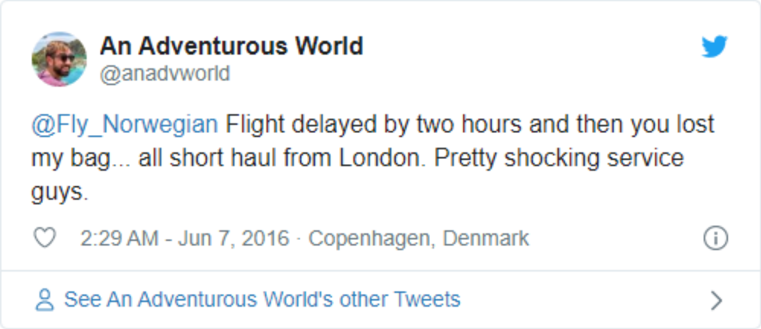
Airline Tweet Analysis
Support Vector Machines and Naïve Bayes for Sentiment Analysis
By: Teresa Cameron

Introduction

Commercial aviation has a large impact on the U.S. economy. More than 5% of the U.S. GDP growth and over 10 million jobs are created by this industry (Airlines for America 2021). The business decisions made by this industry have a far-reaching impact, so it is important for decision makers to have access to timely and comprehensive information.

Many people have strong opinions about the airline industry. Whether they are upset about the customer service, flight disruptions, or the gradually shrinking seating space, people willingly broadcast their sentiments on social media platforms. This data can be useful to gauge how customers feel towards the airline industry and the issues decision makers should allocate their resources. The following tweets provide an example of customers who use Twitter to highlight problems with service.

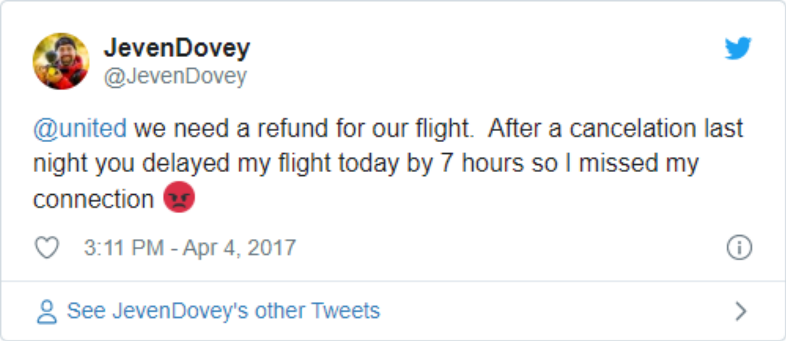


An Adventurous World
@anadvworld

@Fly_Norwegian Flight delayed by two hours and then you lost my bag... all short haul from London. Pretty shocking service guys.

♡ 2:29 AM - Jun 7, 2016 · Copenhagen, Denmark

See An Adventurous World's other Tweets



JevenDovey
@JevenDovey

@united we need a refund for our flight. After a cancelation last night you delayed my flight today by 7 hours so I missed my connection 😡

♡ 3:11 PM - Apr 4, 2017

See JevenDovey's other Tweets

If an airline company is able to quickly and accurately identify the sentiment of tweets, it could directly respond to users and solve their problems while retaining loyalty. By engaging in social listening or monitoring online conversations in social media, businesses can also remain informed of potential future issues, trends, and information regarding competitors.

It is extremely time consuming and costly for companies to hire individuals to review a large number of tweets. One method that can be used to save time and money is to use machine learning (ML). Machine learning is using computers to find patterns, such as sentiment, in large amounts of data. The ML models are trained using the text of tweets with sentiment labels and then used to predict the sentiment of a large number of unlabeled tweets. Being able to quickly capture and sort social media posts by sentiment would allow airline companies to better address customers' needs and allow decision makers to make more informed business decisions.

Analysis and Models

About the Data

The data consisted of a CSV file with 14,640 tweets labeled as positive, neutral, or negative. Figure 1 shows the sentiment distribution of the tweets.

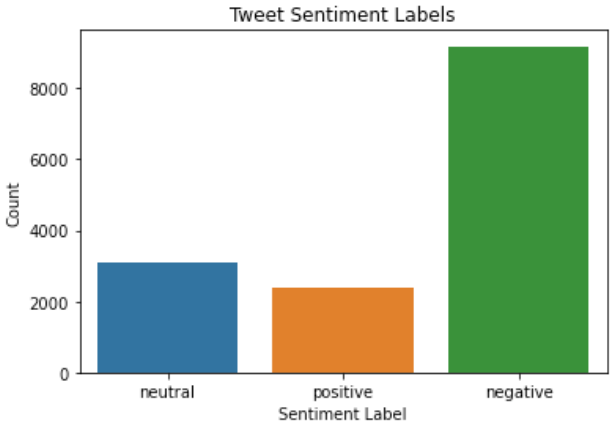


Figure 1

Considering there were significantly more negative tweets than neutral or positive, the models were trained on the original data and a subset of the data that had an equal amount of each label.

The data was vectorized three different ways. The first was a unigram frequency distribution, the second was a binary frequency distribution, and the third was a unigram tfidf. Next the data was divided into training and testing data with 60% of the data being used for training and 40% to be used for testing.

Models

The classifiers used were Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Support Vector Machines. The Multinomial Naïve Bayes classifier was used to create two models. One based on the values of the data being the frequency distribution and the other the values were tfidf.

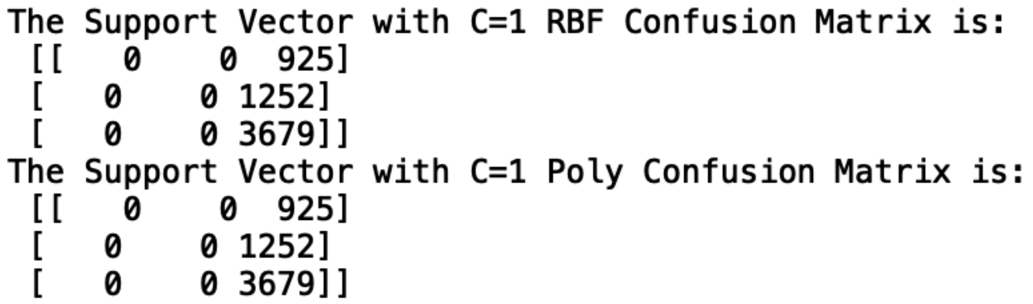
The Support Vector Machine classifier was run with the C at 1 and 10, for the kernels set to linear, rbf, and poly, and on the data values as term frequency and tfidf.

Results

For Naïve Bayes, the classifier with the highest accuracy score was Bernoulli at 77% with the original dataset. The models created with the data that had an equal distribution of labels produced slightly less accurate results.

For Support Vector Machine classifier, the one with the highest accuracy was the one with a linear kernel, c equal to 1, and trained on tfidf data at 77%. Increasing C decreased the accuracy of all models except RBF. Also, the accuracy of the models did not change using the TFIDF data or different values of C for the classifier using the poly kernel. Most of the other classifiers tied with an accuracy of 63%.

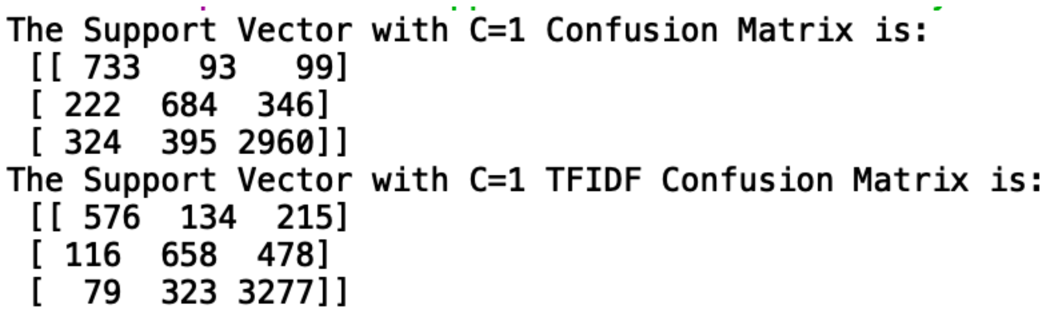
The reason for this can be seen in the confusion matrices. Figure 2 shows the confusion matrix of the support vector classifier with RBF and poly kernels.



```
The Support Vector with C=1 RBF Confusion Matrix is:
 [[   0    0  925]
 [   0    0 1252]
 [   0    0 3679]]
The Support Vector with C=1 Poly Confusion Matrix is:
 [[   0    0  925]
 [   0    0 1252]
 [   0    0 3679]]
```

Figure 2

The classifiers with an accuracy of 63% predicated all the testing data was negative. The confusion matrices for the support vector classifiers using C=1 and frequency distribution and tfidf data are shown in Figure 3.

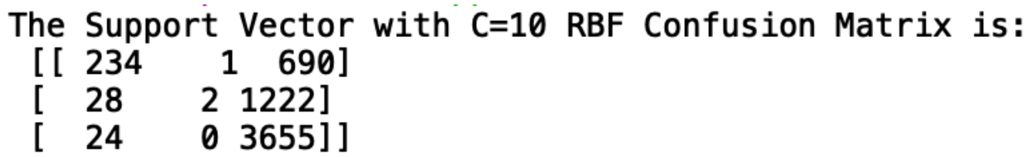


```
The Support Vector with C=1 Confusion Matrix is:
 [[ 733   93   99]
 [ 222  684  346]
 [ 324  395 2960]]
The Support Vector with C=1 TFIDF Confusion Matrix is:
 [[ 576  134  215]
 [ 116  658  478]
 [  79  323 3277]]
```

Figure 3

Although the first SVM classifier was better at categorizing the positive and neutral tweets, the second was better at correctly identifying negative tweets from those that were positive. It only misidentified 79 negative tweets as positive while the first classifier misidentified 324 in the same category.

Another interesting find was the SVM classifier that had a C of 10 and used the RBF kernel. Figure 4 shows the confusion matrix.

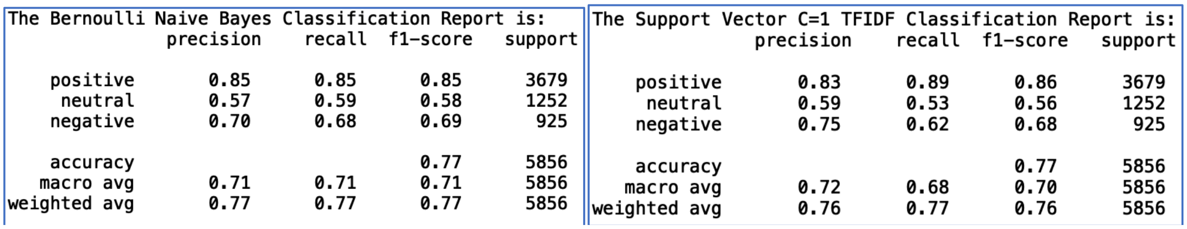


```
The Support Vector with C=10 RBF Confusion Matrix is:
 [[ 234    1  690]
  [  28    2 1222]
  [  24    0 3655]]
```

Figure 4

The model did not simply identify the testing data with one or two labels, but it incorrectly identified many of the text labeled as neutral. It was even more effective at identifying negative tweets from positive than the SVM using TFIDF data.

The Bernoulli Naïve Bayes and SVM with TFIDF data classifiers received very close accuracy scores. The Classification Reports can be reviewed to better understand these models. Figure 5 shows both Classification Reports.



```
The Bernoulli Naive Bayes Classification Report is:
              precision    recall  f1-score   support

    positive       0.85      0.85      0.85      3679
     neutral       0.57      0.59      0.58      1252
    negative       0.70      0.68      0.69       925

    accuracy                           0.77      5856
   macro avg       0.71      0.71      0.71      5856
weighted avg       0.77      0.77      0.77      5856
```

```
The Support Vector C=1 TFIDF Classification Report is:
              precision    recall  f1-score   support

    positive       0.83      0.89      0.86      3679
     neutral       0.59      0.53      0.56      1252
    negative       0.75      0.62      0.68       925

    accuracy                           0.77      5856
   macro avg       0.72      0.68      0.70      5856
weighted avg       0.76      0.77      0.76      5856
```

Figure 5

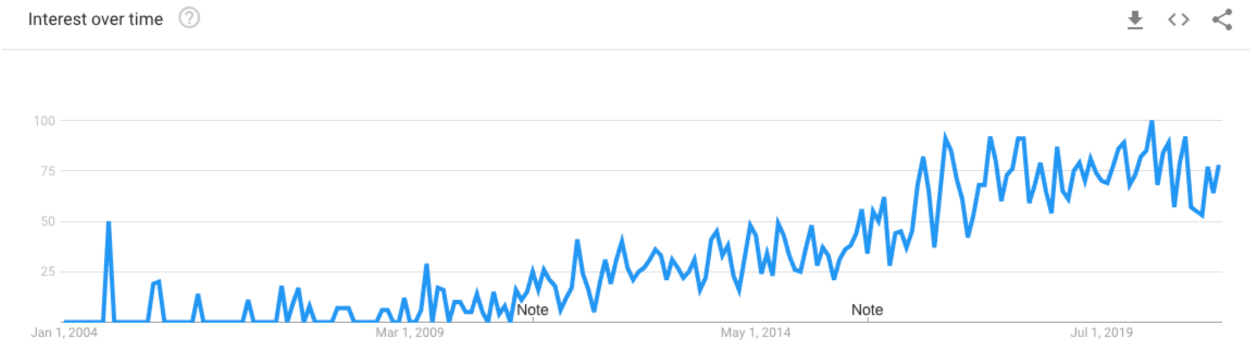
Since the recall for positive labels was higher in the SVM classifier, this model correctly identified more of the positive tweets. However, the Bernoulli Naïve Bayes correctly identified more of the negative text, which there were more of. This led to a slightly higher weighted average f1 score.

Conclusion

Social media allows businesses, and their customers to communicate more efficiently. Twitter users can post their questions and experiences using the "@" symbol to notify companies directly. The ability of a company to monitor, analyze, and respond to Tweets helps retain loyalty and improve brand equity.

With over 45,000 flights and 2.9 million airline passengers every day, US airline companies need to find effective ways of interacting with customers using social media (Air Traffic By the Numbers 2020). Sentiment analysis using ML models can assist companies with this task by quickly sorting tweets to determine the urgency of a response.

As shown by the Google Trends graph below, interest in sentiment analysis has increased over time. Understanding ML models that assist with sentiment analysis is a much-needed skill that will probably continue to increase with time.



References

Economic Impact of Commercial Aviation By State 2021, Airlines for America, accessed 1 April 2021, https://www.airlines.org/data/

Air Traffic By the Numbers, 21 September 2020, Federal Aviation Administration, accessed 1 April 2021, https://www.faa.gov/air_traffic/by_the_numbers/