Analysis

Gathering Data

This step is important because the quality and quantity of data we have gathered will directly determine how good the predictive model can be. In this case the data we collect will be based on the airline sentiment analysis and the relationship between customer sentiments and customer satisfaction, customer loyalty and business profitability.

Data Preparation | Exploratory Data Analysis | Visualization

At first we will put all our data together and randomize the order. We are going to make a determination of predicting Airline Tweet Sentiment Forecasting independent of the name of the user who tweeted.

```
Number or Rows and Columns are (14640, 15)
```

After checking number of null values in all the columns. It was found that the number of redundant columns which have the highest number of null values are tweet_coord, airline_sentiment_gold, and negativereason_gold.

Categories for airline_sentiment are negative, neutral and positive.

Getting the number of Tweets for each airline, United airline has the highest number of tweets i.e 3822 tweets and Virgin America has the lowest i.e. 504 tweets, And all other airlines have tweets between 2200 to 3000.

Looking at the barplots for the mood count of each airline, US Airways, United and American airlines comparatively get negative reactions and tweets for Virgin America are balanced.

Moving forward for the words used in Positive and Negative Tweets, we need to find out most frequent words in negative tweets and positive tweets.

By creating a wordcloud for negative and positive tweets, we would be able to differentiate which negative / positive sentimental word has been used most frequently. Bigger the word, more the word has been used in the tweets.

Most Frequent words for negative sentiments: flight, time, cancelled, airport, help, Plane, flightled, customer, service, still, amp, delayed, now, ridiculous etc. These keywords might help airlines to track the problems faced by the users during their travel experience for the particular airline and helpful to improve the service as per the user's comfort.

Most Frequent words for positive sentiments: flight, love, great, guy, good, time, thank, awesome, appreciate, amazing, service, got, crew, customer service, seat, staff, always, helpful, wait, really, etc. These keywords will help airlines to maintain the service or will be helpful for future decisions to provide the discount offers to the users.

What are the reasons for negative sentimental tweets for each airline ?

Total 10 negative reasons : Bad Flight, Can't Tell, Late Flight, Customer Service Issue, Flight Booking Problems, Lost Luggage, Flight Attendant Complaints, Cancelled Flight, Damaged Luggage, longlines

Customer Service Issue is the main negative reason for US Airways, United Airlines, American Airlines ,Southwest, Virgin America.

Late Flight is the main negative reason for Delta

Interestingly, Virgin America has the least count of negative reasons (all less than 60)

Contrastingly to Virgin America, airlines like US Airways, United Airlines, American Airlines have more than 500 negative reasons (Late flight, Customer Service Issue)

Is there a relationship between negative sentiments and date ?

The dataframe is from 2015-02-17 to 2015-02-24. Visualizing if the date has any effect on the sentiments of the negative tweets.

After Plotting this and get better visualization for negative tweets. American Airlines has a sudden upsurge in negative sentimental tweets on 2015-02-23, which reduced to half the very next day 2015-02-24. (I hope American is doing better these days and resolved their Customer Service Issue as we saw before)

Virgin America has the least number of negative tweets throughout the weekly data that we have. It should be noted that the total number of tweets for Virgin America (504) was also significantly less as compared to the rest airlines, and hence the least negative tweets.

The negative tweets for all the rest airlines is slightly skewed towards the end of the week

Selecting the Model

Predicting the tweet sentiments with tweet text data with

Decision Tree Classifier

Random Forest Classifier

Calculating accuracies, plotting the confusion matrix and comparing the models.

We have applied the classification algorithms like Decision Tree Classifier and Random Forest Classifier. The data is split in the standard 80:20 ratio.

We have plotted the **confusion matrix** for predicted sentiments and actual sentiments (negative, neutral and positive) for both the classifiers

Random Forest Classifier gives us the best accuracy score (0.77), precision scores according to the classification report.

The confusion matrix shows the TP,TN,FP,FN for all the 3 sentiments (negative, neutral and positive), Here also **Random Forest Classifier** gives **better** results than the **Decision Tree Classifier**.

Logistic Regression

But we have got the **winner**, predicting sentiments from tweet text data, **Logistic Regression** with the Test Accuracy of 0.8042.

In case of negative sentiments, out of 1456 negative sentiments 1263 were predicted correctly.

In case of neutral sentiments, out of 420 neutral sentiments 266 were predicted correctly.

In case of positive sentiments, out of 320 positive sentiments 237 were predicted correctly.

```
calculating results...
Logisitc regression Train accuracy is: 0.9535
Logisitc regression Test accuracy is: 0.8042
```