PSTAT 135 Final Report: Sentiment Analysis of Airline Tweets

University of California, Santa Barbara
Spring 2020

Prepared for
Dr. Kloke

Prepared by
Ana Caklovic 3718673
Daniel White 3919586
Natalie Rozak 5101563
Randy Zhu 8554123

Table of Contents

List of Tables and Figures

Report

Introduction

Our research goal was to build machine learning models that classify tweets about airlines into the categories "positive" or "negative." To do this, we created a corpus of common words from tweets about airlines and used Spark and Python to analyze the data. While constructing our models, we explored the data to learn about trends between variables, discover commonly used words in airline tweets, and learn which terms are important for determining sentiment.

Sentiment analysis of airline tweets is useful for social media managers of airlines and airports. They can use our model to determine whether a set of tweets are positive or negative, then "like" tweets that are flagged as positive and respond to tweets that are flagged as negative. If a program gets implemented to automatically "like" positive tweets, then our model can be a time-saving tool for airline companies. Furthermore, the insight we generated about influential words for classification will help these employees understand the limitations of our model. Airline travelers can also benefit from the exploratory data analysis we performed as they choose an airline for their next trip.

Data Overview

We used a dataset from Kaggle called "Twitter US Airline Sentiment." Kaggle datasets are available to the public so we are authorized to use this data. Furthermore, the data was acquired in a legal matter since Twitter allows scraping. This dataset contains tweets about airline travel experiences from February 2015, along with several attributes related to these tweets.

Types of Variables

After cleaning, our dataset contained 14595 rows and 5 columns. Here are the variables in our final dataset:

1) `text`: the tweet written by the twitter user (string)
2) `label`: identifies tweets as positive or negative (string)
3) `negativereason`: gives the cause of an unpleasant experience (string)
4) `airline`: links tweets with a particular flight company (string)
5) `retweet_count`: gives the count for the number of retweets (integer)

Also, we created a corpus of 349 words using the tweets stored in the variable `text` and used this vector in the machine learning component of our project.

Summary Statistics and Visualizations of Training Dataset

The word cloud for the `text` variable shows "flight," "united," "usairways," "americanair," "southwestair," and "jetblue" are the six most common words (see Figure 1). We expect to see these words when reading text about air travel, so our data seems like a good representation of tweets related to flying. The histogram of text length shows that the distribution is heavy tailed, meaning there are many values at every point in the domain of the graph (see Figure 2). Tweets range from 1 to 36 words, and the average length of a tweet is 17.55 words. Our `label` column has 3852 “positive” values and 6371 “negative” values. The dataset is somewhat imbalanced since only 37.68% of our training observations are in the positive class.

The bar chart for `negativereason` shows customer service issue is the largest category with 1992 observations and late flight is the second largest category with 1175 observations (see Figure 3). Damaged luggage and long lines were the smallest categories with 48 and 121 data points, respectively. We also used a bar chart to learn about categories of the `airline` variable. There are a similar number of tweets about Delta, US Airways, Southwest, and American (see Figure 4). United has the greatest amount of tweets (2669 observations); Virgin America has the least amount of tweets (356 observations).

Summary statistics were used to understand the distribution of `retweet_count`. The minimum number of retweets was 0, and the maximum number of retweets was 32. Since the mean and standard deviation are very small, most tweets were not retweeted.

Statistic	Value
mean	0.0879
standard deviation	0.7559
minimum	0
maximum	32

Table 1: *Summary Statistics for* `retweet_count`

Methods

Data Preprocessing

Transforming data ensures that our analysis uses properly structured information and that the information is free from erroneous points. This step is essential because underlying trends in the data may be masked by noise, and Python cannot study data that contains improper values. To clean our data, we removed duplicate and missing observations, corrected categorical data, and selected columns to investigate. Text data needs to be transformed into a corpus of words for machine learning. We accomplished this by expanding contractions, tokenizing the text, removing stop words, lemmatizing words, and constructing a vocabulary. The features of our model consists of a corpus of words and the response of our model labels the data as “positive” or “negative.” Finally, we split our data into training and test sets: 70% of the data was used as training and 30% of the data was used as test.

Exploratory Data Analysis

Exploratory data analysis allows us to inspect important aspects within the data and observe it from many different angles, by means of statistical and visualization techniques. This proves to be a significant step in our analysis because we need to make sure the data is indeed what it claims to be, without making any assumptions about its contents. Our data visualization techniques include the generation of word clouds, histograms, bar charts, and pie charts; running Principal Component Analysis; and clustering the data with Latent Dirichlet Allocation.

Machine Learning

We tested five models:
1) Random Forest
2) Gradient Boosted Tree (GBT)
3) Logistic Regression (OLS, Lasso, and Ridge)
4) Naive Bayes
5) Linear Support Vector Machine (LSVM)

For each model, we first ran five-fold cross-validation in order to tune the parameter(s). The evaluator used to assess the model performance was the area under the precision-recall curve. Research indicated that area under the PR is a better evaluator than area under the ROC curve with an imbalanced classification problem. The best result from the cross-validation was used to build the final model. Results on the test set were analyzed with a confusion matrix: we calculated the accuracy while examining how many misclassified observations were false positives and how many were false negatives.

Data Analysis

Data Cleaning

First, we removed problematic observations. Our data frame started with 14837 rows, and after this stage, we had 14595 rows. Duplicates were not included in a dataset because they overrepresent repeated observations. We eliminated 52 rows that were exact duplicates and 34 rows that differed only by `tweet_id`. Python and Spark cannot operate on datasets with missing values; to solve this problem, we removed the 156 rows with missing values for `text`. Rows with missing values for `text` are not useful since our machine learning models make predictions solely based off of the tweet. Null values in other columns were ignored and we subsetted the data frame for complete rows during exploratory data analysis.

Then, we prepared our data frame for future use. We created a binary classification problem for machine learning by converting “neutral” `label` values to “positive.” Selecting columns for exploratory data analysis was our final step: we chose to analyze `label`, `negative`, `airline`, `retweet_count`, and `text`.

Text Transformation

To construct tweets for machine learning, we first expanded contractions of the `text` column. Words like “don’t” were converted to “do not.” Next, we tokenized the string of text. Breaks occurred at all spaces and non-alphanumeric characters were removed. All letters got converted to lowercase so to ensure all words of the same spelling are considered equivalent. Digits were changed to words and characters got eliminated from the data. We removed common terms like "the" since stopwords occur in both positive and negative tweets, adding noise to the data.

We considered building our corpus of words with unigrams and bigrams. We found that our bigrams did not provide additional information, and we chose to create our model with unigrams instead. For example, "cancelled" and "cancelled flight" have the same meaning when referring to airline travel.

Lastly, we formatted our data for machine learning. The text data got converted to a vocabulary of words that appears in 0.5% of the documents. This vocabulary is represented by a sparse vector of 349 indices. The `label` column also got converted from a categorical variable to a numeric variable: "negative" became 0 and "positive" became 1.

Exploring Data Through Visualizations

Using the data frame for exploratory data analysis, we created visualizations to learn about the composition of our dataset. The bar chart of the proportion of positive and negative reviews shows Virgin America has the largest percentage of positive tweets, and US Airways has the largest percentage of negative tweets (see Figure 5). The heatmap for the negative reason of a tweet shows that American, United, and US Airways have a large number of tweets complaining about customer service issues (see Figure 6). Virgin America has a small number of complaints in all categories; this is expected since most tweets about Virgin America are positive. Long lines is an uncommon complaint for all airlines.

To explore `negativereason` further, we created pie charts showing the percentage of data points in each category (see Figures 7, 8, 9, and 10 for airlines referenced below and DataExploration.ipynb for all). Compared to other airlines, American had the largest proportion of tweets corresponding to a customer service issue and Delta had the largest proportion of tweets corresponding to a late flight. The airline with the greatest percentage of tweets related to lost luggage was United and the airline with the second largest percentage of tweets related to lost luggage was Southwest.

Also, we subsetted the data to learn about trends related to tweets that were retweeted. 129 positive tweets got retweeted, and they were retweeted a total of 293 times; 408 negative tweets got retweeted, and they were retweeted 606 times. From this data, it seems like people are more inclined to retweet negative information. When looking into the negative reasons for retweeted tweets, we found the largest customer service and late flight were the two largest retweeted categories (see Figure 11). These two categories account for the largest percentage in original tweets, so it seems believable that they are also the largest retweeted categories.

Principal Component Analysis

Our data frame for machine learning has 349 dimensions, and we were worried that there may be multicollinearity between variables. We ran Principal Component Analysis to see a dataset with a reduced dimension could capture most of the variance. However, our scree plot of the first 20 principal components showed each component only accounted for a small percentage of the total variance (see Figure 12). The cumulative percentage of variance explained by the first 20 principal components was only 37.41%.

Since 20 principal components were not adequate for describing the dataset, we also tried 50, 100, 150, and 200 principal components. Here are our findings:

Number of Principal Components	Cumulative Percentage of Variance Explained
50	53.16%
100	68.83%
150	78.89%
200	86.27%

Table 2: *Principal Component Analysis of Train Data*

Since many principal components are needed to describe a large percentage of the variance, we decided not to use the reduced dataset. The cost of lost interpretability outweighs the benefit of working with a reduced dimension. After running our analysis, we learned about important words for determining sentiment and this would not have been possible if we used principal components.

Latent Dirichlet Allocation

Latent Dirichlet Allocation was used to cluster our machine learning dataset into topics. We maximized the log likelihood of the data to choose the best number of topics. The largest log likelihood was -344670.84, and it occurred with two topics. The topics somewhat split the data into positive and negative words (see Figures 13 and 14). Words in the first topic, like "thank" and "great" are probably common in positive tweets. Negative tweets are likely to contain many words in the second topic, like "cancelled" and "delayed" will probably appear in negative tweets. However, this was an incomplete separation, for example "thanks" appears in both topics.

We believed that there were more topics in the documents, like "delayed flight" or "bad customer service." Because of this, we ran LDA for larger numbers of clusters, like 6 and 9. However, the words did not separate into clear topics when more clusters were created. We suspect that this is because airport tweets have common words like "American Airlines" embedded within all topics. It is challenging for the model to decipher different clusters when topics share many similar terms.

Machine Learning

An analysis of the five models on the test set revealed the following results. We used area under PR to tune hyperparameters and accuracy (number of correctly classified observations divided by the total number of observations) to select the model:

1) **Random Forest**: Out of 4,372 observations, it misclassified 1,394. It had **68.12%** accuracy and the area under PR was 70.79%. This model had the <u>lowest accuracy</u>.

	0	1
1	1370	223
0	2755	24

Table 3: *Confusion Matrix on Test Set for Random Forest*

2) **Gradient Boosted Tree**: Out of 4,372 observations it misclassified 947. It had **78.34%** accuracy and the area under PR was 77.54%.

	0	1
1	591	1002
0	2423	356

Table 4: *Confusion Matrix on Test Set for GBT*

3) **Naive Bayes**: Out of 4,372 observations it misclassified 973. It had **77.74%** accuracy and the area under PR was 58.48%.

	0	1
1	522	1071
0	2328	451

Table 5: *Confusion Matrix on Test Set for Naive Bayes*

4) **Logistic Regression**:
   a) **OLS**: Out of 4,372 observations it misclassified 925. It had **78.84%** accuracy and the area under PR was 78.90%.

	0	1
1	465	1128
0	2319	460

Table 6: *Confusion Matrix on Test Set for OLS*

   b) **Lasso**: Out of 4,372 observations it misclassified 921. It had **78.93%** accuracy and the area under PR was 79.20%.

	0	1
1	469	1124
0	2327	452

Table 7: *Confusion Matrix on Test Set for Lasso*

   c) **Ridge**: Out of 4,372 observations it misclassified 890. It had **79.64%** accuracy and the area under PR was 79.36%.

	0	1
1	590	1003
0	2479	300

Table 8: *Confusion Matrix on Test Set for Ridge*

5) **Linear Support Vector Machine**: Out of 4,372 observations it successfully classified 3485. It had **79.71%** accuracy and the area under PR was 79.22%. This model had the highest accuracy.

	0	1
1	495	1098
0	2387	392

Table 9: *Confusion Matrix on Test Set for LSVM*

A comparison of the five models indicates that the LSVM model had the best performance by a very small margin. The LSVM model had the highest accuracy, and its area under PR was only slightly smaller than the Ridge Regression model (which had the highest area under PR). The team chose accuracy as the deciding metric in this instance because it allows for more interpretability. In addition, there is no significant consequence for misclassifying positives more often than negatives when working with airline tweets, so long as the model does an adequate job at classifying both cases. Logistic Regression, GBT, and Naive Bayes models had very similar accuracies to LSVM. Only the Random Forest model had substantially worse accuracy.

Analysis of Results

We also extracted feature importance from the Random Forest and GBT models and found the top ten features for each model. The top features the two models had in common were: “cancelled,” “hold,” “usairways,” “jetblue,” and “southwestair.” It’s interesting to note that three of these words are airline names.

Model	**Top Ten Features**
Random Forest	thanks, thank, hour, jetblue, http, usairways, hold, co, southwestair, cancelled
GBT	flight, delayed, southwestair, americanair, worst, hold, bag, usairways, cancelled, jetblue

Table 10: *Feature Importance for Random Forest and GBT*

In addition, we extracted coefficients from the Logistic Regression models (OLS, Lasso, and Ridge) and LSVM. The words corresponding to the five largest and five smallest coefficients of each model were:

Model	**Words of largest positive coefficients**	**Words of smallest negative coefficients**
OLS	http, awesome, thank, amazing, hi	fail, co, worst, ridiculous, unacceptable
Lasso	awesome, thank, amazing, thanks, hi	disappointed, suck, unacceptable, ridiculous, worst
Ridge	thank, awesome, amazing, thanks, hi	terrible, worst, ridiculous, suck, disappointed
LVSM	thank, awesome, thanks, hi, amazing	delayed, terrible, suck, nothing, disappointed

Table 11: *Words corresponding to coefficients of OLS, Lasso, Ridge, and LSVM*

As expected, positive coefficients mostly corresponded to positive words, while negative coefficients mostly corresponded to negative words. Since these coefficients had large weights, these words are important for predicting whether an observation is positive or negative. Words like “awesome,” “amazing,” and “thank” were common in positive tweets; words like “disappointed,” “worst,” and “ridiculous” were common in negative tweets.

Conclusion

Initial exploration of the data revealed many trends and which airlines travelers should consider. Virgin America led the way with the largest proportion of positive tweets, while US Airways trailed behind with the largest proportion of negative tweets. Many tweets complained about American, United, and US Airways customer service issues, while United proved to have the greatest percentage of tweets related to lost luggage. Customer service and late flights emerged as the most common concerns amongst the tweets. Most tweets were not retweeted, however Twitter users retweeted negative tweets more often than positive tweets.

After creating our machine learning models, we analyzed the results to see which observations were misclassified by every model. Every model misclassified 0.68% of the negative tweets and 21.90% of the positive tweets. Negative tweets that were always misclassified included words like "fleek" and "thanks" (see Figure 15). "Cancelled" and "delayed" were found in misclassified positive tweets (see Figure 16). Our team suspects that some of these tweets were mislabeled so the misclassification may not indicate a shortcoming of our models. Unfortunately, we did not have a key for joining the `results` data frame with our original data frame so we could not manually inspect the tweets to see if they were mislabeled.

Analysis of Linear Support Vector Machine Model

Our final model was a Linear Support Vector Machine. With a 79.71% accuracy, this model had the best performance on the test set. Further analysis of the machine learning results showed that the largest and smallest coefficients corresponded to positive and negative emotion words respectively, as expected.

Word Clouds of the LSVM results showed that "jetblue" and "southwestair" were the most common correctly classified words for positive tweets, while "united" and "flight" were the most common correctly classified words for negative tweets (see Figures 17 and 18). However, "flight" and "united" were also the most common misclassified words for positive tweets, and "jetblue" and "southwestair" were two of the most common misclassified words for negative tweets (see Figures 19 and 20).

The raw prediction variable indicated that the model had greater confidence in correctly classified negative tweets when compared to correctly classified positive tweets. The histograms of these variables shows that correctly classified negative tweets have more values at larger raw prediction values (see Figures 21 and 22). The mean and median of correctly classified negative tweets were 1.51 and 1.37, respectively. Correctly classified positive tweets only had a mean raw prediction of 0.88 and a median raw prediction of 0.80. On the contrary, the model misclassified positive tweets with greater confidence than it misclassified negative tweets, as seen in the histograms (see Figures 23 and 24). The mean raw prediction for misclassified positive tweets was 0.53 and the median raw prediction for misclassified positive tweets was 0.68. Meanwhile, misclassified negative tweets had a mean raw prediction of 0.45 and a median raw prediction of 0.45.

Limitations

As seen in our analysis of the LSVM model, the model sometimes misclassified positive tweets related to United, and it sometimes misclassified negative tweets related to JetBlue and Southwest. Before responding to online posts, social media managers of airlines should manually inspect negative tweets about United with small raw prediction scores to ensure that the tweet was not accidentally misclassified. Similarly, they should inspect positive tweets about JetBlue and Southwest with low raw prediction scores.

Also, attributes such as `airline` seem to have been manually entered, and this may have biased our exploratory data analysis and thus limited our study. When analyzing machine learning results, we found certain airlines mentioned that were not part of the `airline` variable. For example, the company JetBlue was referred to in many tweets but it was not one of the possible airline labels.

Furthermore, this data set only included tweets from February 2015, and we expect that airline tweets may have seasonal trends that weren't accounted for. For example, our data showed very few complaints about long lines, and this is probably because February is not a busy travel month. However, people on Twitter are likely to express dissatisfaction about long lines experienced while traveling during the holidays.

Further Studies

A potential research direction could be studying seasonal trends by using a data set of tweets gathered across a longer period of time. Airline travel experiences differ depending on the time of year, and we expect tweets to reflect these changing trends. By including data about when a tweet was posted, one could see if seasonal trends influence results and if important words in machine learning models differ depending on the season.

Another future research direction—if the location of tweets is tagged correctly—could be to explore whether certain locations are more likely to have positive or negative tweets about specific airlines. For example, Atlanta is the Delta hub, so people flying to Atlanta may have a positive regard towards this airline. One could also examine the positive and negative reasons linked to every location. One location could have a large amount of tweets about customer service issues. This would allow for airline companies to see if there are certain "problem areas" that need improvement.

References

Brownlee, J. (2020, January 8). *Tour of Evaluation Metrics for Imbalanced Classification*. Machine Learning Mastery. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

Chambers, B. and Zaharia, M. (2018). *Spark: The Definitive Guide*. O'Reilly Media Inc.

Data Camp. *TOPIC MODELING IN R* [Slides]. Retrieved from https://s3.amazonaws.com/assets.datacamp.com/production/course_11368/slides/chapter4.pdf

Drabas, T. and Lee, D. (2017). *Learning Pyspark*. Packt Publishing.

Kelvin, A. (2020, February 28). *Adding Strictly Increasing ID to Spark Dataframes*. Github. https://albertusk95.github.io/posts/2020/02/adding-unique-row-id-pyspark-dataframe/

Kloke, J. (2020). *PSTAT 135 Lecture Notes*. Personal Collection of J. Kloke, University of California Santa Barbara, Santa Barbara CA.

Li, Z. (2019). *PSTAT 131 Lecture Notes*. Personal Collection of Z. Li, University of California Santa Barbara, Santa Barbara CA.

Matplotlib: Visualization with Python [Documentation]. Retrieved from https://matplotlib.org/3.2.1/index.html

Mayo, M. *Text Data Preprocessing: A Walkthrough in Python*. KDnuggets. https://www.kdnuggets.com/2018/03/text-data-preprocessing-walkthrough-python.html

Mirza, D. (2017). *CS 8 Lecture Notes*. Personal Collection of D. Mirza, University of California Santa Barbara, Santa Barbara CA.

Oh, S. (2019). *PSTAT 134 Lecture Notes*. Personal Collection of S. Oh, University of California Santa Barbara, Santa Barbara CA.

Spark Overview [Documentation]. Retrieved from https://spark.apache.org/docs/latest/

Twitter (2015). *Twitter US Airline Sentiment* [Data file]. Retrieved from https://www.kaggle.com/crowdflower/twitter-airline-sentiment

WordCloud for Python documentation [Documentation]. Retrieved from https://amueller.github.io/word_cloud/

Ye, X. *Study Notes on the Latent Dirichlet Allocation* [Notes]. Retrieved from http://www.cis.jhu.edu/~xye/papers_and_ppts/ppts/LDA_study_notes_Xugang.pdf

Appendix

Instructions for Running Code

Before running the code, confirm that Tweets.csv is in the same folder as the Jupyter notebook files. Then, please run the files containing code in this order:

1) PackageDownloads.ipynb
2) DataPreprocessing.ipynb
3) DataExploration.ipynb
4) MachineLearning.ipynb
5) AnalyzeResults.ipynb

Also, it is important to note that these notebooks take several hours to run.

Figures

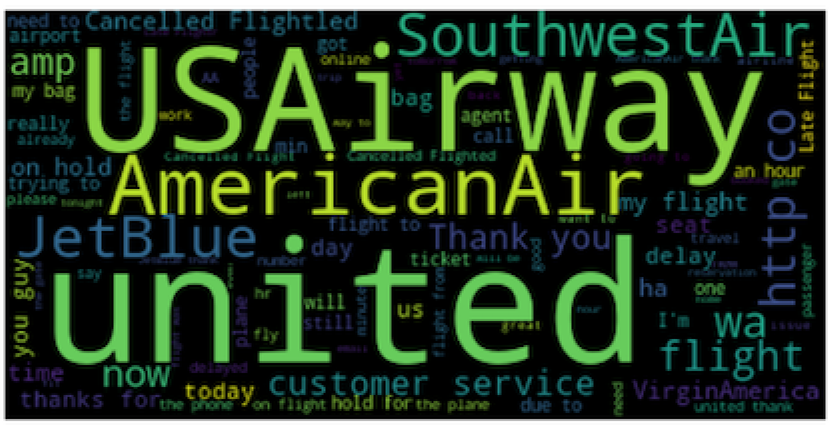

Figure 1: *Word Cloud of* `text`

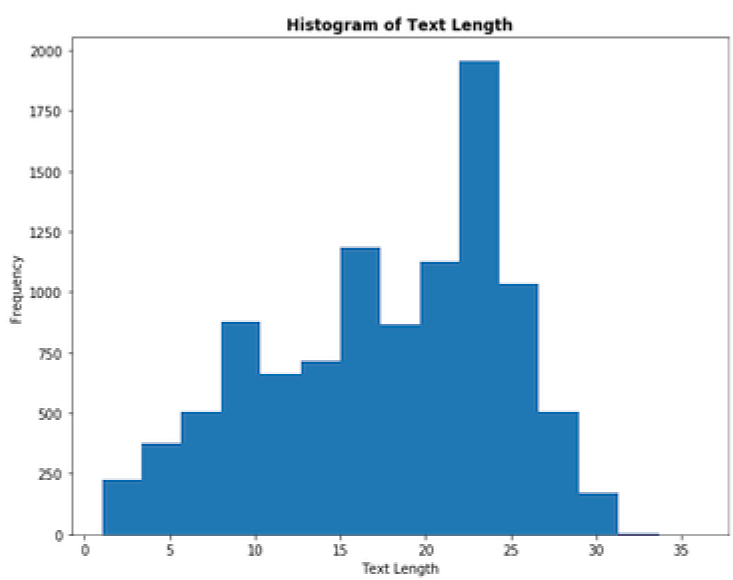

Figure 2: *Histogram of Text Length*

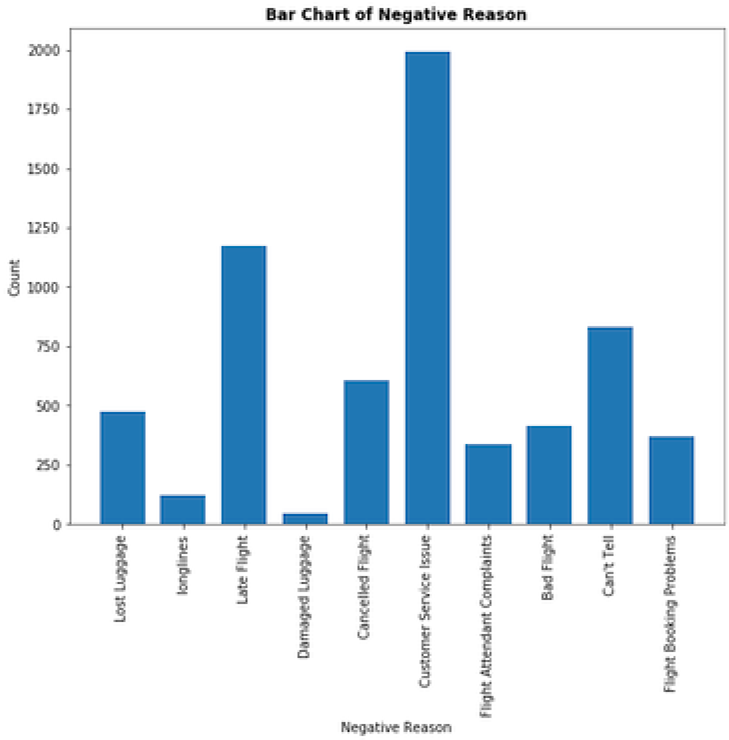

Figure 3: *Bar Chart of Negative Reason*

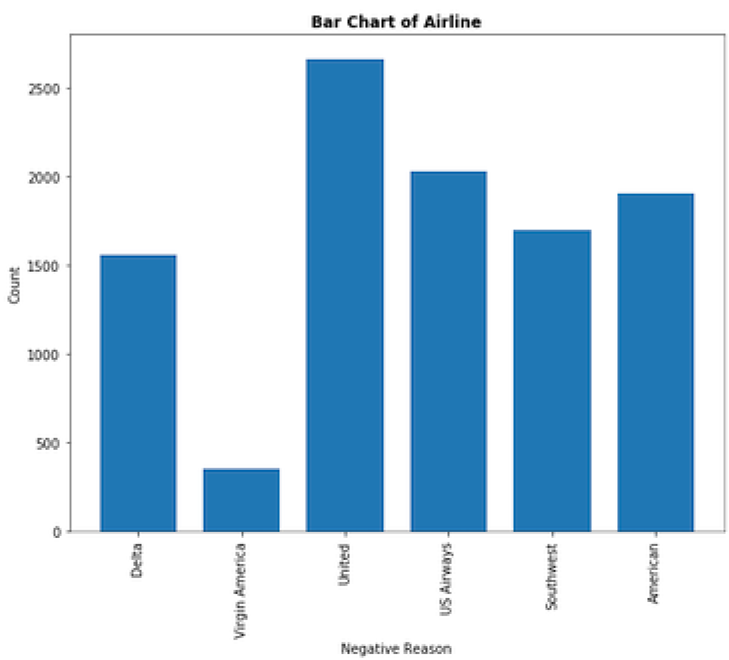

Figure 4: *Bar Chart of Airline*

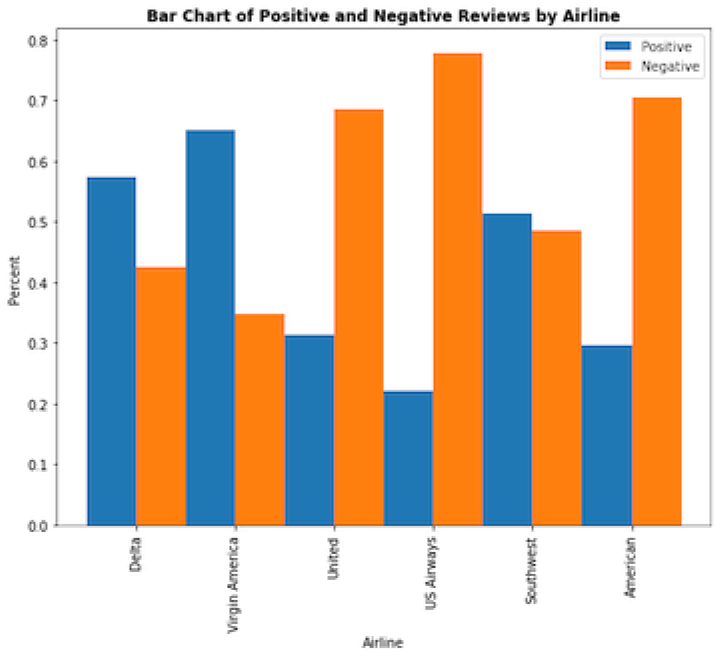

Figure 5: *Bar Chart of Positive and Negative Reviews by Airline*

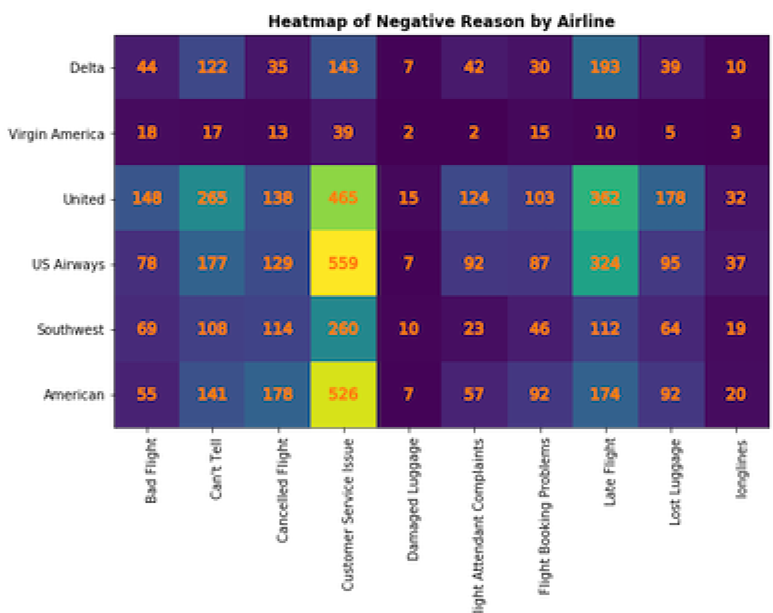

Figure 6: *Heatmap of Negative Reason by Airline*

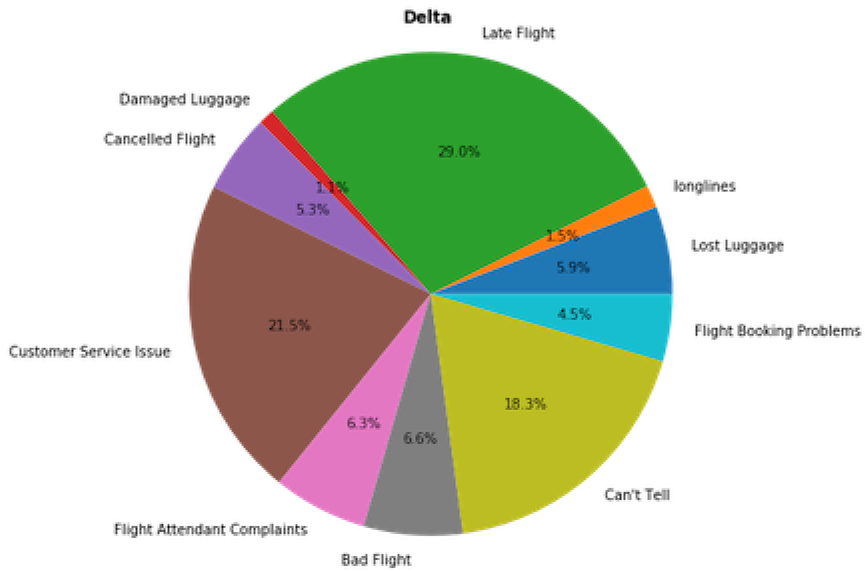

Figure 7: *Pie Chart of Negative Reason for Delta*

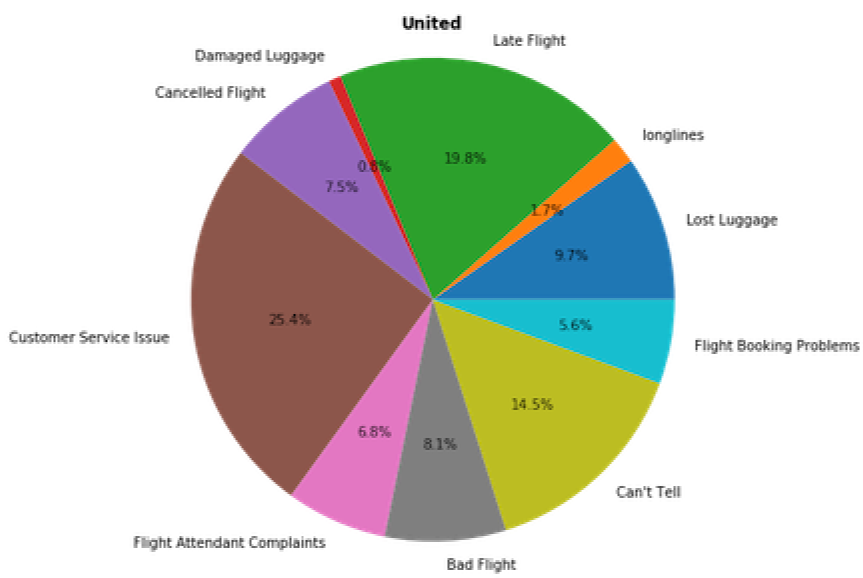

Figure 8: *Pie Chart of Negative Reason for United*

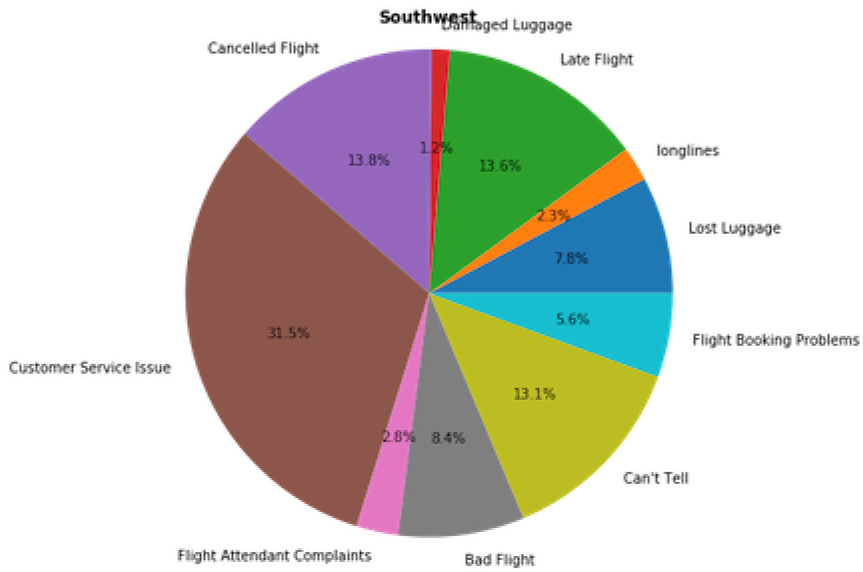

Figure 9: *Pie Chart of Negative Reason for Southwest*

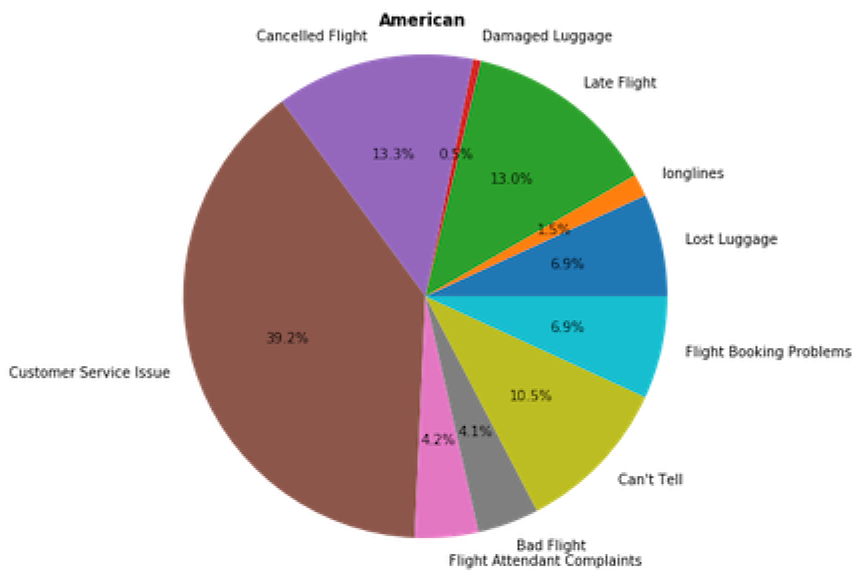

Figure 10: *Pie Chart of Negative Reason for American*

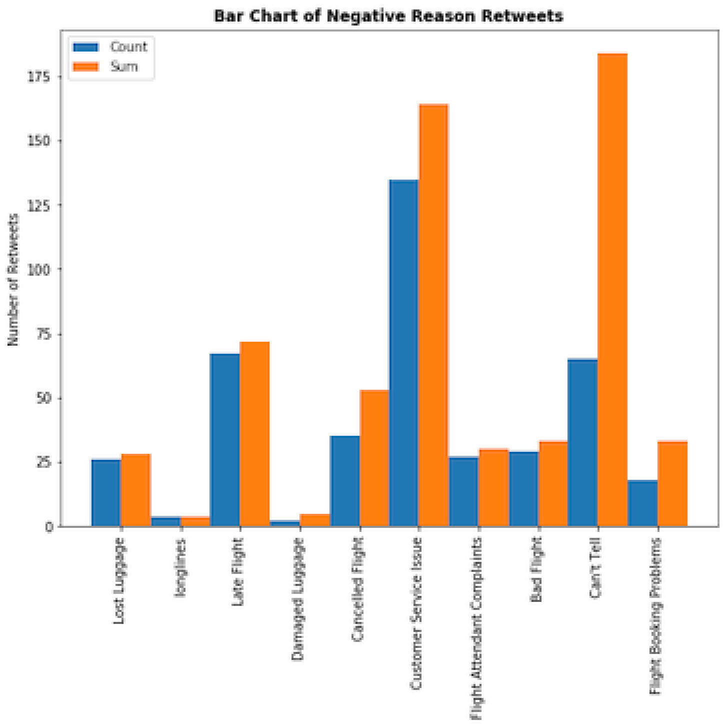

Figure 11: *Bar Chart of Negative Reason Retweets*

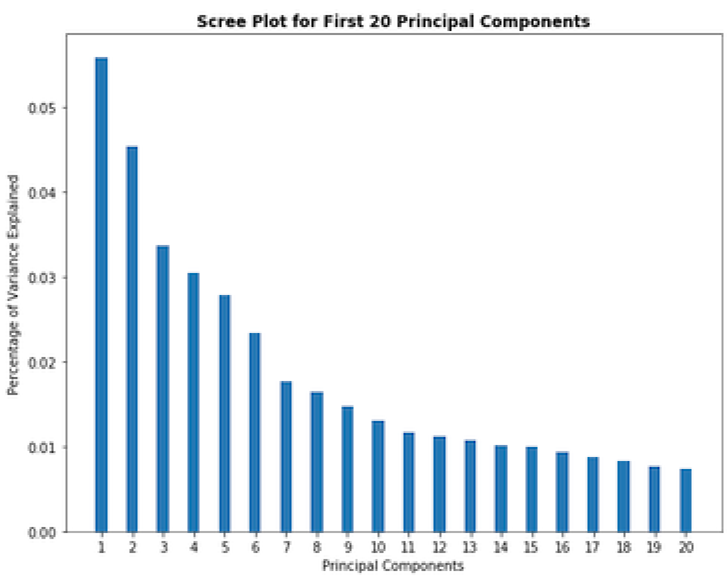

Figure 12: *Scree Plot for First 20 Principal Components*

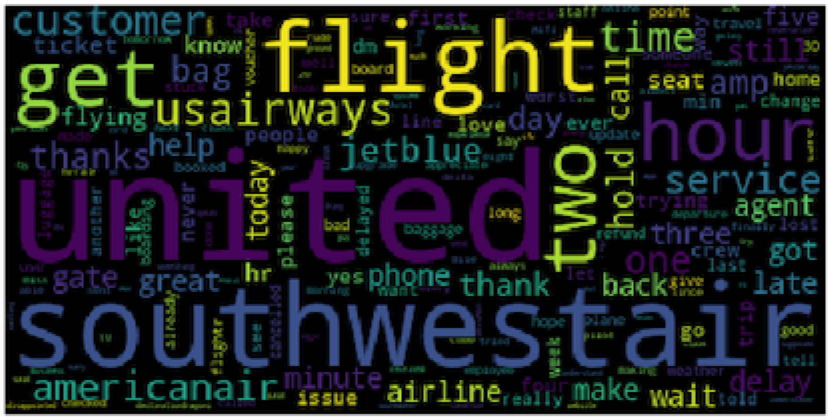

Figure 13: *Word Cloud of First Cluster from LDA*

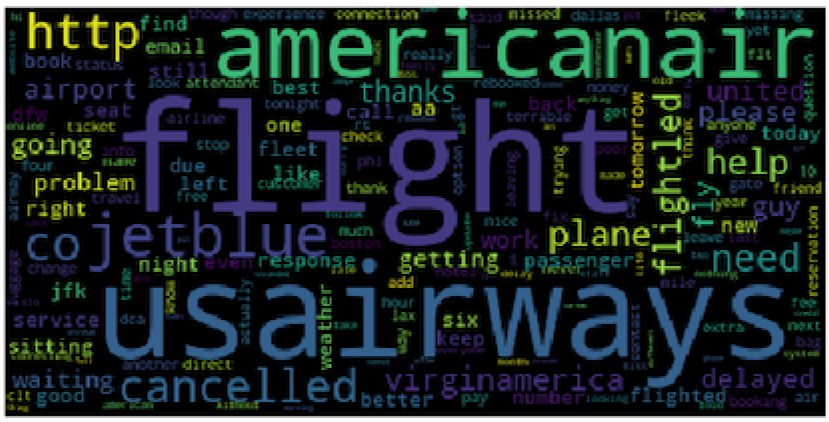

Figure 14: *Word Cloud of Second Cluster from LDA*

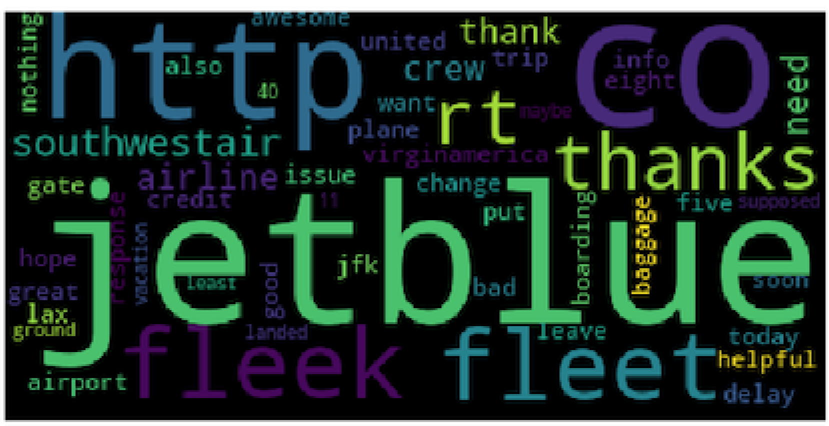

Figure 15: *Word Cloud of Misclassified Negative Tweets*

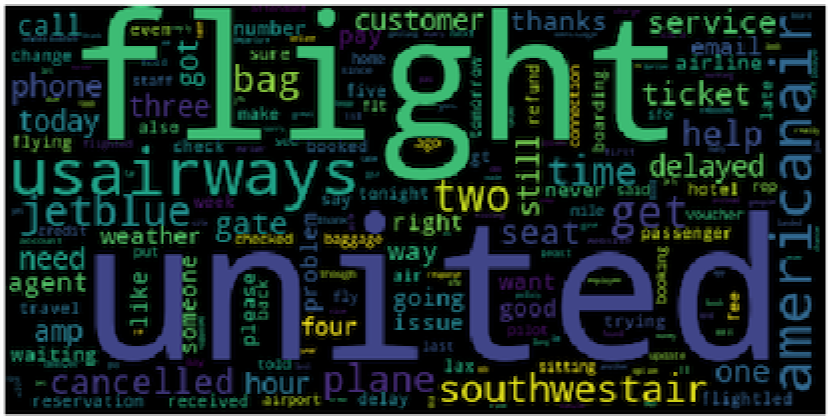

Figure 16: *Word Cloud of Misclassified Positive Tweets*

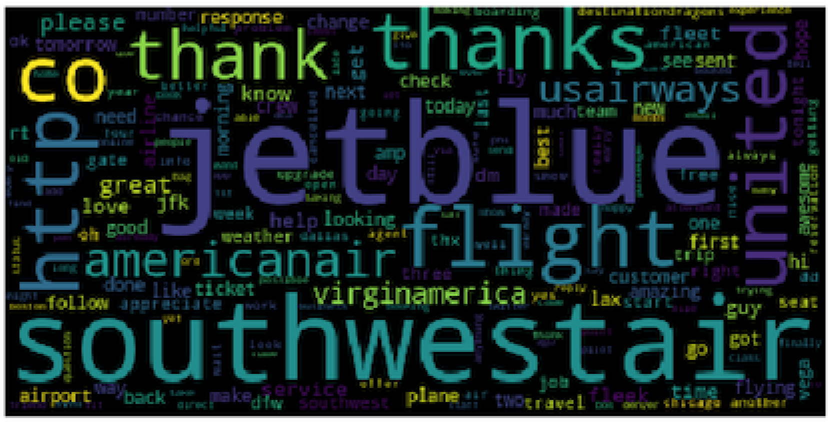

Figure 17: *Word Cloud of Correctly Classified Words in Positive Tweets from LSVM*

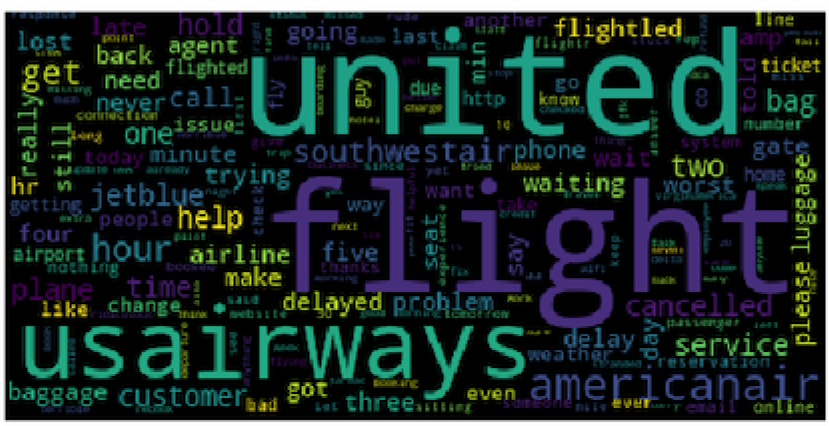

Figure 18: *Word Cloud of Correctly Classified Words in Negative Tweets from LSVM*

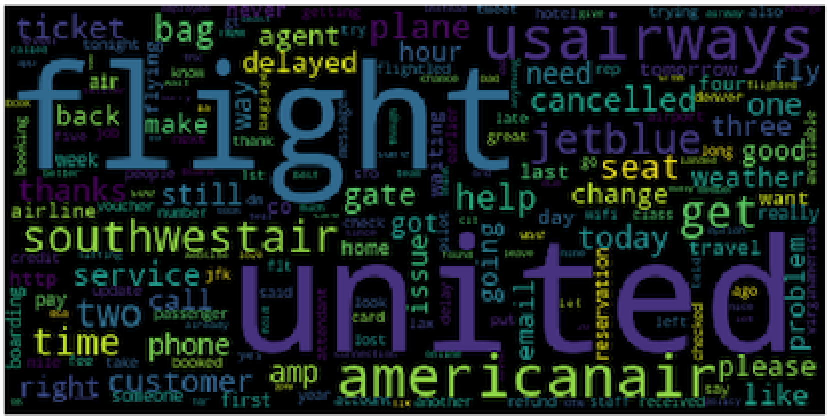

Figure 19: *Word Cloud of Misclassified Words in Positive Tweets from LSVM*

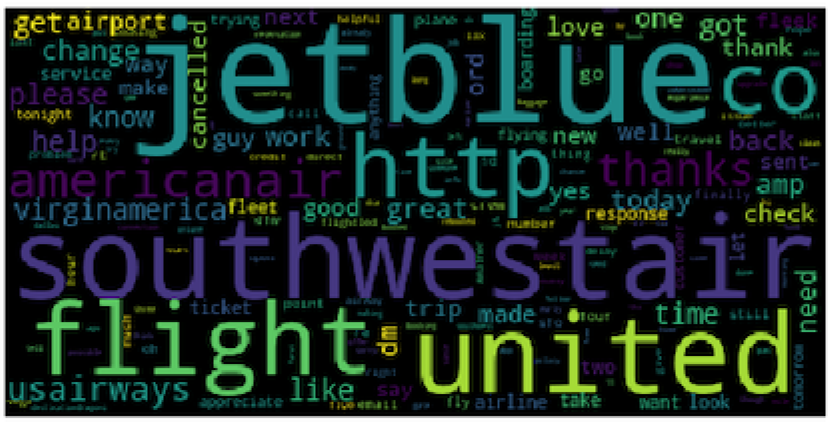

Figure 20: *Word Cloud of Misclassified Words in Negative Tweets from LSVM*

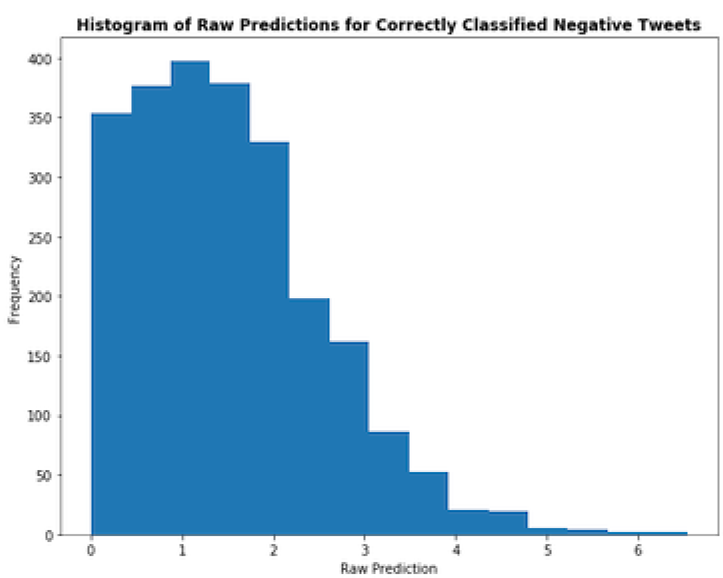

Figure 21: *Histogram of Raw Predictions for Correctly Classified Negative Tweets*

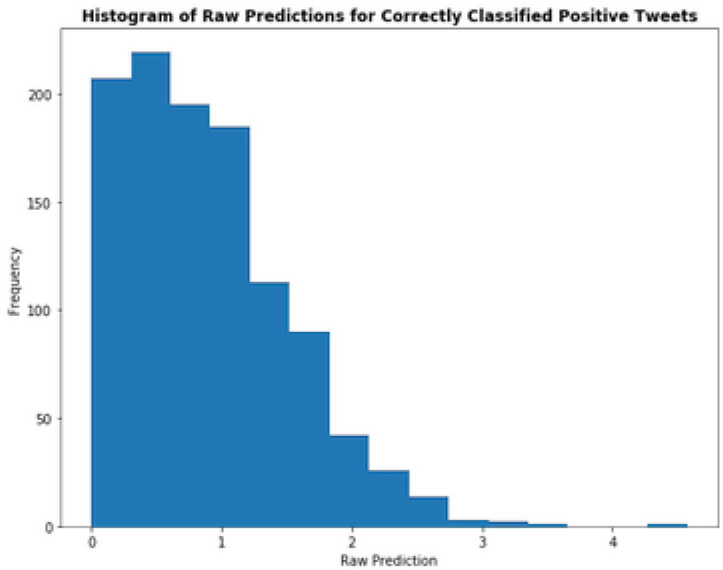

Figure 22: *Histogram of Raw Predictions for Correctly Classified Positive Tweets*

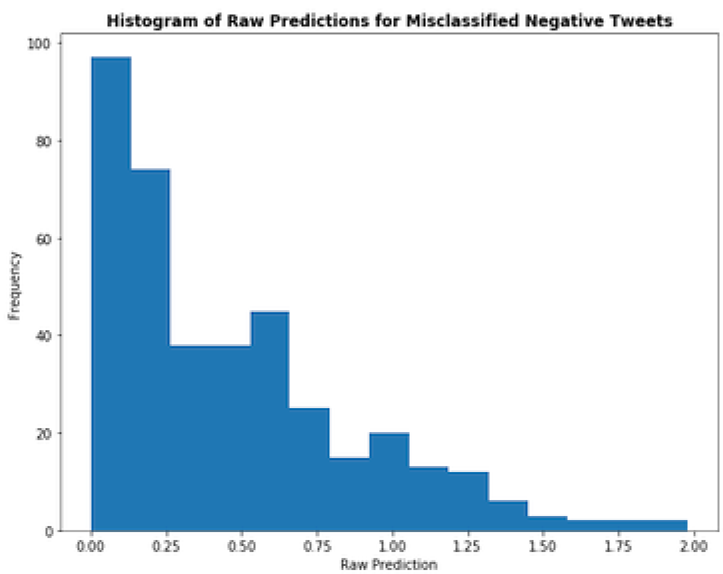

Figure 23: *Histogram of Raw Predictions for Misclassified Negative Tweets*

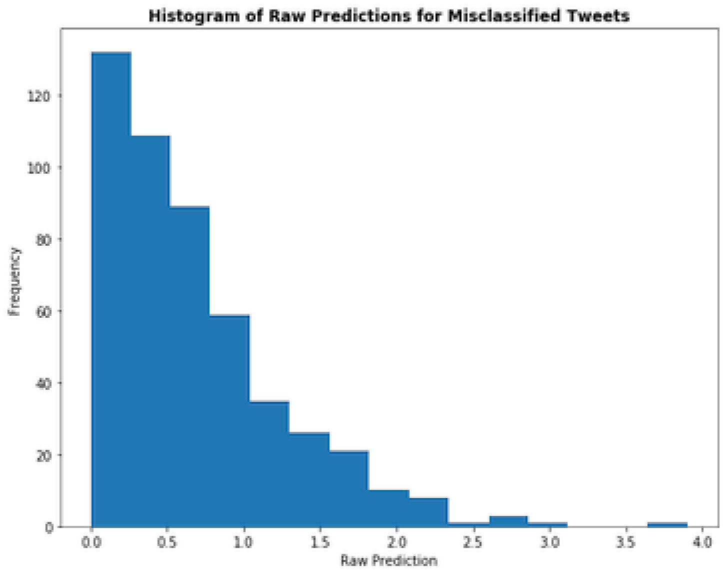

Figure 24: *Histogram of Raw Predictions for Misclassified Positive Tweets*