**Objective:**

Comparing US based airlines performance using twitter sentimental analysis. The sentiment analysis will be performed on each of the airlines separately and also on the entire dataset.

**Project Problem:**

Analyzing the US based airlines performances by analyzing the tweets of the airlines. Identifying the sentiments of tweets and classifying them as neutral, negative and positive sentiment for each airline. Then tokenizing the tweets as unigrams, bigrams and trigrams and finding their significance on the sentiments.

**Dataset:**

Data is obtained from the airline sentiment dataset from Kaggle <https://www.kaggle.com/crowdflower/twitter-airline-sentiment>

**The process followed for Data Cleaning**

1. Load data to R studio
2. Replace twitter handle with blank and filter the dataset by airlines

The tweets contained the airlines twitter handle. We must first remove the twitter handle as they should not be used in the text analysis.

1. Build a corpus, and specify the source to be character vectors

Here we convert the text into a word corpus using the function Vector Source. A word corpus enables us to eliminate common words using the text mining package tm. Removing the corpus specific stopwords

1. Convert the corpus to lowercase
2. Remove punctuations from corpus
3. Remove URLs
4. Remove anything expect English and Space
5. Remove Stopwords
6. Remove extra whitespace 10. Make a copy of the corpus 11. Converting corpus to dataframe

**Term Document Matrix**

1. Create a term document matrix

We convert the word corpus into a document matrix. The Document matrix can be analyzed to examine most frequently occurring words

**Word Frequencies**

We find the most frequent words and we create a Word Cloud of tweets. We are limiting the maximum words to 100 and plotting the top 10 frequent words.

1. Find Frequent Terms
2. Creating a word cloud of frequent term
3. Plotting the top 10 frequent words
4. Plot the frequency of the words on linear / log scale

Plotting the frequency of top words in the logarithmic scale.

1. Plotting Bigrams on words frequency

we will construct bigrams and plot the top 15 bigrams on a logarithmic scale. Bigrams are two-word phrases

1. Plotting trigram on word frequency

We will construct trigrams and plot the top 15 trigrams on a logarithmic scale. Trigrams are three-word phrases

**Sentiments**

1. Retrieve data for the chosen airlines
2. Perform sentiment analysis on the tweets of the chosen airlines

**Document Term Matrix**

1. Creating a Vcorpus
2. Creating a Unigram, Bigram and Trigram DTM from Vcorpus

The unigram/bigram/trigram are single/double/triple word phases from the document term matrix is created and the sparse terms are removed. The tweet ID, airlines and the tweets are added to the dataframe created. The terms are placed across the columns and their occurrence across each tweet are indicated either 0 or 1.

**Preparing the model**

1. Joining the unigram, bigram, trigram dataframe with sentiments dataframe.
2. Mapping the tweet reason to check its significance on the sentiment
3. Mapping the tweet location to check its significance on the sentiment

**Unigram/Bigram/Trigram Modeling**

1. Creating a dataframe with tweet reason, timezone, sentiment frequency, sentiments and unigrams/bigrams/trigrams.
2. Performing logistic regression for each of unigram/bigram/trigram with positive/negative/neutral sentiments with the variable tweetreason.

**Splitting the dataset**

1. Splitting the dataset into training and testing dataset in 70:30 ratio
2. Applying the balancing algorithm for balancing the unbalanced dataset.
3. Mapping both the test and training dataset

**Cross Validation and Bootstrappping**

1. Preforming CV and bootsrapping for logistic regression for the variable tweetreason for all the nine models
2. Plotting the ROC curve for logistic regression
3. Performing CV for Naive Bayes for the variable tweet reason
4. Performing CV for Naive Bayes for the variable tweet reason and sentiment frequency
5. Plotting the ROC curve for Naive Bayes
6. Performing CV for Naive Bayes for the entire training dataset excluding time zone as it is insignificant
7. Plotting ROC Curves and Precision Recall Curves for Entire training dataset
8. Performing CV for Naive Bayes for the entire testing dataset
9. Plotting ROC Curves and Precision Recall Curves for Entire testing dataset

**Observation**

We have found that the unigram/bigram/trigram positive performed well for all the airlines with ROC value greater than 0.5 on the training dataset.

For the test dataset, unigram sentiments were not significant while bigram positive performed well while bigram/trigram negative performed poorly. Those model that were not fitting were excluded from the observation and has been commented out in the appendix.

**Limitation**

Even though Naive Bayes and Logistics Regression has been applied for looking at the significance of unigram/bigram/trigram on the sentiments, multinomial regression could have provided better insights on the significance of the variables. Conversion all words into lowercase might have diminished the significance of certain variables like USA etc.