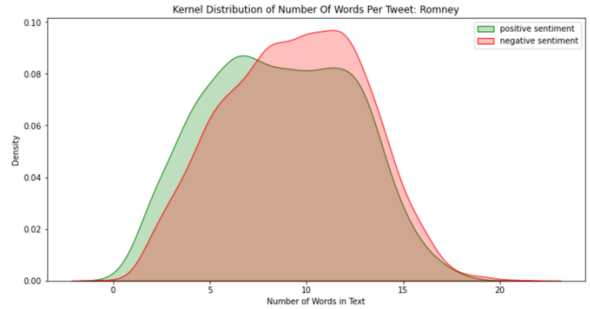

Figure 6: Kernel Distribution of Number of words per tweet Romney

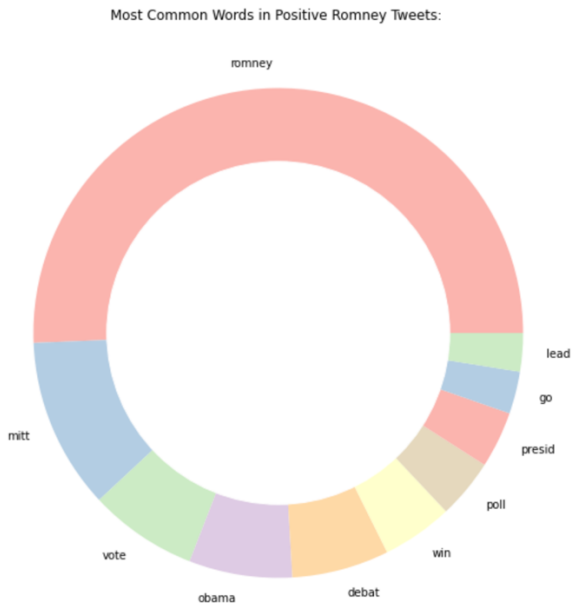

Figure 7: Most common words in Positive Romney tweets

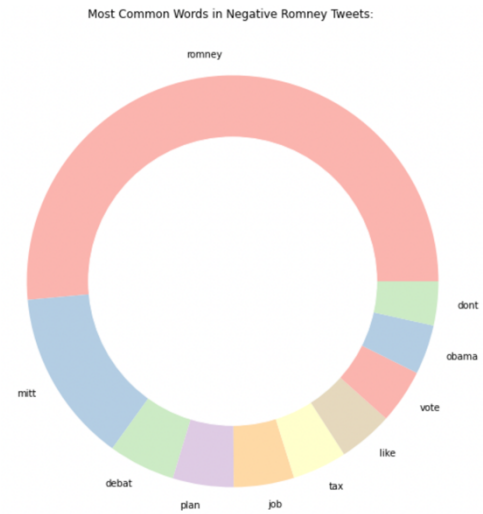

Figure 8: Most common words in Negative Romney tweets

CLASSIFIERS EMPLOYED

BernoulliNB: BernoulliNB implements naïve bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions. This class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a BernoulliNB instance may binarize its input. BernoulliNB gives one

of the highest accuracies in our classification.

SVM: Support vector machine is another popular classification which was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. It belongs to supervised learning method. SVM can solve linear and non linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. SVM along with Naïve bayes and LSTM had the best accuracies in our project.

Decision tree classifier: Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions. Each leaf in the tree represents a class.

Random forest classifier: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

SGD Classifier: This estimator implements regularized linear models with Stochastic Gradient Descent. Stochastic gradient descent considers only 1 random point while changing weights unlike gradient descent which considers the whole training data. For best results using the default learning rate schedule, the data should have zero mean and unit variance.

MLP Classifier: Multilayer perceptron is a class of artificial neural network. MLP consists of at least 3 layers of nodes: an input layer, a hidden layer and an output layer. Except for input nodes, each node is a neuron that uses a nonlinear activation function.

LSTM: Long short-term memory is a RNN architecture used in deep learning. A common LSTM unit consists of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and three gates regulate the flow of information into and out of cell. LSTMs were developed to deal with vanishing gradient problem that can be encountered when training traditional RNNs. LSTM model gave us one of the highest accuracies.

EVALUATION

The evaluation parameters of the classifiers are accuracy, precision, recall and f1 score.

Figure 9 shows F1 scores for Obama. SVM gave us the best result for this evaluation.

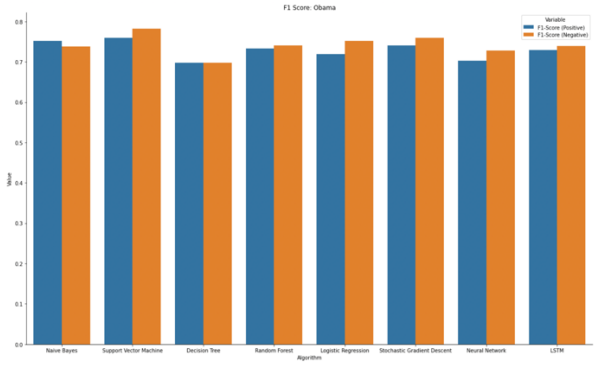

Figure 9: Obama F1 score

Figure 10 shows precision for Obama. SVM gave is the best result for this evaluation.

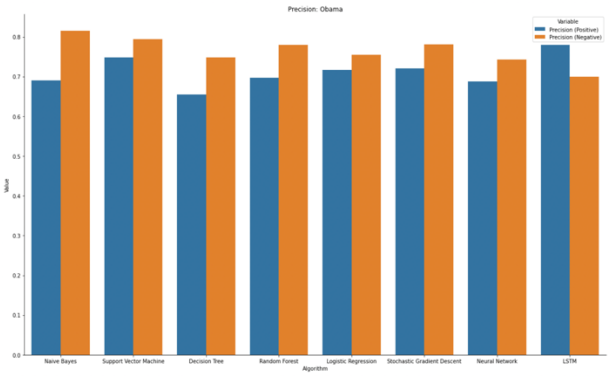

Figure 10: Obama Precision score

Figure 11 shows recall for Obama tweets, SVM gave us best results for evaluation.

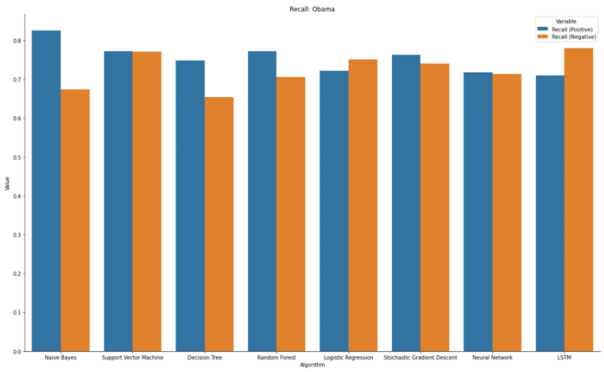

Figure 11: Obama recall score

Figure 12 shows Accuracy for Obama, SVM gave us the best results for evaluation.

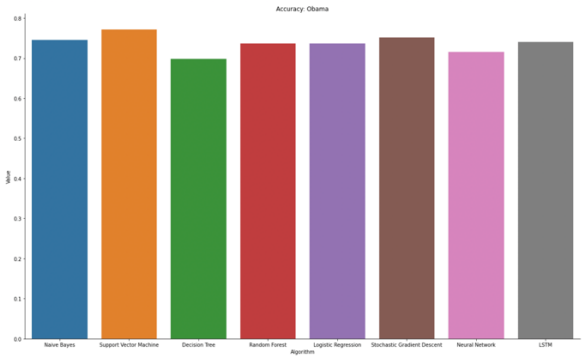

Figure 12: Obama Accuracy

As we can see in Figure 13, For the F1 scores of Romney, LSTM gave us the best result for the evaluation.

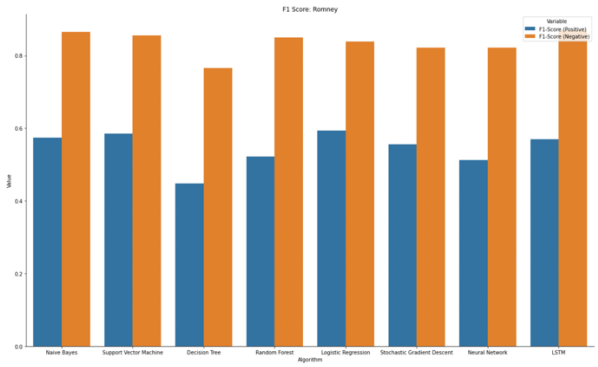

Figure 13: Romney F1 score

As we can see in Figure 14, For precision score of Romney, LSTM gave us the best result for evaluation.

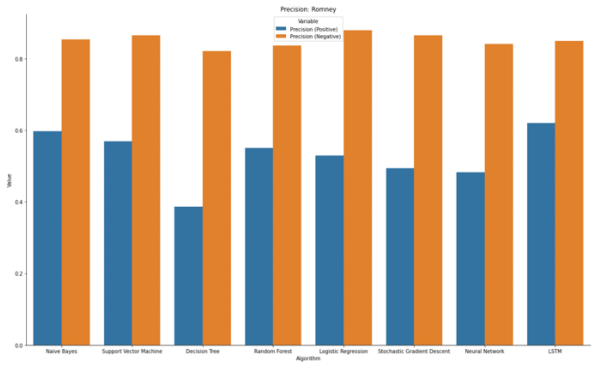

Figure 14: Romney Precision

As we can see in Figure 15, For recall score of Romney, SVM, Naïve Bayes and Logistic Regression gave us the better scores.

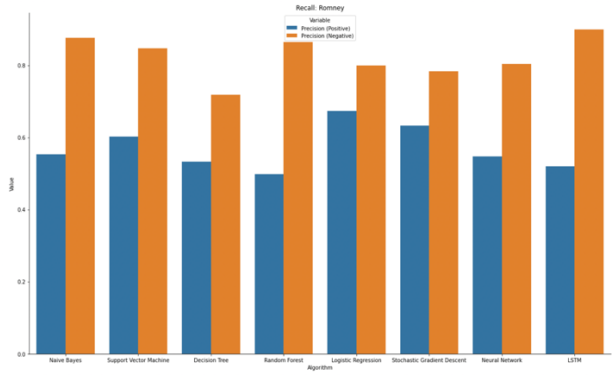

Figure 15: Romney Recall Score

As we can see in Figure 16, For accuracy score of Romney, LSTM gave us the best results for evaluation.

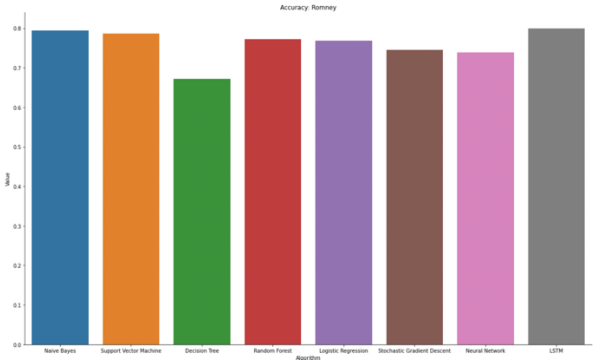

Figure 16: Romney Accuracy

CONCLUSION

We experimented with 8 classifiers as shown above. We used 10-fold cross validation for training the dataset and trained the classifiers. SVM classifier worked best for the Obama tweets achieving an accuracy of 77.20 % and LSTM classifier worked best for the Romney tweets achieving an accuracy of 79.97%.

REFERENCES:

1. Scikit-learn: https://scikit-learn.org/stable/
2. Nltk: https://www.nltk.org/
3. Towards data science: https://towardsdatascience.com
4. Wikipedia: https://en.wikipedia.org/wiki/Wikipedia