**Team Member:**

Anuja

Byungchan Lee

Mark Hu

Belah

Qiwei Lin

**Twitter Airline Text Sentiment Analysis**

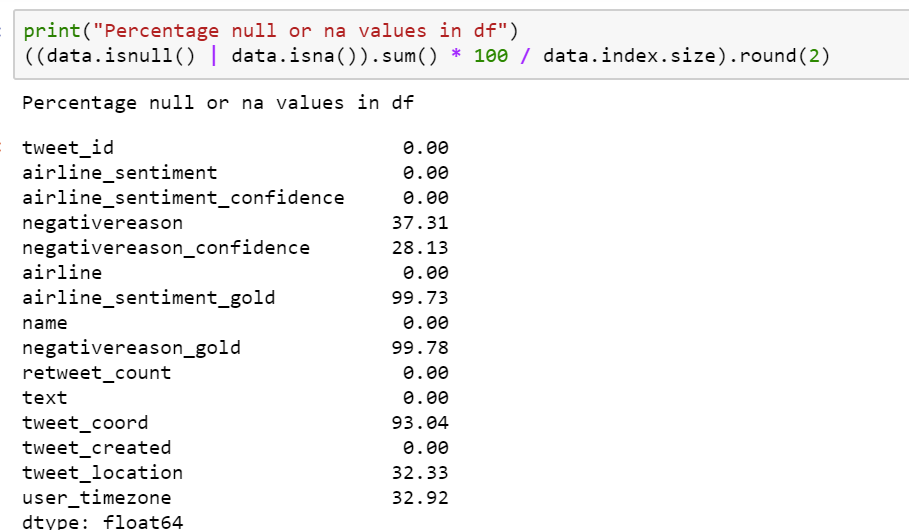
**ABSTRACT:**

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistic to identify and extract information in source material. Nowadays, sentiment analysis is a topic of great interest and development since it has a lot of practical applications. It can help the companies to analyze their market, their survey responses, product reviews, social media comments, then get important insights on their product, brand, and service. Also, business enterprise and managers can leverage opinion polarity and sentiment topic recognition to gain a deeper understanding of the companys’ values and the overall market scope while monitoring online conversations. From our online research, it is estimated that nearly 80% of the world’s data (data likely comes from social media, text, chats, surveys, articles, emails, etc.) is less organized in pre-defined manners and unstructured. If we do not spend time to analyze these texts, then it will be difficult for the business to understand and make valuable use out of them. Then you will say, analyzing all kinds of texts are time consuming, expensive and difficult, it’s not worth it. However, sentiment analysis system allows the business to make sense of these unstructured data by automating processes, saving much more hours of analyzing the data, then get the actionable insights immediately, which in other worlds, make the whole process of text analysis efficient, cost-efficient and effective. What’s more, humans are affected by emotions, but sentiment analysis is a centralized system. Hence, it can companies to apply the same criteria to all of their data, which can help to reduce errors and improve data consistency.

**Background:**

People around the world currently socially active on Facebook, Twitter, Instagram, etc. They share all kinds of information online like things that happen or has happened, their personal feelings, their insights about different events, their innovation ideas, and even their own experience. Of course, the business and the companies are aware of this kind of social media arise, and are willing to use these information in their favors. Because of this, our team wants to collect the data from social media to start our sentiment analysis. We decide to choose the airline topic by making use of tweets. We were inspired by an incident that happened on the evening of April 9th, 2017, which is the time that United Airlines forcibly removed a passenger (he had a real seat) from their overbooked flight. This incident was filmed by other passengers on the plane and immediately went viral on Twitter, Facebook. After 24 hours later, the data showed the video was viewed 6.8 million times and shared over 87,00 0 times. Among the social media, Twitter (popular microblogging service where users create status messages) is the most active platform since this platform can express all kinds of opinions about this topic, and allows feedback to be aggregated without manually intervention. In order to keep the customers, that led to the apology statement tweeted by the CEO and gave comfort and feedback to United Airlines customers. Nowadays, brands of all shapes and sizes have strong and meaningful interactions with customers, leads, and even competition on social networks. Not only the companies can monitor the customers’ needs, their business are also supervised by the customers. Also, over 100,000 flights carry passengers to and from different destinations in one day, and it’s fair to day wir travel brings out a mixed bag of emotions on people that they can tweet online.That’s why we want to detect sentiment polarity by collecting airline text data from Twitter. In this way, Consumers can use our analysis to research product and service before booking the flights; business market team can use our sentiment analysis to research public opinion of their company’s customers’ services and products; business managers can use our analysis to gather critical feedback about the newly problems or the hot topic. Since the airline topic and tweets are too large to navigate and download, we decided to use the publicly available Twitter data sets.

**Data Preprocessing and Cleaning:**



1. Removing Null values:

The null values in the data can be misleading many times and hence its important to remove them. We removed columns airline\_sentimet\_gold, negativereason\_gold, tweet\_coord as it had more than 92% null values.



1. Removing Twitter Handles (@Virgin Airlines)

The tweets contain lots of twitter handles (@Virgin Airlines), that is how a Twitter user acknowledged on Twitter. Hence, we have removed all these twitter handles from the data as they don’t convey much information.

We have created a new column tidy\_tweet, it will contain the cleaned and processed tweets. We have passed “@[\w]\*” as the pattern to the remove\_pattern function. It is a regular expression which will pick any word starting with ‘@’.

1. Removing Punctuations, Numbers, and Special Characters

We got rid of the punctuation, numbers and even special characters since they wouldn’t help in differentiating different kinds of tweets. Here we will replace everything except characters and hashtags with spaces.

1. Removing Greek characters if any

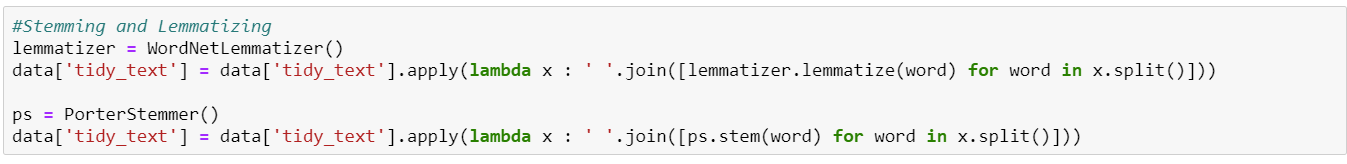
We got rid of the Greek characters using unidecode function as they are not useful in conveying any information

1. Replace with obvious words:

We replaced words like luv to love, wud to would or lyk to like as these words could be misleading for the model sometimes.

1. Removing Stop-words

A stop words are commonly used words (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore. We would not want these words taking up space in our database or take up our valuable processing time. So we remove these stop words using an inbuilt dictionary of words.

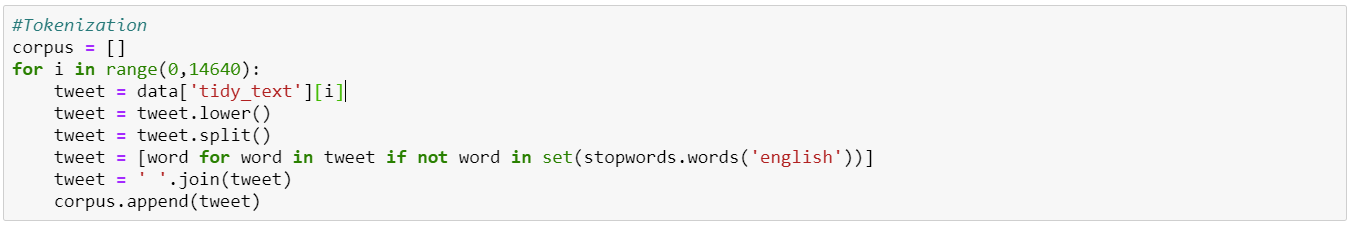


1. Stemming

Stemming is a rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word. For example – “play”, “player”, “played”, “plays” and “playing” are the different variations of the word – “play”. Hence, we have reduced these kinds of words to their root word. Here we reduce the total number of unique words in our data without losing a significant amount of information.

1. Lemmatization

Lemmatization is the process of converting a word to its base form. The difference between stemming and lemmatization is, lemmatization considers the context and converts the word to its meaningful base form, whereas stemming just removes the last few characters, often leading to incorrect meanings and spelling errors. Hence, we prefer lemmatization, but we have used both here.



1. Tokenization

We tokenize all the cleaned tweets tidy\_text in our dataset. Tokens are individual terms or words, and tokenization is the process of splitting a string of text into tokens. From a corpus ie collection of clean text we can extract unique tokens and use it to create bag-of-word or TF\_IF feature.

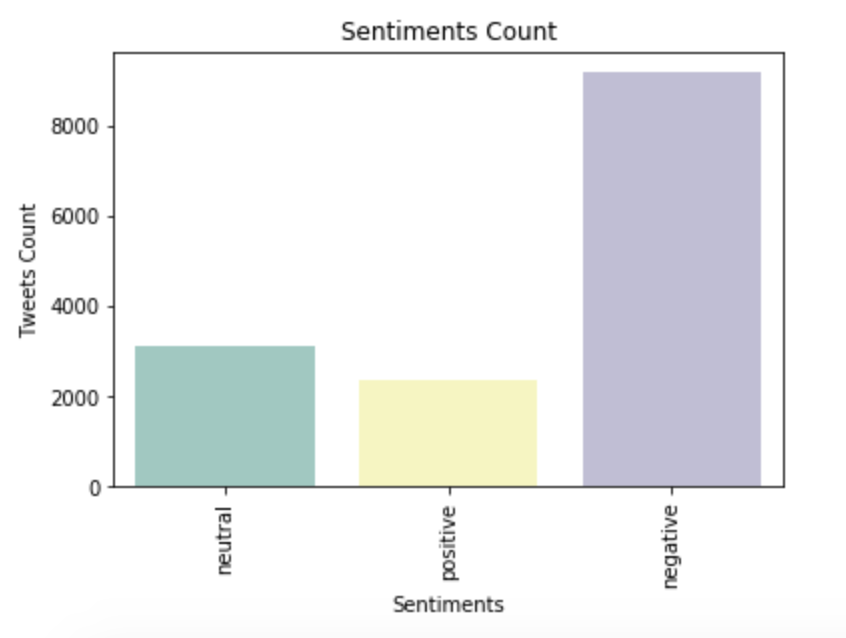
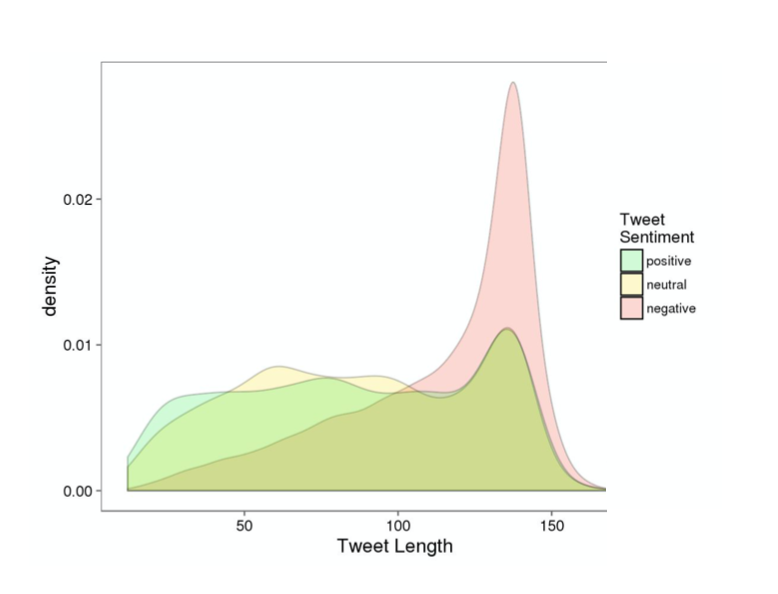
**Visualizations & Insights:**

Our data set from the Kaggle contains a variety of tweets about the six airline companies (Virgin America, US Airways, United, Southwest, Delta and American Airlines) in the U.S. with a total number of 14,640 tweets, each of them was labelled according to their sentiment polarity as : positive, negative and neutral.

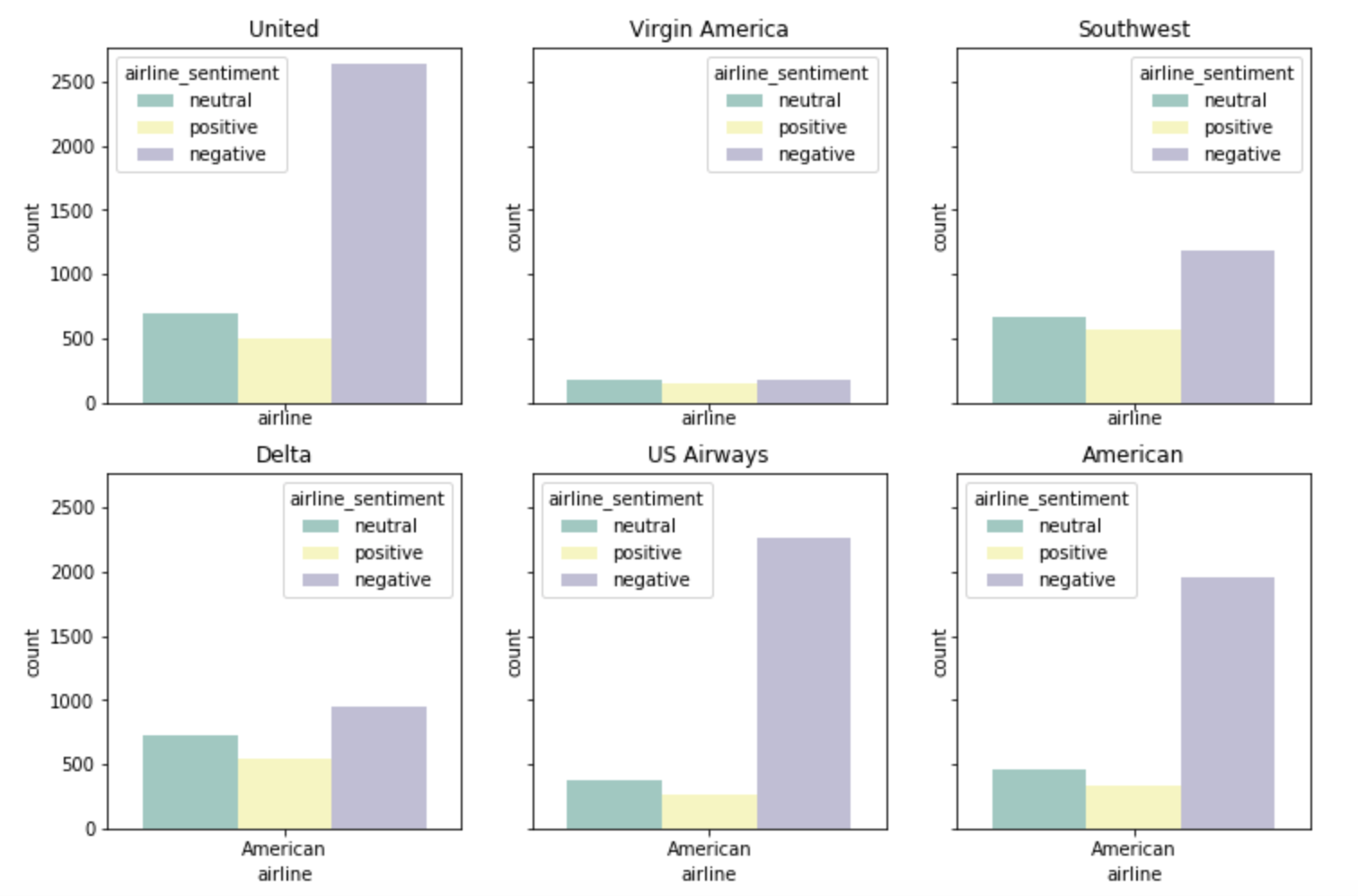
**Insights(Graph#1):** From the counts of sentiment in the graph and we found that more than 60% of the sentiments are negative, followed by neutral reviews and positive reviews to be the last. This is something that needs real attention from the airlines.

**Insights(Graph#2):** Interestingly, we found that tweet length and density for the negative tweets is the maximum which shows that people use more words to express negative feelings.

**Graph#1 : Sentiment Count**   **Graph#2 : Length & Density**

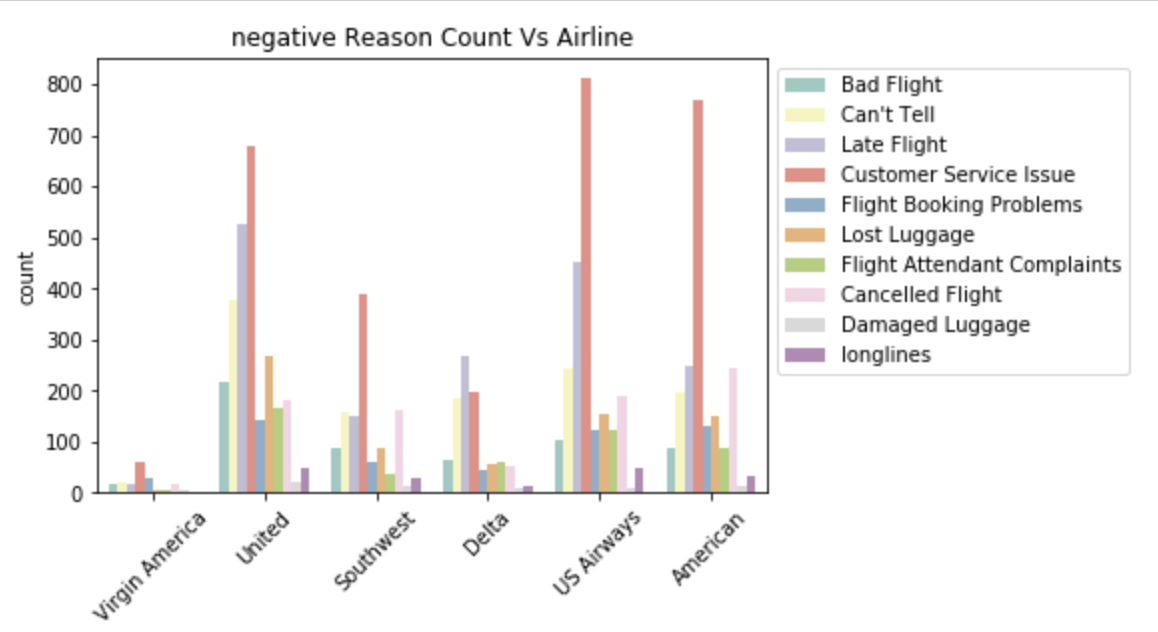
 

**Insights(Graph#3):** To gain more insights we created the bar graphs of the count of sentiment for each airline separately. From the graph, it can be inferred that United Airlines, US Airways, American airline got the maximum number of negative reviews. Virgin America fared better as compared to all other airlines with the least number of negative reviews (about a third of the total).

**Graph#3 : Sentiment Count by Airlines**  

**Insights(Graph#4):** Now the question comes to why so many negative reviews by the customers, to find this we did the graph of count of negative reasons for each airline. Overall, Customer Service Issue came out to be the main negative reason for all airlines with Late Flight being ranked 2nd. To be specific, Customer Service Issue is the main negative reason for all airlines except Delta. Late Flight is the main negative reason for Delta which is an even greater issue.

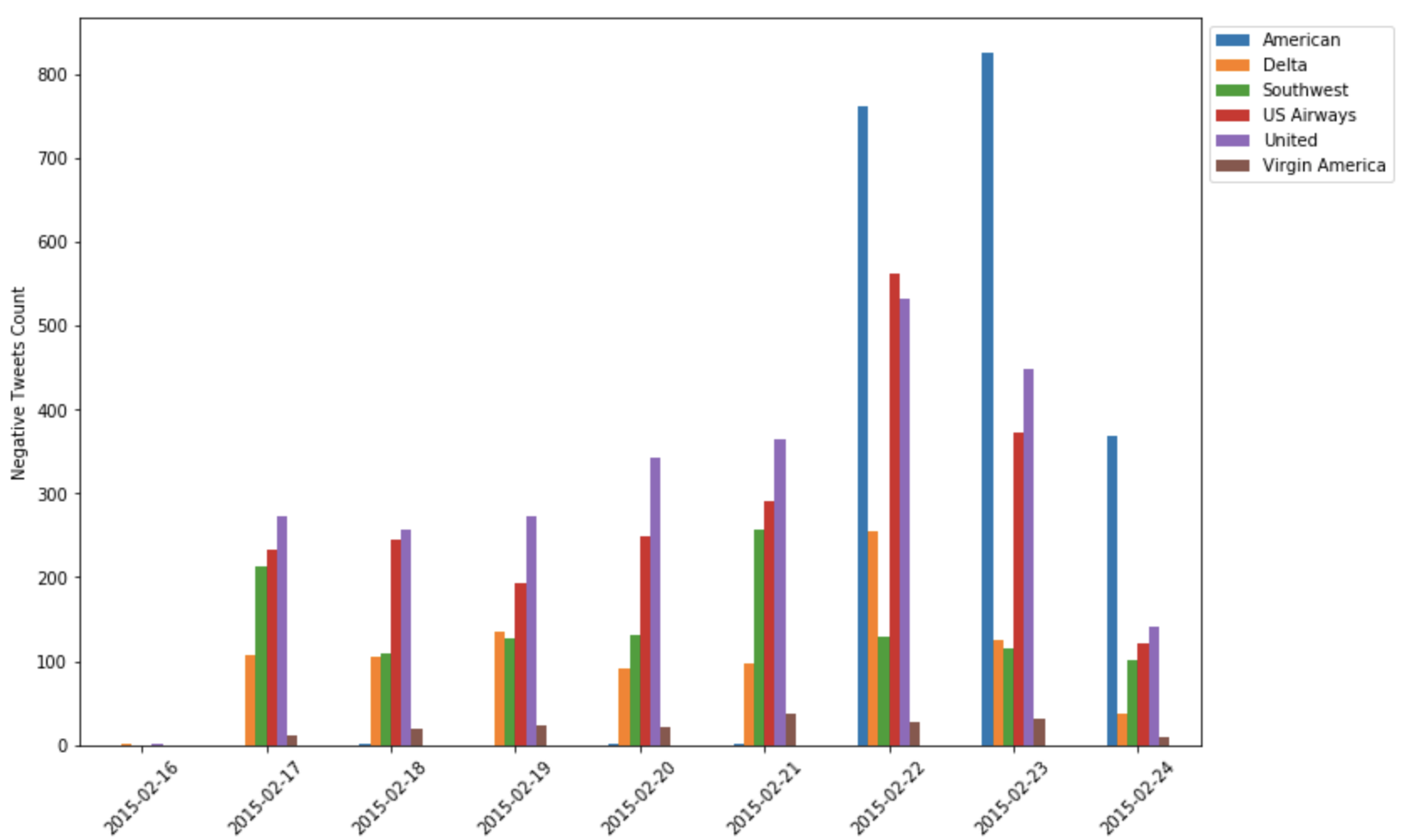
**Graph#4 : Negative reasons by Airlines**



**Insights(Graph#5):** The dataframe has data from 2015-02-17 to 2015-02-24, so we plotted the relationship between negative sentiments and date. After analyzing the graph, we found that American Airline has a sudden upsurge in negative sentimental tweets on 2015-02-23 and 2015-02-24, which reduced to half the very next day.

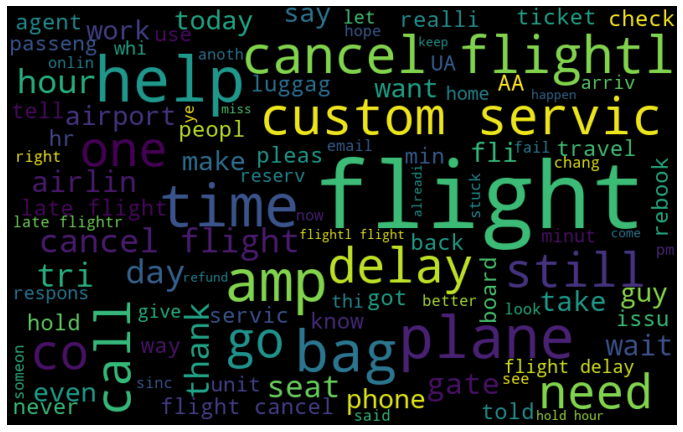
Moreover, the negative tweets for all the rest airlines is slightly skewed towards the end of the week. It might be because of the fact that more customers use the airlines during the weekends.

**Graph#5 : Negative reasons by Date**



**Insights(Graph#6):** The word cloud for negative sentiments inferred that major customer pain points are “**delayed flight**”, “**customer service**” and “**cancelled flight**”. This cross-validated the negative reasons that we inferred from the graph#4.

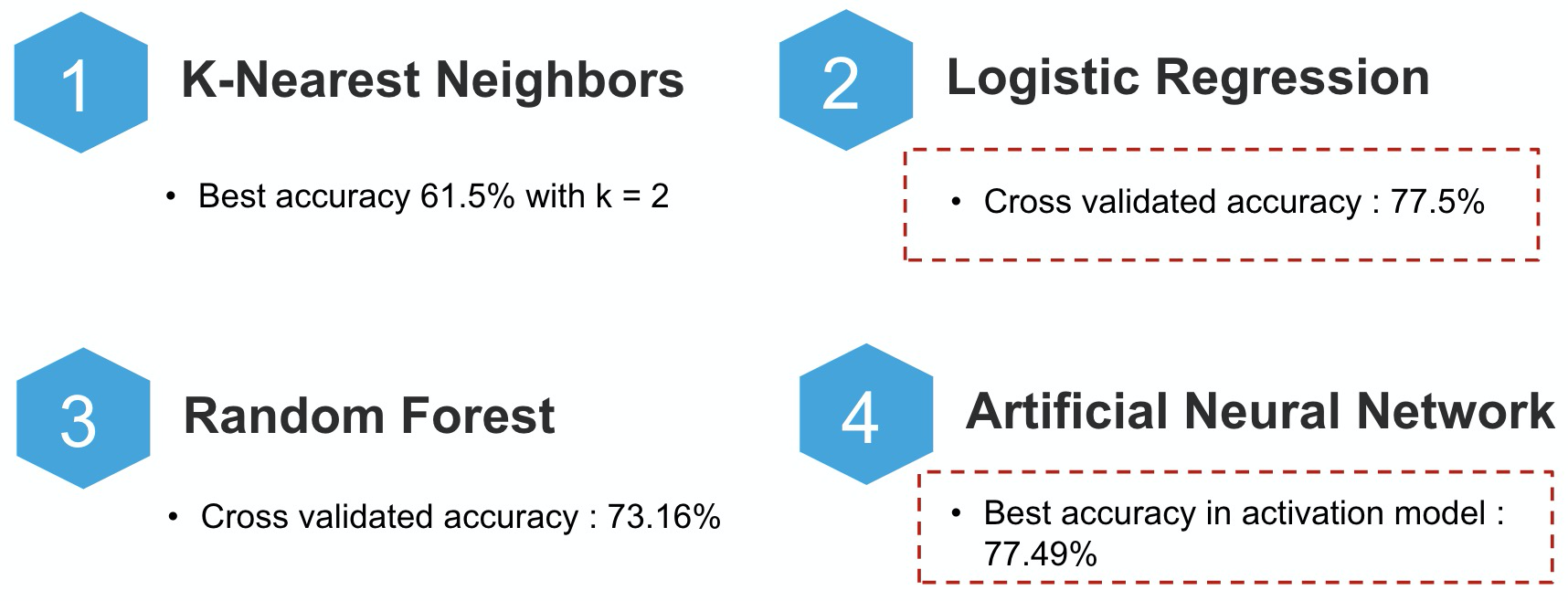
**Graph#6 : Word Cloud of Negative Airline Sentiments**

****

**Modeling and Data Analyzing:**

We got these 14641 tweets about airlines with tagged sentiment for each tweet, which allow us to use to train a model and learn to predict a tweet’s sentiment from its text.

For sentiment analysis, we chose four model, K-Nearest Neighbors, Logistic Regression, Random Forest and Artificial Neural Network, for tweet text classification based on their predicted sentiment. By training and testing models, both Artificial Neural Network and Logistic Regression models provided 77.49% and 77.5% accuracy respectfully. K-Nearest Neighbors gave 61.5% accuracy with K = 2, and Random Forest is 73.16%. Therefore we believed Artificial Neural Network and Logistic Regression models are appropriate models to predict airline-relative tweets’ sentiment prediction.

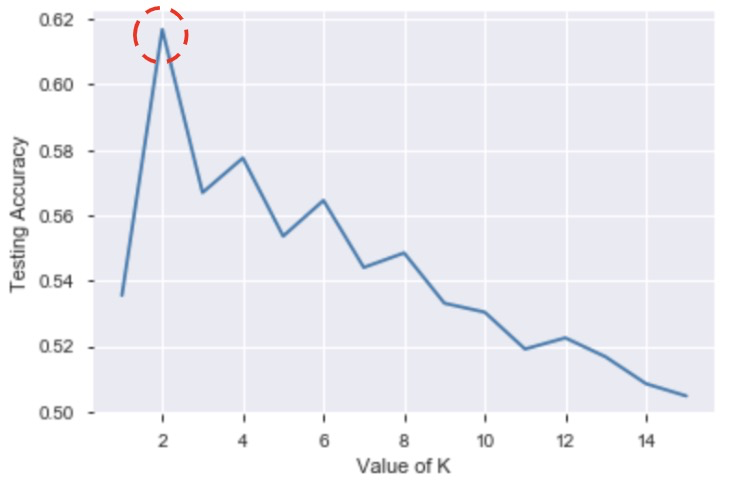


In pre-processing phase, we transformed needed data, ‘Tidy\_text’ and ‘Sentiment’ into numeric form. By adding ‘airline\_sentiment’, we tagged all positive sentiment as 2, negative as 0, and neutral as 1. For ‘Tidy\_text’, after we cleaned text, we used CountVectorizer and TfidfVectorizer in Sklearn package to convert words into a matrix of token counts. Right here, we got converted ‘Tidy\_text’ as input and ‘airline\_sentiment’ as sentiment output, so we started doing supervised analysis to classify with the above four models.

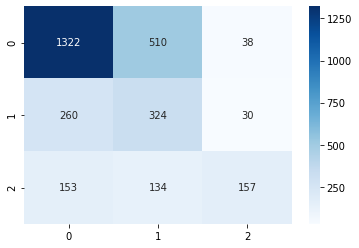
In general we split the data into 80% ‘train set’ and 20% ‘test set’ of database.

**K-Nearest Neighbors (KNN)**

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve classification. The KNN algorithm assumes that similar things exist in close proximity. Similar data points are close to each other. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics calculating the distance between points on a graph. Therefore, the key point to use KNN model is to determine the right value of K, which means how many groups we hope model to classify into. We used KNeighborsClassifier package in Sklearn to test all K values from 1 to 15 and plot each accuracy into below.



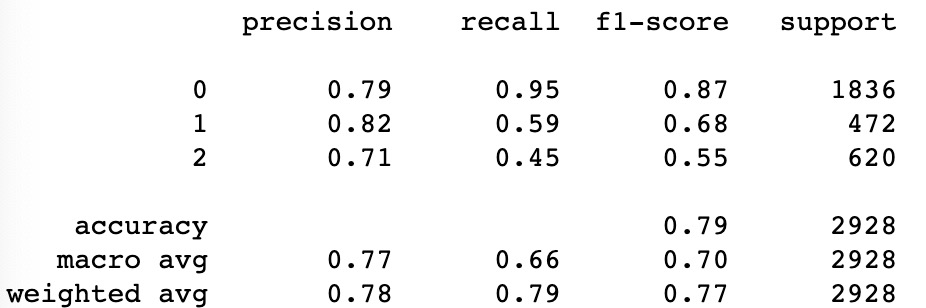
Obviously in the plot, the KNN model with K = 2 has the highest accuracy, 61.58% in this range, and the accuracy keep decreasing with increase of K, so we concluded that the best performance of KNN model is 61.58% when K = 2. This is the metrics from test set, which the model has not seen before.

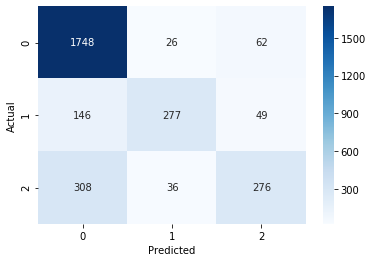


We see 1322 of negative texts, 324 of neutral texts and 157 of positive texts, have been predicted right. We noticed KNN model can predict most negative texts right. This accuracy is not good enough to use. Also, we think in original data, all sentiments have been split into 3 groups, positive, negative and neutral, but the best KNN performance showed when K = 2. We concluded that KNN model is not suitable for this data.

**Logistic Regression**

Logistic regression models the probabilities for classification problems. It’s an extension of the linear regression model for classification problems. LogisticRegression function has been included in Sklearn, so we just needed to input our ‘Train Set’ to train the model and used ‘Test Set’ to see the performance. Below is the result from once test.

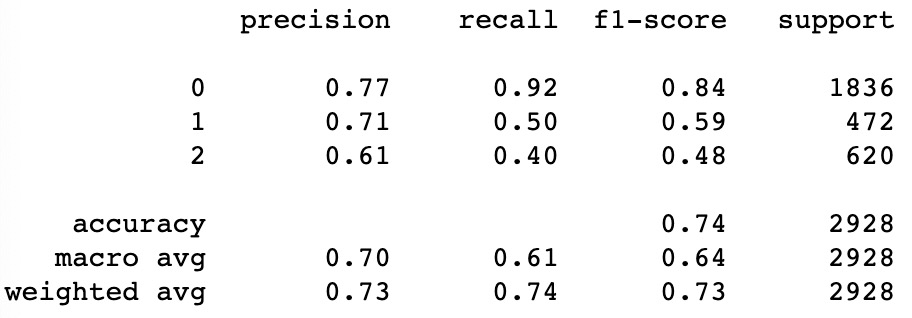


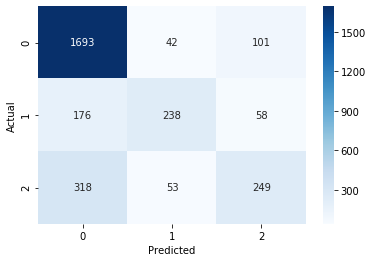


Logistic Regression gives 87% accuracy on negative texts, 68% accuracy on neutral texts and 55% accuracy on positive texts. For getting a more accurate result in the real world, we ran Logistic Regression model 4 times with different cross validation, and we look the maximum average accuracy of Logistic Regression, which is 77.5%. Although Logistic Regression is mainly used to model a binary dependent variable, it performs well on tweets sentiments prediction.

**Random Forest**

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. Similarly, we call RandomForestClassifier function in Sklearn package to train and test on our ‘Train Set’ and ‘Test Set’. Below is the result from once test.





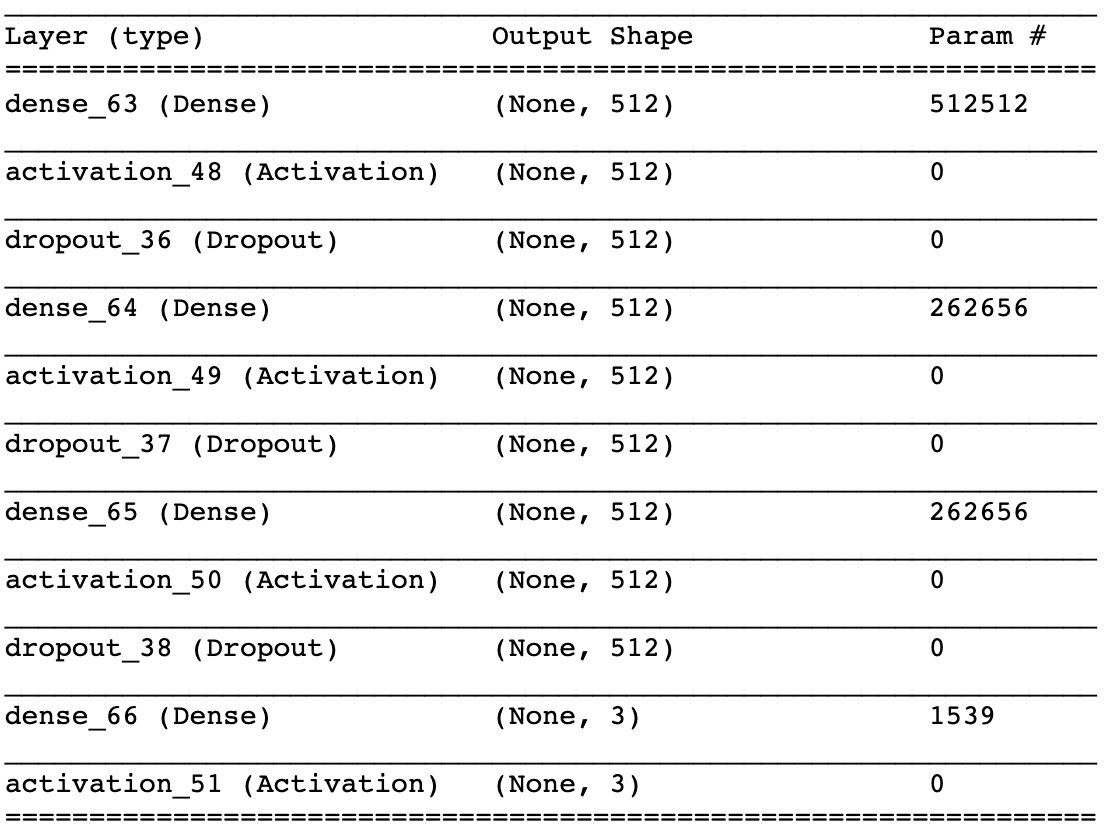
Compared with Logistic Regression model, Random Forest shows worse accuracy which is overall 74%, with 84% negative accuracy, 59% neutral accuracy and 48% positive accuracy. After running 4 more times test with different cross validation, we get the maximum average accuracy is 73.15%. By now, Logistic Regression provides better performance than Random Forest model. We think if we need bigger ‘Train Set’ to increase the accuracy of random forest model in further.

**Artificial Neural Network (TensorFlow)**

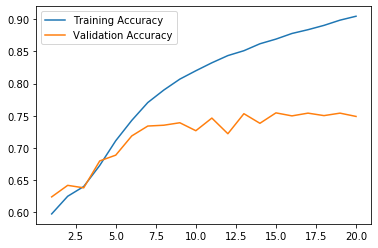
TensorFlow provides powerful tools to build the neural network, which is created by stacking layers—this requires three main architectural decisions: How to represent the text? How many layers to use in the model? How many *hidden units* to use for each layer? We test 5 different structures of model, base model, regularized model, dropout model, reduced model, and activation model to see which structure is the most efficient way to predict airline tweets’ sentiment. Based on our test results, we find out our ‘activation model’ shows the best performance with average 77.49% accuracy.



And our model structure looks like this:



It has 11 different layers in its structure. Dense layers are regular densely-connected NN layer. Activation layers we computed rectified linear function in it. Dropout layers allowed us to drop elements with probability rate. We set the probability is 0.3. Input that are kept are scaled up by = 1 / (1 - 0.3), otherwise outputs 0. The scaling is so that the expected sum is unchanged. We used different combinations of these three types of layers to figure out the most efficient one.



This graph shows the training process of the model. The blue line stands for training accuracy and yellow line stands for validation accuracy. To avoid over-fitting, we keep the model training accuracy not to be 100% right. At the same time, we noticed the testing accuracy looks like reach the top around 75% accuracy. By running evaluating function for this model, we get 77.49% accuracy on sentiment prediction.

**Summary:**

In conclusion, this analysis confirmed our assumptions on how effective sentimental analysis is. By comparing K-Nearest Neighbors, Logistic Regression, Random Forest and Artificial Neural Network model, by now we believe both Logistic Regression and Artificial Neural Network perform well on sentiment prediction. Also, they are both for better results depict clearly the sentiment of the mass crowd and thus the airline companies can easily interpret the dataset and benefit from it by improving on the aspects that get dislikes and get complained and give out negative comments by their target customers. In further steps, there is a scope of improvement in this analysis, we need more training data to deep more on these models. Because Artificial Neural Network contains further more layers and tools in its package, we think Artificial Neural Network has a big probability to breakthrough today’s limitation and reach a higher accuracy on sentiment prediction.What’s more, from our research and analysis, we discover that most negativity content or complains from airplanes are customer service, and Delta performed the best service of all the airline companies we collected.