UNIVERSITY OF MICHIGAN LINGUISTICS

Using Sentiment Analysis to Train A Classifier to Predict Tweets: Results and Reflection

Alan Zhong

December 19, 2019

CONTENTS

1 Introduction

1.1 Big Data and its Implications on Customer Service

We live in an era where companies have an unprecedented amount of access to market and user data. Due to innovations in cloud and edge computing, data collection and analysis can occur more frequently and accurately in closer proximity to users. The development of these technologies alongside widespread adoption of social media has allowed companies to interact with their user-bases on a personal level previously unseen. Through social media, companies can gauge customer reactions to new products or policies in real time while simultaneously being able to make announcements en mass immediately. While many use-cases exist for how technology can affect company policy and efficiency, this report will focus on applying machine learning principals to bolster social media engagement; replacing the need for human-operated customer support.

1.2 Importance of Customer Support and Outreach in the Airline Industry

Due to the nature of the hospitality industry, customer feedback is extremely important. Specifically within the airline industry, customer feedback has historically been exceedingly negative. Passengers complain about a number of things involving every point of the airline-passenger user experience including long lines, delays, layovers, and bad food. Use-cases for machine learning in hospitality can span from resolving individual customer disputes to testing if a new internal workflow policy has any measurable impact. Because of the rise of new competitors, the airline industry remains volatile in terms of approval as seen in Fig 1.1.

Motivated by internal competition, the industry's bad reputation, and complicated operational workflows, airline companies are looking to machine learning and big data for solutions. In order to gauge the efficacy of integrating natural language processing into customer service, 14,640 tweets involving six major airlines were analyzed. Specifically the airlines included in the data set are: United, Virgin America, US Airways, America, Delta, and Southwest.

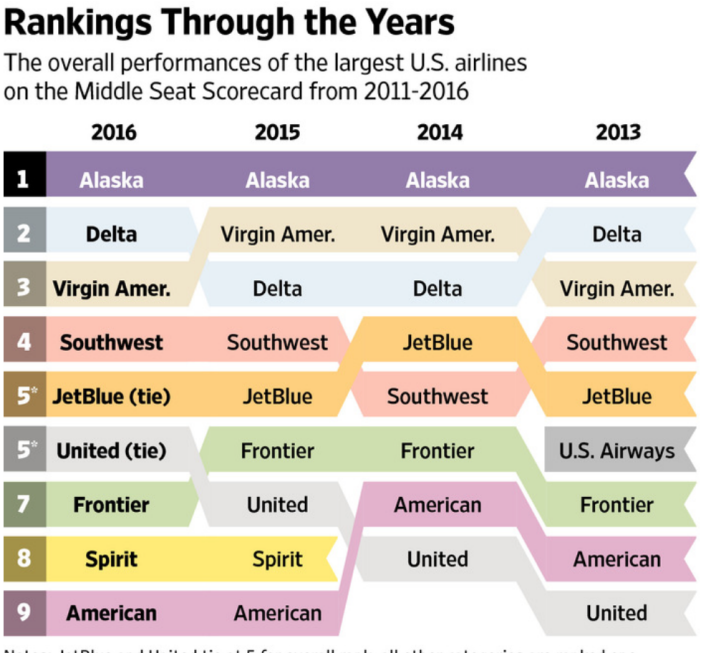

Figure 1.1: Rankings of Major Airlines Based on Customer Approval Between 2013 and 2016 (Pizzarello)

2 Materials

2.1 Twitter Data

With the goal of evaluating how effective natural language processing would be in regards to the airline industry, 14,640 Tweets involving six major airlines were collected. The twitter data can be found at this URL: *https://raw.githubusercontent.com/kolaveridi/kaggle-Twitter-US-Airline-Sentiment-/master/Tweets.csv*. Specifically the airlines included in the data set are: United, Virgin America, US Airways, America, Delta, and Southwest. IDs for these tweets were collected and hydrated using the Twitter API. Sentiment Analysis was performed on the tweets through the Text Analytics Azure API; isolating and labelling keywords with positive, neutral, or negative connotation such as long line, delay, good customer experience, or slow.

The Tweets were organized in a .csv file into a table with 15 columns: 'tweet id', 'airline sentiment','airline sentiment confidence', 'negative reason', 'negative reason confidence', 'airline', 'airline sentiment gold', 'name', 'negativereason gold', 'retweet count', 'text', 'tweet coord', 'tweet created','tweet location', and 'user timezone'. Some of these columns were taken directly from the hydrated tweets, such as 'tweet id' and 'tweet coord'. Others were

evaluated through the Azure Text Analytics API such as 'airline sentiment confidence' and 'negative reason'.

2.2 Analyzing the Tweets

Before the data is ready to be split into training and test sets, the tweets were analyzed. It was found that United, US Airways, and American had the greatest share of mentions as seen in Fig 2.1.

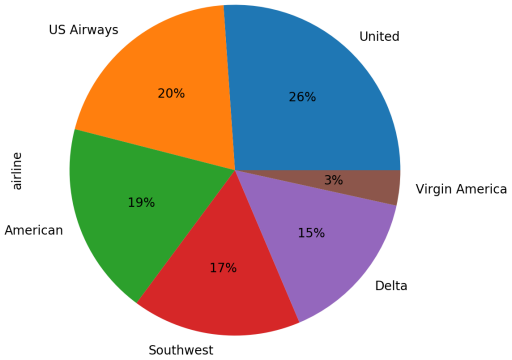

Figure 2.1: Share of Mentions by Airline

As hypothesized, the tweets were overwhelmingly negative as seen in Fig 2.2. Because of the larger sample size of the negative mentions, confidence levels for the negative tweets were the highest, followed by neutral then positive as seen in Fig 2.3. Finally the tweets concerning each airline were divided by sentiment level. As seen in Fig 2.4, the three most popular airlines: United, US Airways, and American had majority negative tweets. Concentrations of positive and neutral tweets in the three other airlines was higher, but the raw number of positive and neutral tweets remained constant.

Before our data can be used to train a classifier to predict sentiment, our tweets need to be processed.

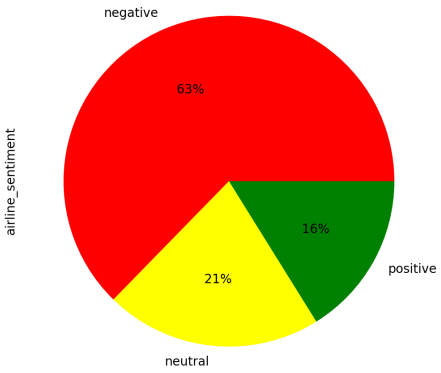

Figure 2.2: Pie Chart of Tweets by Sentiment

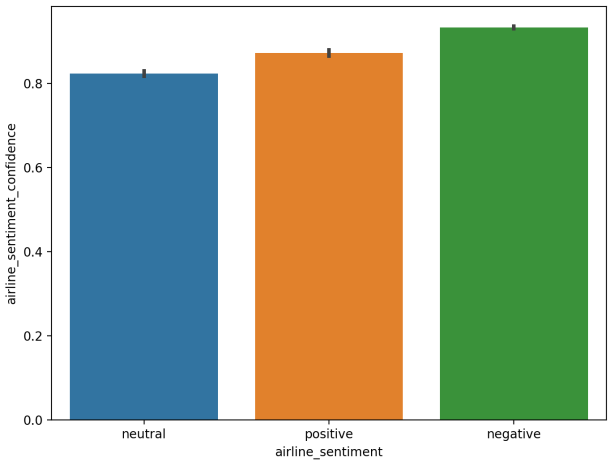

Figure 2.3: Confidence Levels by Sentiment

3 Procedure

3.1 Preprocessing

Due to the informal nature of tweeting, our tweets need to be cleaned of potential slang, punctuation marks, and mispelling. Before cleaning, the tweets were divided into feature and label sets using the panda iloc() method. The feature set will contain column 11 (index 10) of our data which contains 'tweet text'. Meanwhile our feature set will contain each

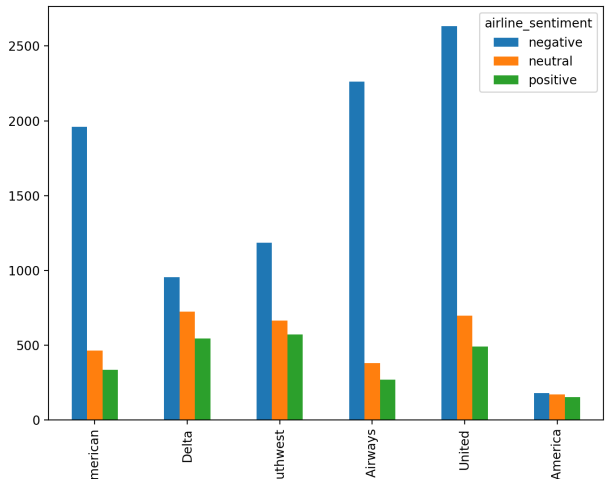

Figure 2.4: Division of Sentiment By Airline

tweet's corresponding sentiment. After separating the tweets in the .csv file into feature and label sets, the feature set was cleaned using regex, otherwise known as regular expressions. First all special characters such as commas, periods, and apostrophes were removed from the tweets. Now single characters, like those that would be left behind by removing an apostrophe, were removed. However this cleaning could result in there being extraneous spaces added to the tweets so those were removed as well. Finally all text was converted to lower-case.

3.2 Representing Text In Numeric Form

In order to train our classifier, our feature and label sets needed to be converted to numeric form to be compatible with statistical methods. Amongst three vectorization options: Bag of Words, TF-IDF, and Word2Vec, TF-IDF was chosen because of its emphasis on importance of a word across the entire document. TF-IDF uses two terms, term frequency which is frequency of a word across the entire document, and inverse document frequency which reveals how common or rare a given word is. Specifically:

$$td - idf(t,d,D) : tf(t,d) * idf(t,D)$$

$$tf(t,d) = \log(1 + frequency(t,d))$$

$$idf(t,D) = \log \frac{1}{count(d \in D : t \in d)}$$

Vectorization was performed using the class TfidfVectorizer in the ScikitLearn library.

3.3 Training the Classifier

After the feature and label sets were cleaned and vectorized, they were split into training and test sets. 80 percent of the data was put into training and 20 percent was put into test. The Random Forest Algorithm in the RandomForestClassifier class of the sklean.ensemble module was used to train the machine learning model. After using the fit() method to train the model, predict() was called in order to evaluate our model.

4 Results

4.1 Evaluating the Model

```
[[1723  108   39]
 [ 326  248   40]
 [ 132   58  254]]
              precision    recall  f1-score   support

    negative       0.79      0.92      0.85      1870
     neutral       0.60      0.40      0.48       614
    positive       0.76      0.57      0.65       444

    accuracy                           0.76      2928
   macro avg       0.72      0.63      0.66      2928
weighted avg       0.75      0.76      0.74      2928

0.7599043715846995
```

Figure 4.1: Output of predict() Method

Above is the output of the predict() method in terminal. Overall, our model correctly predicted the sentiment of 75.99 percent of the time. Predictions of negative sentiment were most accurate, being correctly predicted 79 percent of the time. Additionally 92 percent of the true negative tweets were predicted correctly which was much higher than that of the neutral and positive sets. Amongst the 2928 tweets in our test set, 1870 were negative, 614 were neutral, and 444 were positive.

4.2 Analysis

In order to raise our model's overall accuracy, the test set would need to be more uniform. The negative predictions were the most accurate, which is not surprising considering how many more negative tweets there were. Additionally, 92 percent of the true negative tweets were predicted correctly. If the distribution of the test set had been more uniform then the accuracy of neutral and positive predictions would improve. A different machine learning model could have been chosen as well. The Randomized Forest Algorithm was used because of its compatibility with non-normalized data (data scaled to be between 0 and 1). Other algorithms such as logistic regression or SVN may have provided better results. Additionally utilizing a decision tree algorithm would have resulted in more interpretable results.

5 Conclusion

Moving forward into the era of Big Data, companies will have many more tools and much more information to work with. Bolstered by social media, companies can interact with their customer bases to an unprecedented degree. Gradually, social media is overtaking traditional advertising. An example of Big Data remolding an industry is the airline industry.

Currently the average flying experience, from start to end, is perceived to be extremely negative. Part of this negative reception can be attributed to slow or poor customer service. However, these workflows and systems can be improved by using machine learning processes. With machine learning, social media complaints could be analyzed real-time and users could be directed to a solution quickly.

This report analyzed 14620 tweets concerning six major airlines: United, Virgin America, US Airways, America, Delta, and Southwest. These tweets were labelled with positive, neutral, and negative sentiment, cleaned, vectorized, and finally used to train a classifier. The machine learning algorithm we used was the Random Forest Algorithm.

Our model had overall 75.99 percent prediction accuracy with our test set of 2928 tweets. Due to the overwhelming majority of negative tweets, our model correctly predicted 92 percent of negative tweets. In order to improve results, the test set could have been resampled to include an equal distribution of sentiment. Additionally, a different machine learning model, such as logistic regression or SVN, could have been used to improve results.

Although our results were far from perfect, the potential of utilizing machine learning processes to improve customer service in the airline industry is clear. Internal workflows and systems could be streamlined using this technology and social media reactions could be analyzed real-time.

6 REFERENCES

Pizzarello, Edward. (2017, January 13). Who Are the Best and Worst Airlines of 2016?. https://pizzainmotion.boardingarea.com/2017/01/13/best-worst-airlines-2016/

Usman, Mali. (2019, April 03). Sentiment Analysis with Scikit-Learn. https://stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/