# Twitter Sentiment Analysis Milestone Report

**By William Dew**

## Introduction

Air travel is a huge business in the US. Since Air travel is a service industry Airlines will want to know how they can improve or what they need to do better from their customers. People are always talking about their air travel stories to each other being good or bad. With twitter people can and will tag the airline in their response to their travel.  From twitter data the airlines can see what features are causing positive or negative tweets. And from the data decide what to do as a company.

This project will be focusing on building a model that will analyze the text of a tweet.  There are three sentiment classes that a tweet can be marked as in this data: negative, neutral, and positive.  The goal of the model is to predict, from the text, what sentiment class the tweet is.  Once the unknown tweets are classified the airlines can determine how to improve or see what they are doing right.

### Data

The Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").  The data is on kaggle at https://www.kaggle.com/crowdflower/twitter-airlin.

The twitter data has 14639 rows and 15 columns.  Each column is a characteristic of the tweet and each row is a unique tweet.  The 15 columns can be divided into three types:

1. Qualitative:

*airline_sentiment:* classification of negative, neutral, or positive
*negativereason:* reason the flight was bad (late, canceled, etc.)
*airline:* US airline name
*airline_sentiment_gold:* classification of negative, neutral, or positive
*name:* twitter username
*Negativereason_gold:* reason the flight was bad (late, canceled, etc.)
*text:* text of the tweet
*tweet_coord:* longitude and latitude
*tweet_location:* city or location of tweet
*user_timezone:* timezone of user

2. Quantitative:
   *airline_sentiment_confidence:* confidence number given by human
   *negativereason_confidence:* negative confidence number given by human
   *tetweet_count:* number of retweets
   *tweet_created:* date and time tweet was created

3. Unique- each value in column is unique, no duplicates:
   *tweed_id:* unique id for each tweet.

## Data Wrangling

Before analyzing the text in the tweet the non-text columns will be looked at to see if they give any insight. Some columns have really low non-null object counts. Columns *airline_sentiment_gold* has 40, *negativereason_gold* has 32, and *tweet_coord* only has 1019. These columns have a lot of missing values and won't be useful in analysis. They will be dropped. The max for *retweet_count* is 44 and is too low to be useful and will also be dropped. The two other columns *airline_sentiment_confidence* and *negativereason_confidence* seem to be human confidence of rating the tweets. These columns will not be helpful because of the human bias and will be dropped.

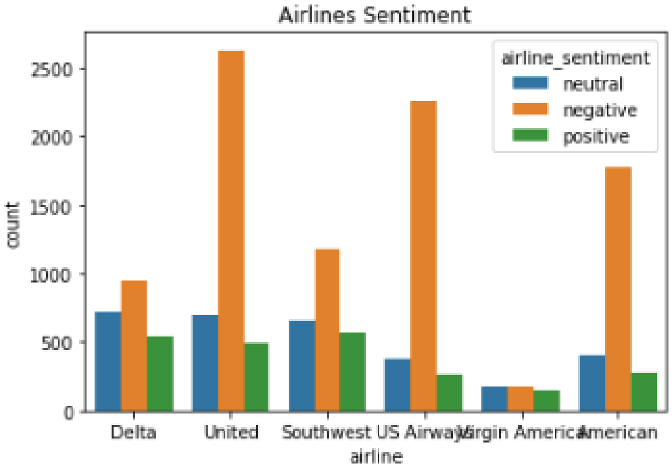

Figure 1

In the graph above we see the airline sentiments of each airline neutral, positive, and negative counts. We see that all but Virgin America have a lot more negative tweets then neutral or positive and in some cases more then both of them combined.  This means the data is unbalanced with negative being the overwhelming class.  This must be addressed when building the model.

The columns *name*, *tweet_created*, *tweet_location* when grouped by *airline_sentiment* have very low counts and might not be a great feature to find correlation with. Looking at *tweet_created* as a different type of date would probably bring the counts up. We could do day of week, hour, or month date.

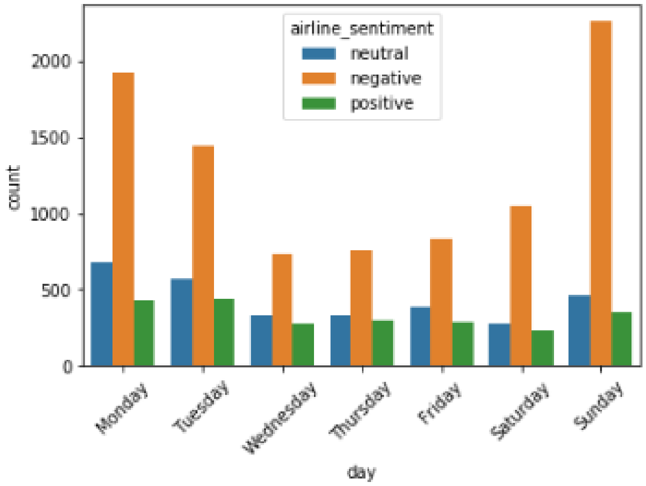

Figure 2

In figure 2 we see that most tweets were created on Monday, Tuesday, and Sunday. This is probably when people are coming home from their vacations.  Tweets by hour of the day have a normalized pattern suggesting people tweet mostly during business hours with declines in the early morning and late night. Looking at the dates of collection the data spans 8 days between February 16 and 24 in 2015.