Student Name(s): Amir Ghaderi, Mezbah Uddin, Md Shariful Islam

Student Number(s): 500794236,500793378,500801351

Course Title: Social Media Analytics

Please attach this cover page as the first page of your completed assignment before submitting it.

Assignment #	**Group Project**
Due Date	**April 3, 2017**
Group # (if applicable)	**7**

Return all assignments through the Course Assignment Submission Page (Unless otherwise specified by the instructor).

I hereby certify that I am the author of this document and all sources used in the preparation of this assignment have been cited in accordance with Ryerson's Code of Student Conduct directly or paraphrased in the document. Sources are properly credited according to accepted standards for professional publications. I also certify that this paper was prepared by me (all group members if it is a group paper) for this purpose.

Sentiment Classification of Twitter Data for American Airlines Customer Service: Delta Airlines Vs WestJet

Amir Ghaderi – 500794236
Mezbah Uddin – 500793378
Md Shariful Islam- 500801351

DS8006 – Social Media Analytics

Ryerson University

Summary:

For this project we conducted an in-depth Twitter analysis of the customer service experience from customers of Delta Airlines and of WestJet Airlines. We collected our data from Twitters Search API and were interested in tweets that were specifically directed at either Delta Airlines or WestJet airlines. We started our analysis by exploring the meta data fields of each of the datasets. We constructed a series of visualizations to help us better understand each of the datasets. Upon completion of the data preprocessing phase we performed sentiment analysis and developed various machine learning classification models. One of the major benefits of our work is that future Airlines can use our machine learning classification models on incoming tweets in order to determine the distribution of positive and negative tweets. In addition to the sentiment analysis, we also performed text analysis on each of the dataset. We split each of our datasets into positive and negative sets and performed word frequency counts on each dataset. Through the use of the negative word frequency counts we are able to discover what are the major issues that Airline customers are concerned with. Finally, we performed network analysis on each dataset in order to determine the structure and size of each dataset.

Amir Ghaderi:

- Dataset Extraction via Twitter
- Metadata exploration and metadata visualization
- Sentiment Analysis and Sentiment Visualizations
- 33% of Report Writing

Mezbah Uddin:

- Machine Learning Classification Model Development
- Text Analytics (word Frequency)
- 33% of Report Writing

Md. Shariful Islam:

- Visualization of Text Analytics
- Network Analysis and Visualizations
- 33% of Report Writing

Problem Statement:

The number of people who are heading to Twitter to contact an airline company is increasing overtime. Whether these individuals are attempting to make a compliant, ask a question, or simply share an experience. Using Twitter for customer service has become an industry standard for airline companies. In fact, according to Milward Brown and Twitter 40% of airlines travelers have used Twitter to contact an airline company. Twitter allows airlines to better communicate with their customers and address any issues immediately. It also provides airlines with an opportunity to take a negative customer experience and turn it into a positive one.

However, with the emergence of any new technologies also brings new challenges. One of the major problems with using Twitter for customer service is the volume of tweets being received. Airline companies are now required to filter through all of the incoming tweets and provide a response that is helpful and time appropriate. This process is not only extremely time consuming and resources intensive, it is also expensive. A study conducted by Yahoo Travel concluded that all major airline companies will eventually respond to an inquiry via Twitter. With all major airline companies using twitter for customer service operations, the need for a more streamlined approach to processing tweets is required.

Through our work on this project we are attempting to provide airlines with a more streamlined approach to processing all of the incoming tweets. Through the use of our machine learning classification model, airlines would be able to instantly determining whether a tweet is positive, negative, or neutral. In addition, they would be able to pass their negative tweets into our text analytics functions and determine, what are the major problems their customers are concerned with.

Literature Review

Wan & Gao (2015) in their paper used ensemble classification strategies to do sentiment analysis of customer service of 16 major airlines twitter feeds. The ensemble methods used majority voting principles of multiple classification methods including Naïve Bayes, SVM, Bayesian Network, decision trees, and random forest. The results showed that the proposed ensemble approach outperforms these individual classifiers in this airline service Twitter dataset. Melville et al (2009) in their paper showed sentiment analysis of blogs combining lexical knowledge with text classifications. Pak and Paroubek (2010) in their paper showed automated method of collecting a corpus that can be used to train a classifier. The classifier was trained by collecting tweets with positive and negative emoticons. The classifier was based on multinomial naïve bayes and able to classify documents with reasonable accuracy. Read(2005) in his paper showed how emoticons can be used to reduce dependency on machine learning classification techniques. The paper identified that traditional machine learning techniques have been successful. So, the paper used emoticons to classify articles collected from Usenet newsgroup. Using classification techniques of Naïve Bayes and SVM achieved maximum of 70% accuracy. Sreenivasan et al (2012) in the paper used official twitter accounts of Malaysian Airlines, Jetblue and Southwest to investigate issues that customers mostly talked about and how responsive airlines authorities are in addressing customer concerns. In addition, the paper provided recommendations as to how airlines can micromanage to increase customer satisfactions. Sentiment analysis with machine learning in R, R blogger site, helped with the machine learning methodologies of classification that have been used in this paper. The blog successfully demonstrated R packages and techniques that can be applied to train a model and classify using it.

Data Collection:

We collected two datasets from Twitter for this project. We decided to use Twitter because all major airlines companies are currently using Twitter for customer service purposes. Both of the datasets were extracted using Twitter's search API and R's TwitterR package. When collecting the two datasets we ensured that we only collected tweets that were directed at either Delta Airlines or WestJet airlines. We were able to accomplish this by using the "@" symbol and connecting it to the airlines official twitter accounts. For tweets direct at Delta Airlines we used "@Delta" and for tweets directed at WestJet airlines we used "@WestJet". We selected "@Delta" and "@WestJet" as our search criteria in order to limit the amount of noise in our dataset.

Preprocessing:

Prior to analyzing our datasets, a significant amount of data exploration was conducted. We constructed 3 visualizations using metadata in order to gain a better understanding of our datasets. The first visualization is a side by side histogram that shows the distribution of the number of characters per tweet for each dataset. In order to create this visualization, we created a new column in our data frames using a function that counts the number of characters in each tweet. The second visualization that was created was a bar plot which represents the number of unique users in each dataset. This illustration provides us with valuable information regarding the size of the overall dataset that would be obtained if we had access to the Twitter's firehose stream. It can be inferred that the higher the number of unique users, the larger the overall dataset is. The third visualization that was created is a bar plot that represents the percentage of tweets that are retweets in each dataset. This visualization provides us with information regarding the amount of original tweets in each dataset.

Analysis

After the initial explorations and investigations of the data frames, the column with tweets only have been exported to a csv to get the labels of the tweets. The study used Sentistrength to get the positive and negative sentiment scores of each tweet. Since we only wanted to keep the tweets with strong positives and negatives, the sentiment scores have been filtered to only keep the tweets where positive is >= 2 and negative = -1 or negative <= -2 and positive = 1. Tweets with negative score of <= -2 were labeled as "negative" and with positive scores >= 2 were labeled as "positive". After filtering and labelling the delta data frame we were left with 151 negative and 376 positive labeled tweets. The WestJet data frame contained 150 positive and 313 negative labels.

Text Analysis

Tweets were cleaned with the textscrubber function to remove punctuations, http links, special characters, capitalization etc. In addition, tdm creator function has been applied to remove stopwords and convert the corpuses into term document matrix. A term document matrix will have frequency of each word with the word itself. Also, the matrix is in descending form, so that most frequent word comes first in the data frame. In order to extract most out of text analytics, analysis have been individually done by dividing each data frame with only positive and negative labels. Therefore, we had four data frames, positive and negative for each airline. After, applying the functions to each data frame, graphs have been produced with the most frequent words appearing in the datasets. However, most frequent words in the data frames appeared to have words like flights, one, get, hate, much, thanks, longer choosing etc. These words have little or less value from the policy perspectives for the airlines. For that reason, we looked more deeply into the frequent words that has more relevance to airlines.

New data frames have been created with carefully choosing ten relevant positive and negative words. Graph 2 in the appendix gives a bar graph of the relevant words and Graph 3 creates word cloud with the positive and negative words for each dataset.

Classifier

One of our objectives for this report was also to create a classifier that automatically detects sentiments of the tweets, so that it can be used by the airlines as a system, rather than using sentistrength each time to get the labels.

So, every day, or in regular frequency, airlines will retrieve tweets using different search criteria from twitter and label the sentiments using the trained classifiers. Also, the classifier can be put on regular training schedule to better predict sentiment levels of the tweets.

The paper uses maximum entropy, SVM, Random forest, bagging and tree methods of classification. Using Rtexttools package models were trained using 400 of the 527 observations for delta and 363 out of 463 observations for westjet. Four fold cross validations was used to validate the results. Table 1 in the appendix illustrates the results of each fold classification results along with mean accuracy for delta airlines. From the table it can be seen that, maximum entropy, bagging and SVM have mean accuracy higher than 80%.

Table 2, reports the result for delta airlines. From the results it can be seen that max entropy and svm methods have accuracy higher than 80%. Whereas, Bagging, random forest and tree methods haven’t performed as good as in the case of WestJet airlines. The difference in result could be explained by the smaller training size of the Delta airlines were a bit smaller than WestJet airlines.

Network Visualizations

Graph 4 in the appendix depicts the networking visualization of Delta airlines. The network density of delta airlines is .002 which indicates that overall connectivity is very low, therefore not much conversations are going on in between the participants of the network. Diameter of the network is 9, indicating the width of the network. Reciprocity of 2% indicates that 2% of the participants are having two way communications. Centralization value of 0.4 is closer to 1, meaning the network is dominated by few central participants. Modularity of .422 is close to 1, directs that network consists of one coherent group of participants who are engaged in same conversations and paying attention to each other.

Graph 5 in the appendix depicts the networking visualization of WestJet airlines. The network density of delta airlines is .001 which indicates that overall connectivity is very negligible, therefore not much conversations are going on in between the participants of the network. Diameter of the network is 14, higher than that of delta, indicating the width of the network. Therefore, the longest of the shortest path between two nodes in the network is greater in this dataset. Reciprocity of is also 2% directs that 2% of the participants are having two way communications. Centralization value of 0.39 is closer to 1, meaning the network is dominated by few central participants. Modularity of .48 is closer to 1 than delta, directs that network consists of one coherent group of participants who are engaged in same conversations and paying attention to each other.

Overall, the network structure of the two airlines are very similar. They both have high centralization & modularity and low network density.

Conclusion

Dependence on twitter as a platform to communicate and keeping track of the customer is very important for global companies like airlines. Airlines customers are from different geographic boundaries and no better way to stay in touch with them other than social media. For that reason, in this paper, we investigated the twitter feed of two big airlines companies of US, Delta and WestJet airlines. We analyzed the twitter data using visualization's, text analytics, machine learning classifications and network analysis.

From the metadata visualizations we got important insights regarding the tweets which proved to be valuable to machine learning and text analytics. Through the text analytics, we investigated word frequency and discovered that most frequent in the context decision making doesn't add much value. So, we further probed the tweets to get a list of most relevant positive and negative words. The relevant word frequency had words like seat, luggage, airport etc. which can be looked at as areas of improvement or appreciations. Using the machine learning, classifiers were trained with five different algorithms, so that predictions can be made accurately. Also, cross validations of four folds, for each algorithm was also done to measure the generalization accuracy. These classifiers can be used by airlines policy makers to get labels of the new tweets that shall be extracted from twitter accounts. Classifiers can also be scheduled to train to better predict labels time to time at policy makers' discretion.

Network analysis provided solid knowledge as to the interactions between participants of the Delta and WestJet airlines. The network structure for two airlines are very similar with both having high centralization and modularity.

Lessons Learned

Overall, throughout the completion of this project, we have gained valuable experiences working together as a group. Firstly, we learned the importance of communicating with each group member and ensuring that we are all on the same page. Prior to starting this project, we all sat down and had an in-depth conversation regarding our topic selection and steps required to complete this project. Without this conversation the completion of this project would not have been possible. Secondly, we learned the importance of staying on schedule. We operated with a very tight timeline that did not have much room for error. In addition, we encountered scheduling difficulties when attempting to find a time that all member of the group are able to meet at. Finally, as a group we learned the importance of work integration. Each member was assigned a section of code and a section to write. Once completed all pieces needed to be bandaged together in a smooth manner.

Amir Ghaderi:

During the completion of this project I have gained valuable experiences that I will be able to add to my portfolio. I have gained experience working and manipulating twitter data in R. I have also gained experience conducting sentiment analysis and developing visualizations using the ggplot2 package. In addition, I learned the importance of selecting appropriate criteria in order to classify a tweet as negative, positive or neutral.

Mezbah Uddin:

Throughout the completion of this project, I gained a significant amount of experience working with machine learning packages. In addition, I learned the importance of selecting a classification model that is compatible with our dataset. Finally, I learned the importance of removing noise when conducting text analysis.

Md. Shariful Islam:

When completing this project, I have gained valuable skills and expertise working with a network analysis software. I learned about the value that a network visualization can bring to a project.

References

[1]Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis", 2015. [Online]. Available: http://sentic.net/sentire2015wan.pdf. [Accessed: 02- Apr- 2017].

[2]P. Melville, W. Gryc and R. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", *proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining Pages*, vol. 09, pp. 1275-1284, 2009.

[3]A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", *Conference: Proceedings of the International Conference on Language Resources and Evaluation*, vol. 2010, pp. 17-23, 2010.

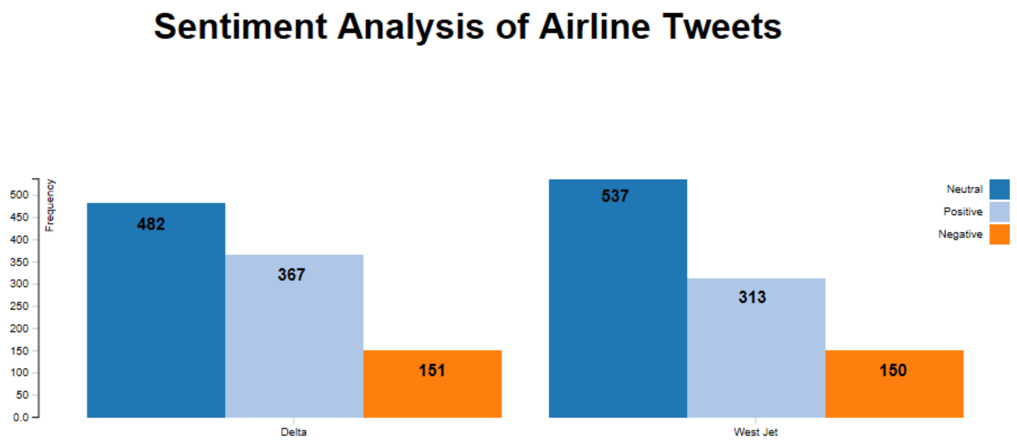
[4]J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification", *ACLstudent '05 Proceedings of the ACL Student Research Workshop*, pp. 43-48, 2005.

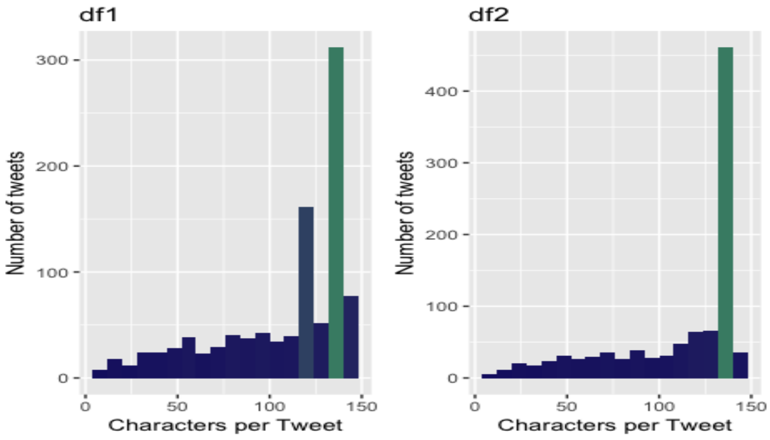
[5]N. Dharmavaram Sreenivasan, C. Sian Lee and D. Hoe-Lian Goh, "Tweeting the friendly skies", 2017. .

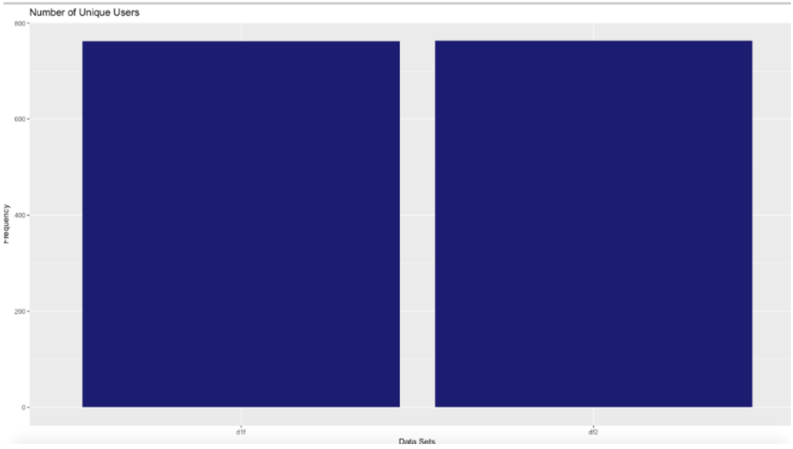
[6]T. Litsa, A. Roberts, W. Conboy and A. Roberts, "The best (and worst) ways airlines use Twitter for customer service | ClickZ", *Clickz.com*, 2017. [Online]. Available: https://www.clickz.com/the-best-and-worst-ways-airlines-use-twitter-for-customer-service/96941/. [Accessed: 02- Apr- 2017].

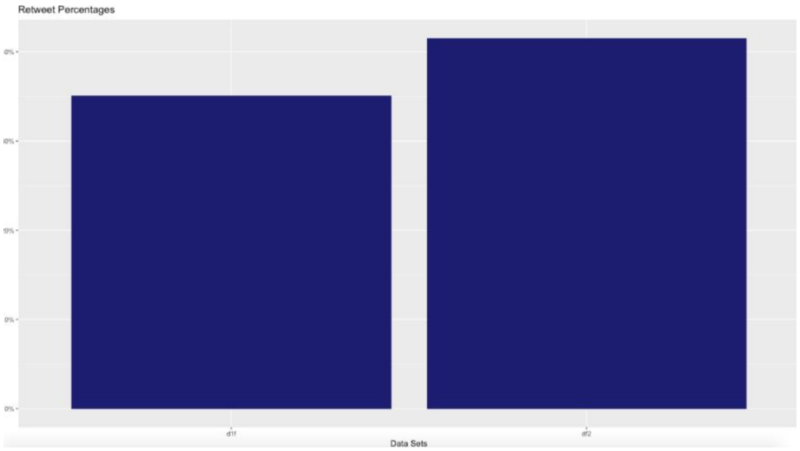
[7]"Forbes Welcome", *Forbes.com*, 2017. [Online]. Available: https://www.forbes.com/sites/grantmartin/2014/10/24/how-to-complain-to-airlines-on-twitter-and-what-youre-doing-wrong/#3e80fff978af. [Accessed: 02- Apr- 2017].

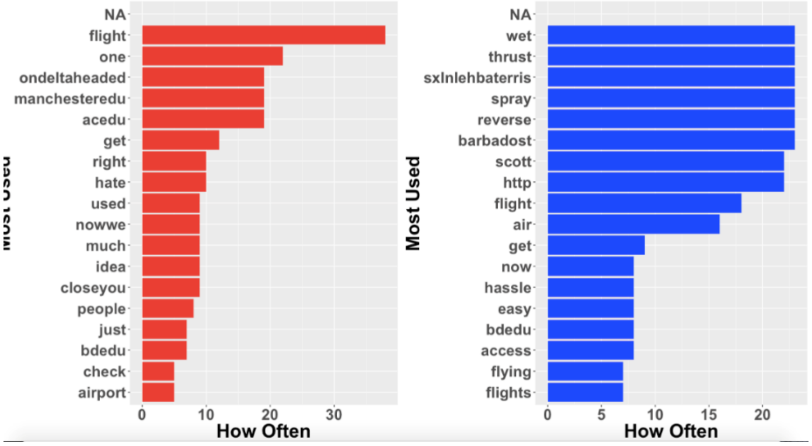
Appendix


Sentiment Analysis of Airline Tweets
Frequency
500
450
400
350
300
250
200
150
100
50
0.0
482
367
151
537
313
150
Delta
West Jet
Neutral 🟦
Positive 🟦
Negative 🟧

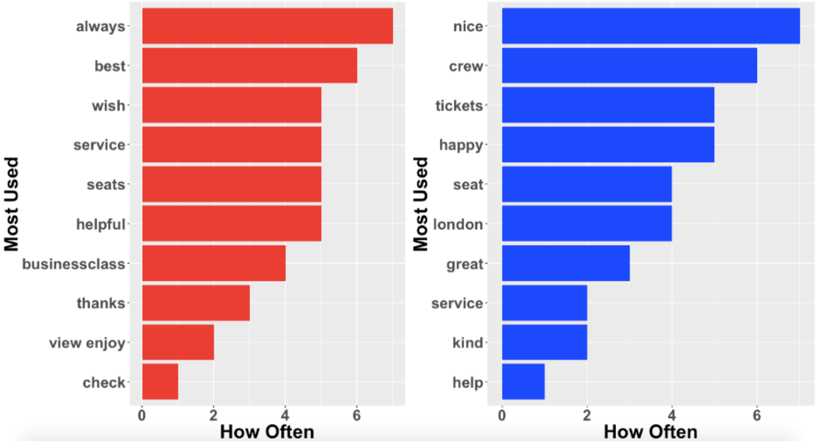

df1
df2
Number of tweets
300
200
100
0
400
0
50
100
150
Characters per Tweet


Number of Unique Users
Frequency
800
600
400
200
0
df1
df2
Data Sets

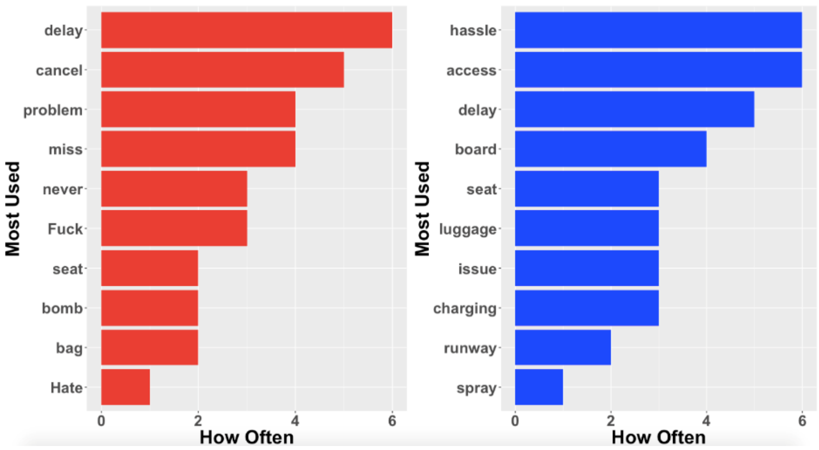




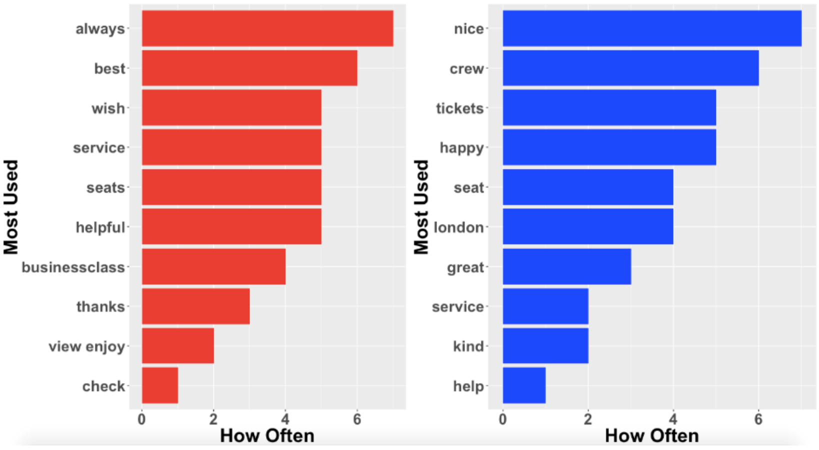
Graph: Most frequent negative word in Delta and Westjet



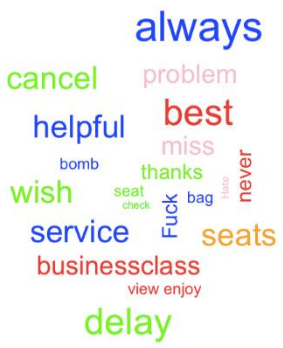
Graph: Most frequent postive word in Delta and Westjet



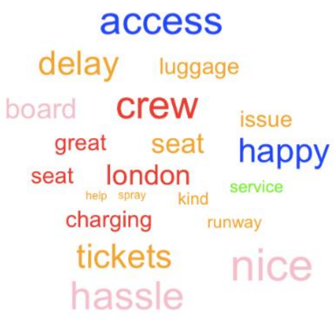
Graph: Most relevant negative words



Graph: Most relevant positive words



Word Cloud Dataframe 1 with relevant words



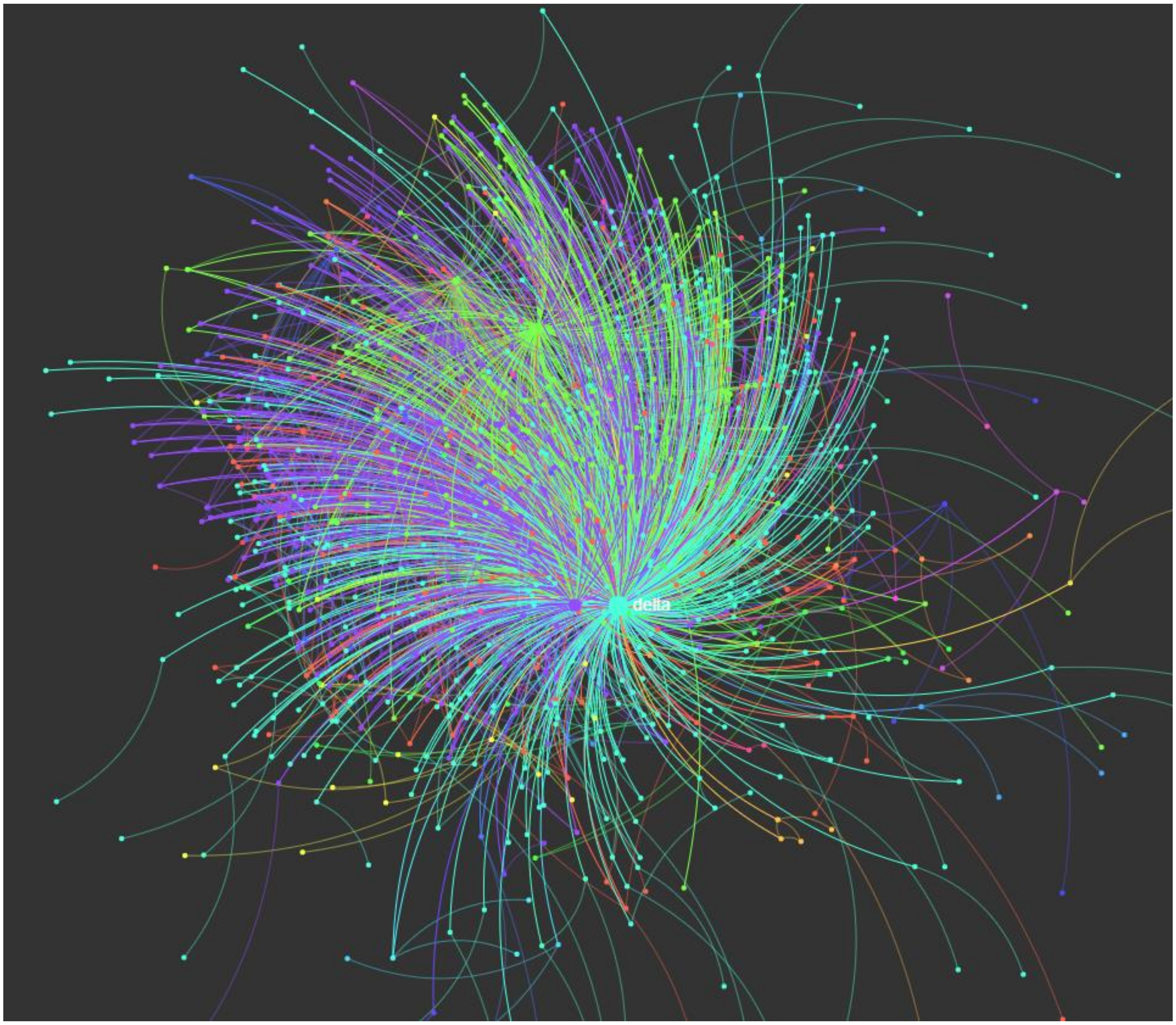
Wordcloud dataframe 2 with relevant words

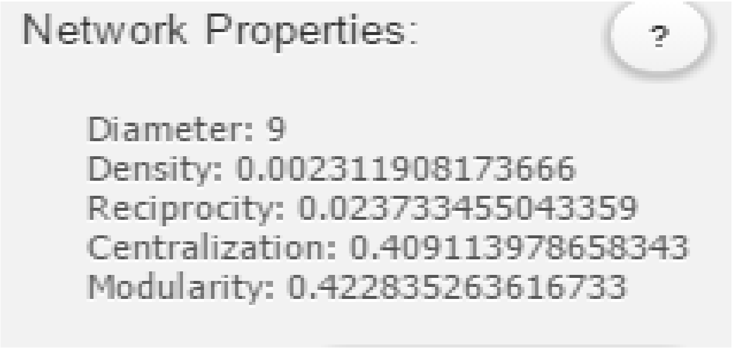
	Fold 1	Fold 2	Fold 3	Fold 4	Average
Max Entropy	100%	100%	100%	100%	100%
Tree Method	77%	74%	79%	73%	73%
SVM	88%	86%	80%	85%	85.1%
Random Forest	84%	76%	85%	73%	79%
Bagging	85%	82%	78%	81%	82%

Table 1 : Classification accuracy for delta airlines dataframe

	Fold 1	Fold 2	Fold 3	Fold 4	Average
Max Entropy	95%	91%	93%	90%	92%
Tree Method	55%	58%	65%	64%	64%
SVM	68%	58%	65%	64%	85.1%
Random Forest	68%	68%	73%	62%	68%
Bagging	65%	61%	65%	63%	63%

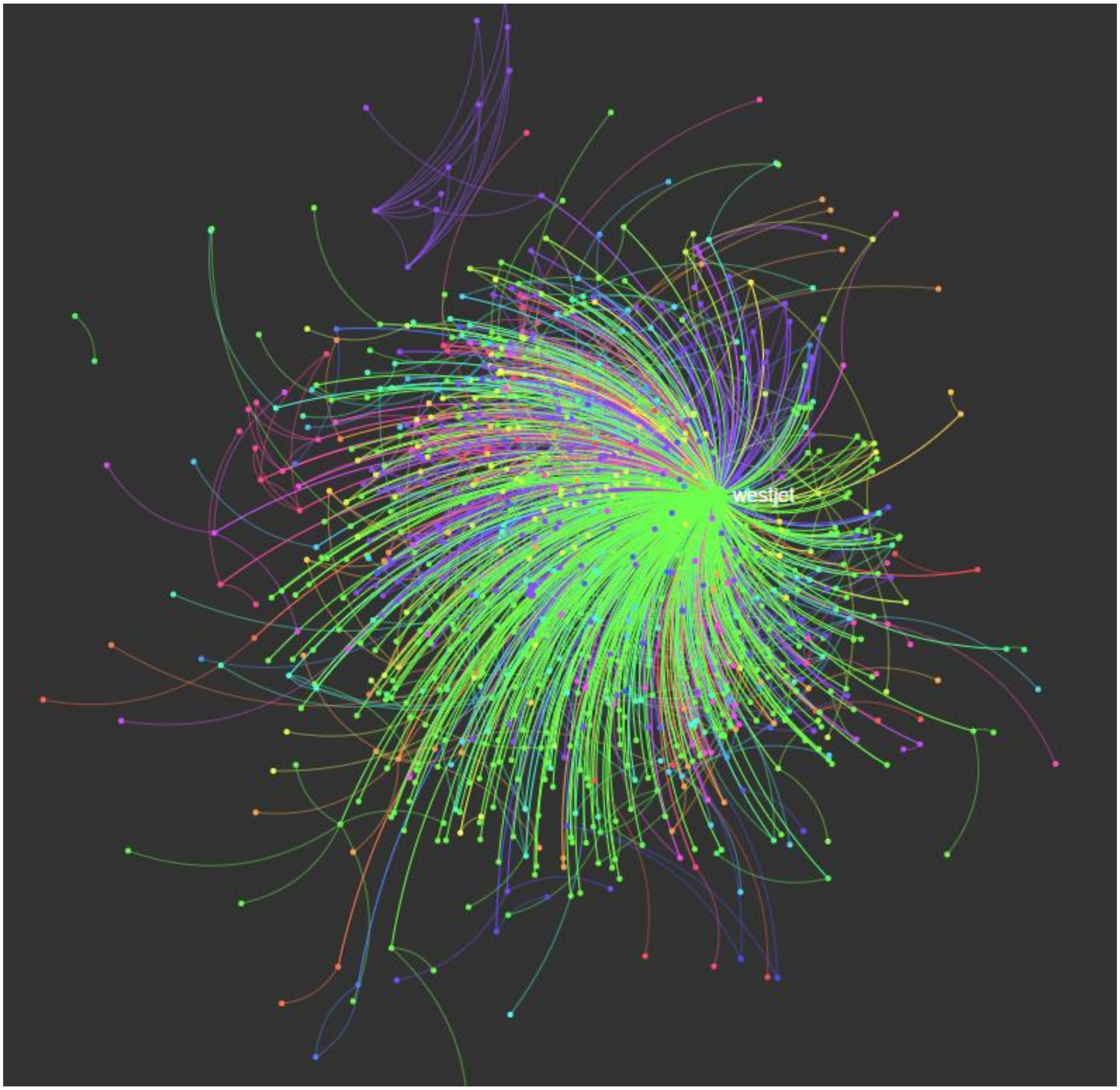
Table 2 : Classification accuracy for westjet airlines dataframe

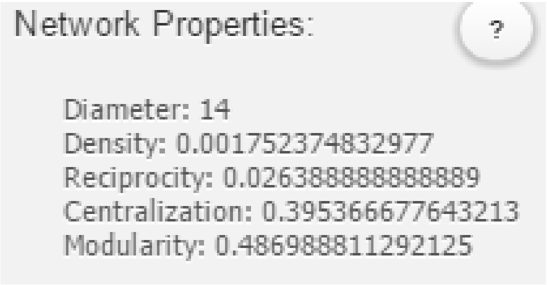




Network Properties:

Diameter: 9
Density: 0.002311908173666
Reciprocity: 0.023733455043359
Centralization: 0.409113978658343
Modularity: 0.422835263616733





Network Properties:

Diameter: 14
Density: 0.001752374832977
Reciprocity: 0.0263888888888889
Centralization: 0.395366677643213
Modularity: 0.486988811292125