Challenge 1

CY6740 – Machine Learning in CyberSecurity

Name: Tejas Krishna Reddy

NUID: 001423166

**Solution:**

**Problem: Twitter US Airline Sentiment Prediction using Machine Learning**

1. Analyzing public sentiment has a vital importance in forecasting the future. Knowing the sentiment would help us understand the future financial stability of that product, it’s sales and its recommendations. Especially in airline industry public sentiment helps us understand the reliability of the airline. Generally, how much delay could we expect from the airline. How good or bad is the comfort/food/luxuries in long flights from this airline etc.

Twitter has been a great social platform in the last decade where people from time to time tweet their experience about products/ or in this case about airlines. Hence, this would be a great platform for us to use these reviews to predict the public sentiments for the airlines.

This would be a great platform where machine learning could show its capabilities. We can use the features given to predict the sentiment of the airline to be either positive or negative. Making it a typical binary classification problem.

1. This dataset from Kaggle has refined features and have tweets about airlines in the year 2015. With this sample we could train a model to predict sentiments for any tweet in the future. Below is the link to the dataset:

[https://www.kaggle.com/crowdflower/twitter-airline-sentiment](https://www.kaggle.com/crowdflower/twitter-airline-sentiment%20)

1. This dataset contains **14641 instances**. It has 16 columns. Among the 16 columns, some features are not helpful in predictions such as tweet\_id, name of the person who tweeted etc. Hence, after a careful observation, **8 features** standout that would help us gauge the **predictor ‘airline\_sentiment’** from the tweet.

Among the 8 features there are **3 numerical features**:

* Positive Confidence: Manually generated numerical (decimal) feature that has an explanation in the website to how it was generated. This feature helps us understand how positive the positive tweets are.
* Negative Confidence: Similar to above for negative tweets.
* Reweet counts: Retweet count helps us understand how many of the other twitter members also feel the same way. This count helps us understand the importance/seriousness of the tweet.

The other **5 features are categorical**: Airline name, text of the tweet (Actual Tweet text), place where it was tweeted, month and timezone where the tweet was placed. These categorical features need to be converted into numerical using one-hot encoding technique to be able to process in machine learning algorithms.   
From the actualText column, we could extract the most repeating important words (after removing the stopwords) and one hot encode that column with important words.

1. Once the categorical variables are converted to numerical, then we could apply machine learning algorithms to the datasets to classify them as positive or negative tweets. After one-hot encoding the number of features we would have would be a lot more and hence, linear classifiers might not perform the best in this regard. For binary classification problems with multiple features, **Logistic Regression** could be a better classifier in classifying the tweets. It’s non-linear fitting would be very intuitive and could reach the global minimum in the cost function optimization. Also, bagging and boosting classifiers would yield better accuracies in this regard due to many one hot encoded variable’s, it would be better if the algorithm has the capacity to choose different importance scores for its features and choose them wisely. Hence, **ensemble bagging algorithms** such as **RandomForestClassifier** and boosting algorithms such as **GradientBoostingClassifier**, **Adaboost** Classifier and **XgBoost** Classifier should yield good results. Apart from this, **multi-layer neural networks** should also do good Inorder to choose better features among many one hot encoded features by training the data through multiple layers and train the weights appropriately.