**Twitter Sentiment and Stock Market Reactions:**

**Evidence from U.S. Airlines during Covid-19**

Team 5

Alan Shen (xs2441)

Haogeng Liu (hl3450)

Hongyi Zhou (hz2700)

Xingchao Gao (xg2347)

December 17, 2021

# Abstract: Paragraph synopsis of research

XXX

# Introduction

In behavioral economics, investor sentiment has frequently been considered an important factor in determining asset prices because it reflects the expectation of market participants on future returns and investment risk (De Long et al., 1990). Traditionally, sentiment is measured by observing analyst estimates, survey data, news stories, and technical indicators such as put/call ratios and relative strength indicators. However, these indicators only reflect a limited subset of investors and usually they are measured untimely. With the rise of the Internet, individual investors are increasingly relying on each other as peer-to-peer sources of information. The biggest revolution in the dissemination of information on the Internet has been the advent of social media platforms such as Twitter, which allow users to post their views about stocks instantaneously to a broad audience. Prior studies have examined the role of social media in capital markets. They find that Twitter sentiment has a significant impact on trading volume (Duz Tan and Tas, 2020) and abnormal returns (Gu and Kurov, 2020).

The main challenge faced by researchers who investigate the impact of firm-level Twitter sentiment on financial markets is that discovering and analyzing all relevant tweets is difficult, if not impossible. Existing studies mainly examine tweets around specific events and perform textual analysis using existing word classifications and the relative frequency of terms “bullish” and “bearish” in tweets to measure investor sentiment. Although straightforward, lexicon-based methodologies are subject to significant measurement errors (Loughran and McDonald 2011). This is especially true for the analysis of posts on social media. The content of Twitter messages is likely to be quite different from that of newspapers and corporate documents.

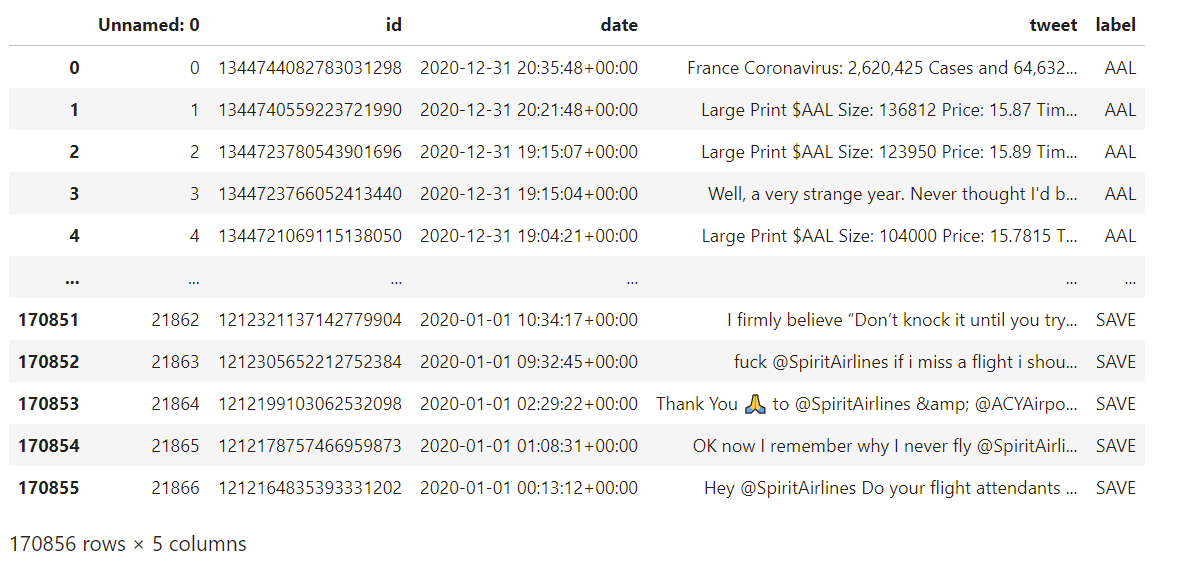
To further understand the effectiveness of using Twitter sentiment analysis as an indicator of stock market reactions, we focus on U.S. airline companies during Covid-19. The airlines have remained one of the hardest-hit global industries since the pandemic. The frequent changes in travel policy and crisis outlook have greatly affected investor sentiment and stock performance during the time, which provide more variations for our analysis.

# Data

For this project, we are focusing on the following six U.S. airline companies:

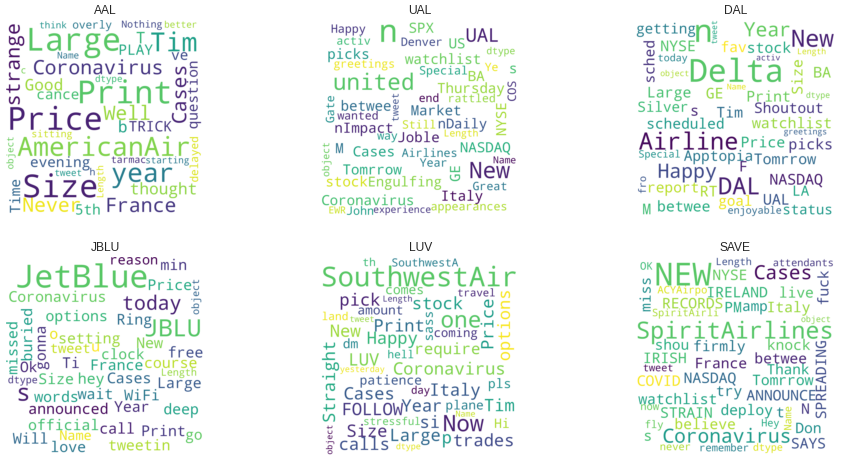
* NYSE: DAL, Delta Air Lines
* NASDAQ: UAL, United Airlines Holdings
* NASDAQ: AAL, American Airlines Group
* NYSE: LUV, Southwest Airlines
* NASDAQ: JBLU, JetBlue Airways
* NYSE: SAVE, Spirit Airlines

Using Twitter API, we extracted 170,856 Tweets that mention these six U.S. airline companies from January 1st , 2020 to December 31st, 2020.



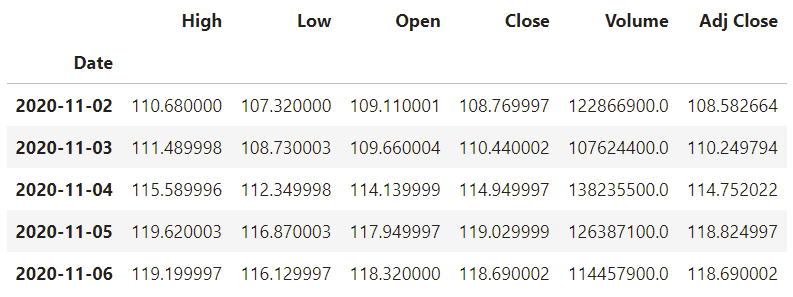
To increase the effectiveness of our NLP learning algorithm, it is important to clean the tweets before using them for analysis. The preprocessing step eliminates the following:

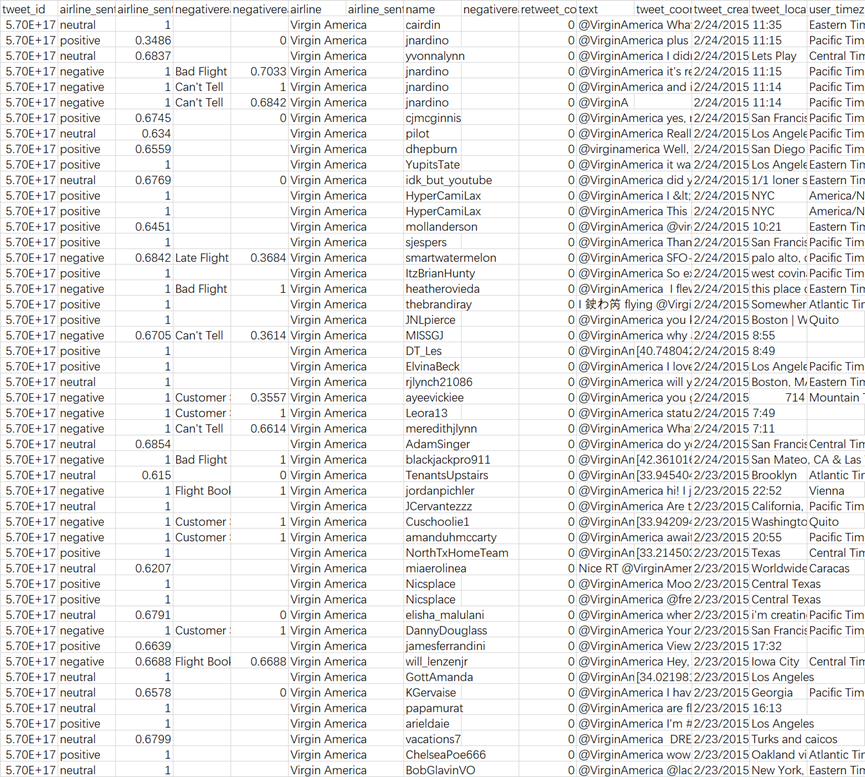
* stopwords (i.e., 'the', 'a', 'and', 'how', etc.);
* tweets with a question mark - questions are not good indicators of sentiment;
* non-letters (including: numbers, emojis, and punctuation);
* URL links;
* accounts referenced using '@' symbol;
* words less than 3 characters;
* retweets - we decided not to count the same tweet twice;



We create word clouds for tweets related to each airline to visualize what the most common words are. Not surprisingly, *“coronavirus”* appears in every airline’s word cloud. And we can see that words expressing feelings, like *“happy”*, *“strange”* and *“nothing better”*, are very common. Also, twitter users talk a lot about money and stock, mentioning words like “stock”, “NASDAQ” and “price”.

We also extracted historical stock price data of each of these six airlines during the same time interval from Yahoo Finance. Yahoo Finance provides data for stock price fluctuations for every company and every day.



To train our model to classify tweets into three sentiment categories: positive, neutral, and negative, we use a dataset from *kaggle (https://www.kaggle.com/benhamner/exploring-airline-twitter-sentiment-data)* as our training set. This data includes 14.6k tweets about 6 major US airlines (American, Delta, Southwest, United, US Airways, Virgin America) in 2015. The author of the data set extracts the sentiment from the tweet and labels each tweet into three sentiment categories: positive, neutral, and negative. Besides, what a negative tweet was disappointing about is also classified.

# Methodology

We try to train a model to classify tweets into three categories: positive, neutral, and negative.

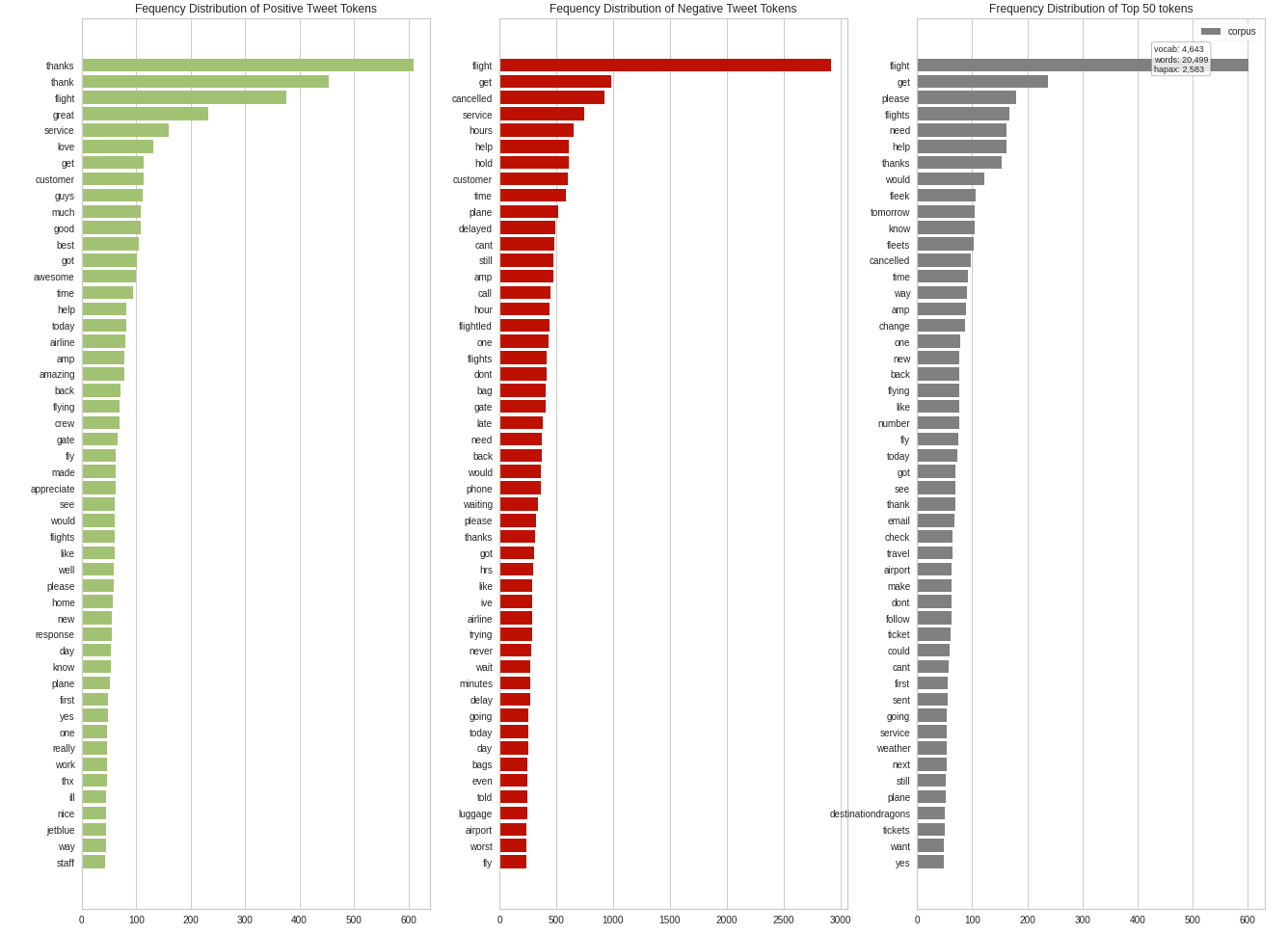
**Text Analysis For Training Data**

First, we calculate the distribution of positive, neutral and negative tweets for each airline company.



We can see from this breakdown graph that for most companies, negative tweets account for the largest proportion.

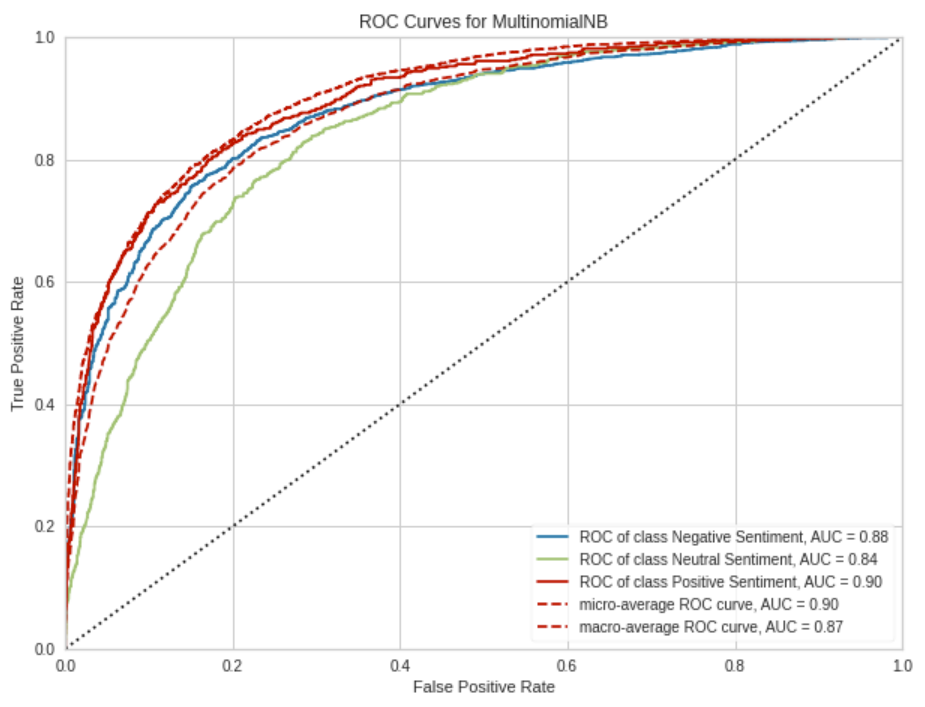
Then, we distinguish the types of words that are most frequently used in positive, negative, and neutral tweets.



**Model Training**

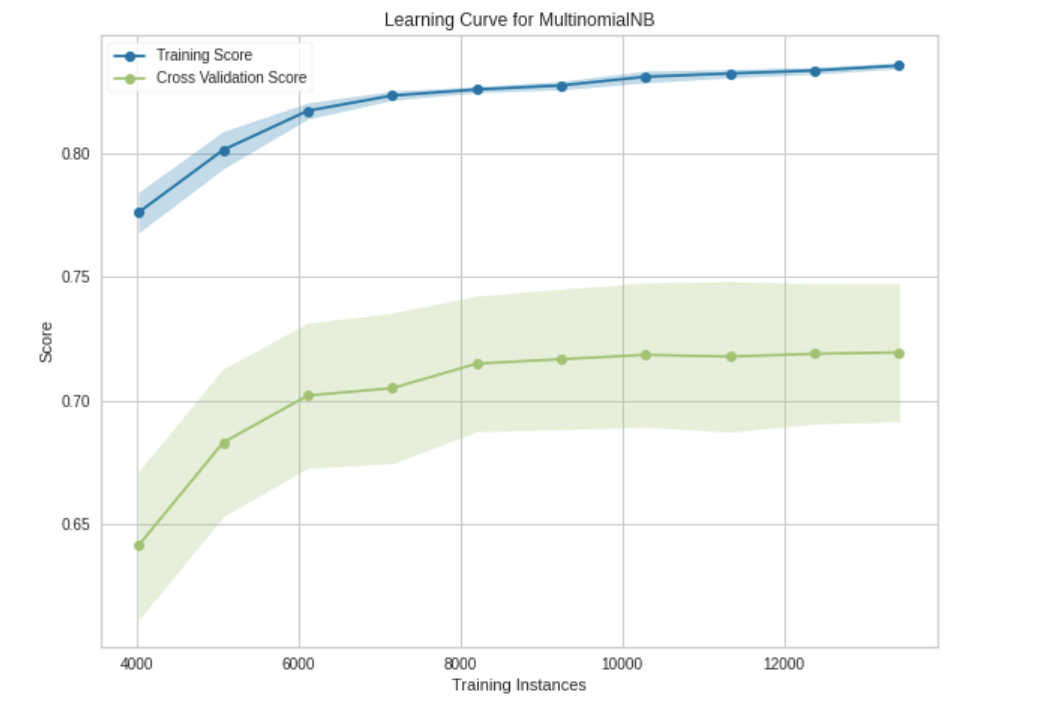
We firstly vectorize the training data using ngram. Then we select the multinomial Naive Bayes Classifier as our model because it is suitable for classification with discrete features. Also, since we did not have that much training data, NB models can be trained faster than other models.

**Testing the Accuracy of the Model**

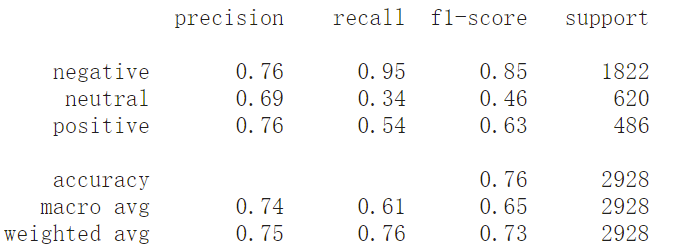
We use a ROC-AUC curve for performance measurement. A ROC-AUC (Receiver Operating Characteristic/Area Under the Curve) plot allows us to visualize the tradeoff between the classifier’s sensitivity and specificity. ROC (Receiver Operating Characteristic) shows the performance of classification at all classification thresholds (positive, neutral, and negative). The ROC is a computation of the relationship between true positives and false positives. AUC (Area Under the Curve) represents the degree or measure of separability (i.e. how well is the model at distinguishing the various classes). The higher the AUC, the better the model typically is. Our ROC curves indicate this model is better at classifying positive tweets than negative and neutral tweets. Overall, the macro avg AUC of 0.88 is a good indication the model is working.

**Visualizing the Classification Model (F1-Score)**

To visualize the learning curve of our classification model, we plot how the f-score changes over many training iterations. We can see that the training and test scores have not yet converged, but the test data seems to have converged at an F1 score of 0.73. This likely means the model's accuracy would not significantly change with more data. This model also primarily suffers from error due to variance (the CV scores for the test data are more variable than for training data) so the model may be overfitting the data.

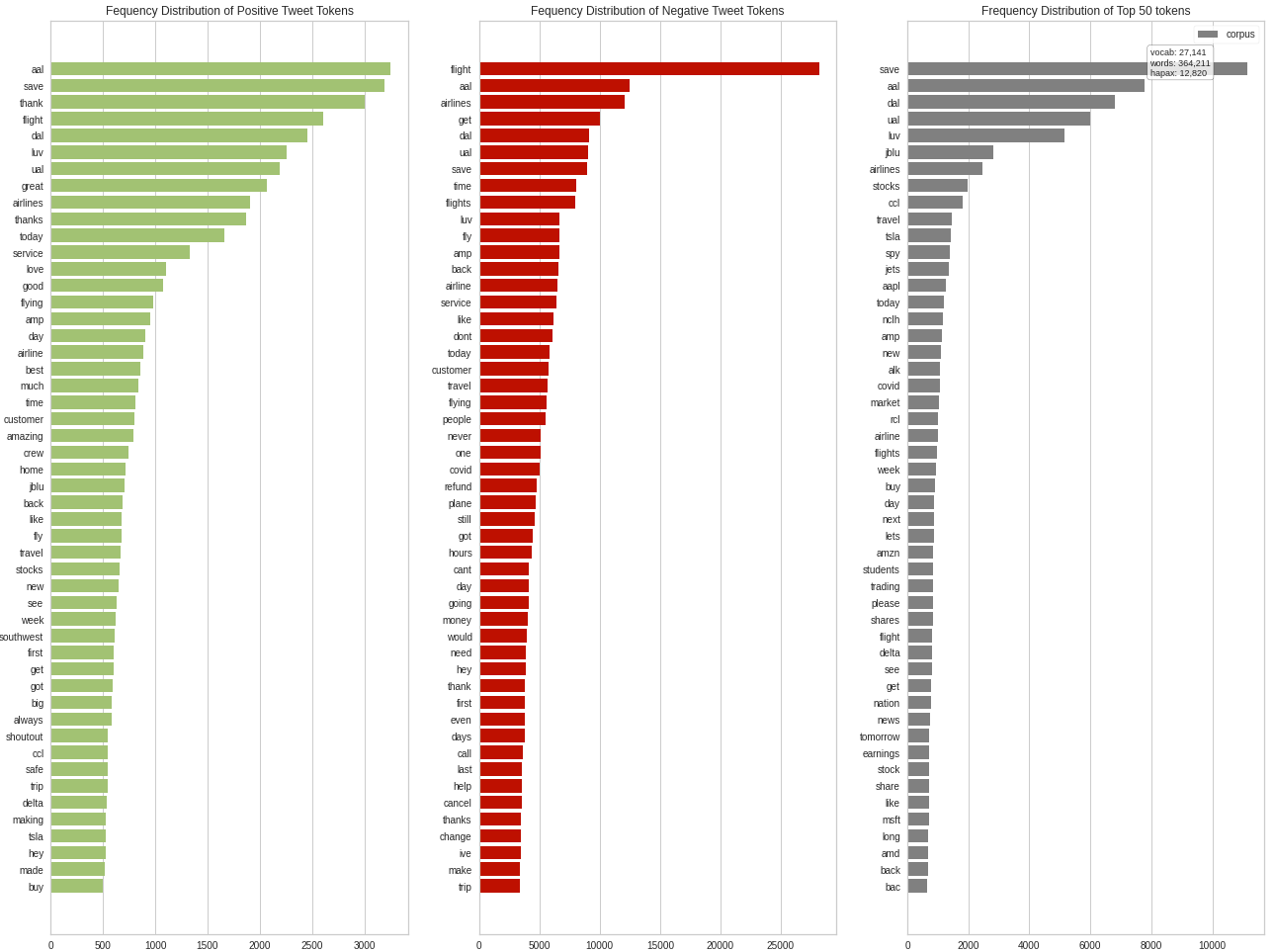


We can also measure the Precision, Recall, and F1-Score. These metrics are important because we know our dataset is imbalanced (there is not an equal number of positive and negative tweets). We do the following calculations:



**Text Analysis for Predicted Labels**

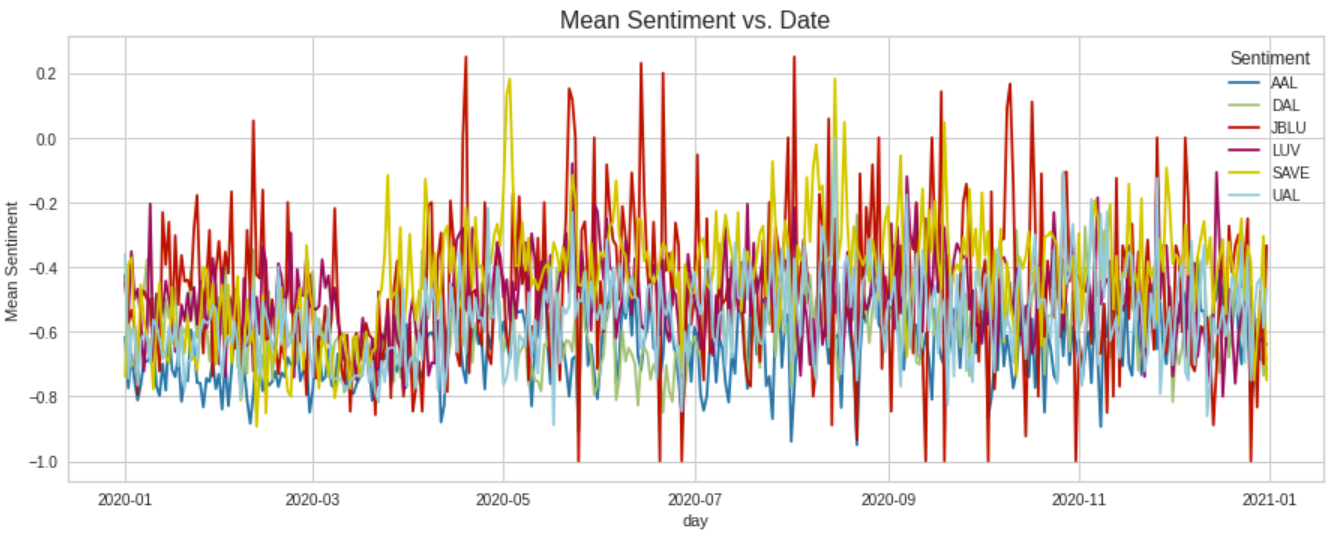
Then we retrain our model using all the data and predict unseen data. We use the same text analysis strategy to gauge the types of words used based on the predicted labels.

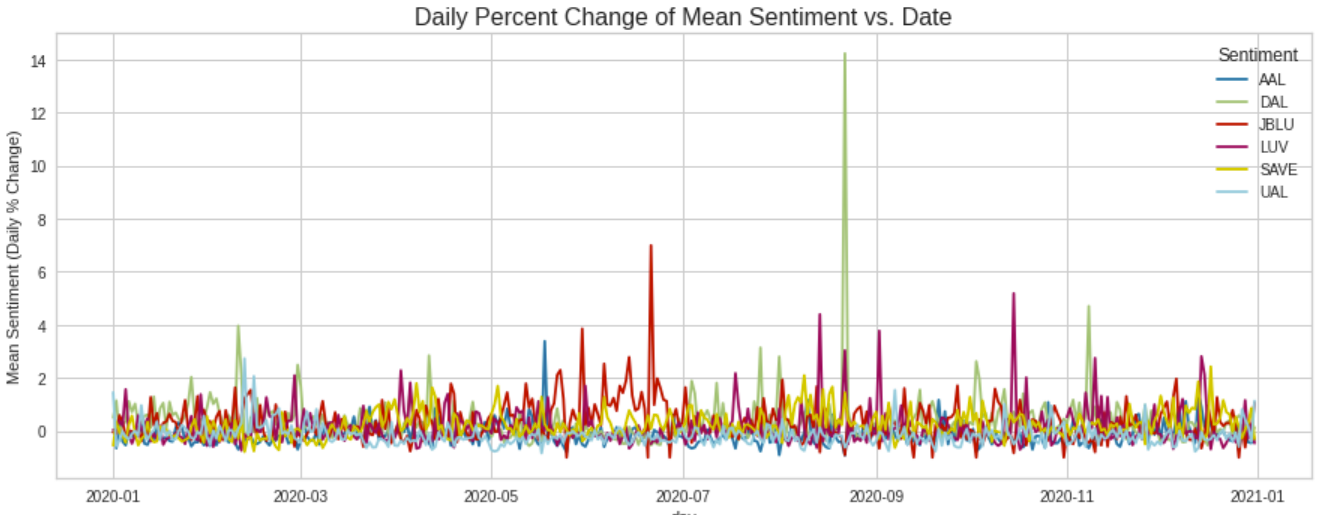


**Twitter Sentiment vs. Stock Price**

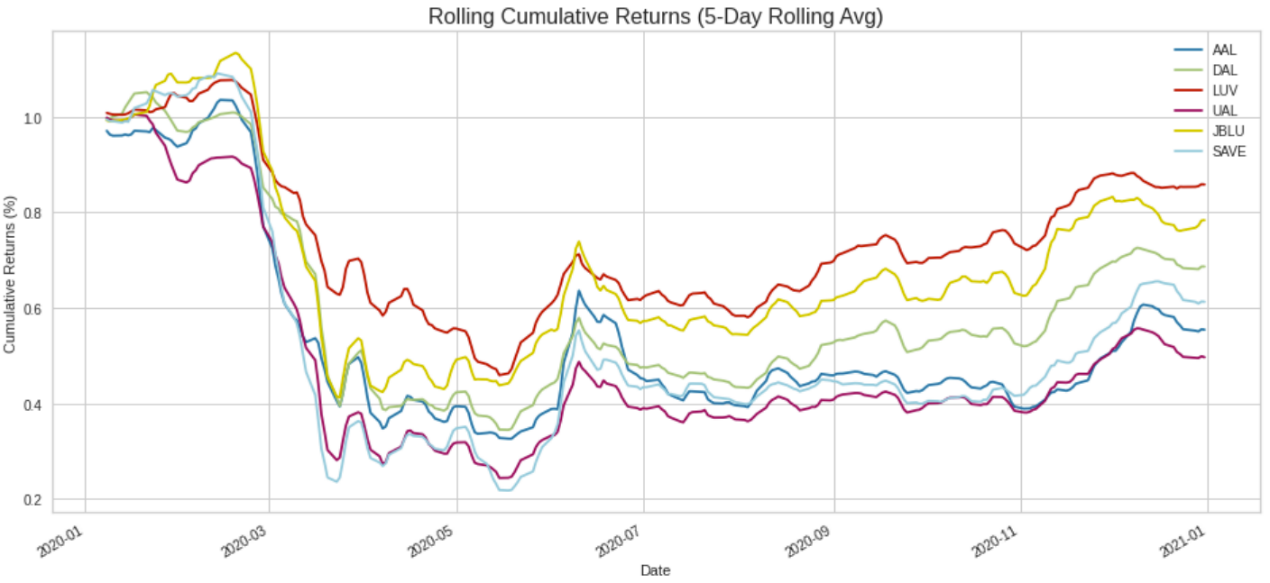
To gauge how Twitter feels about each airline, we take the mean of the predicted sentiment classification. Many of these tweets contain feedback from customers' flight experience. We can see that American Airlines has the worst average sentiment score by a large margin.

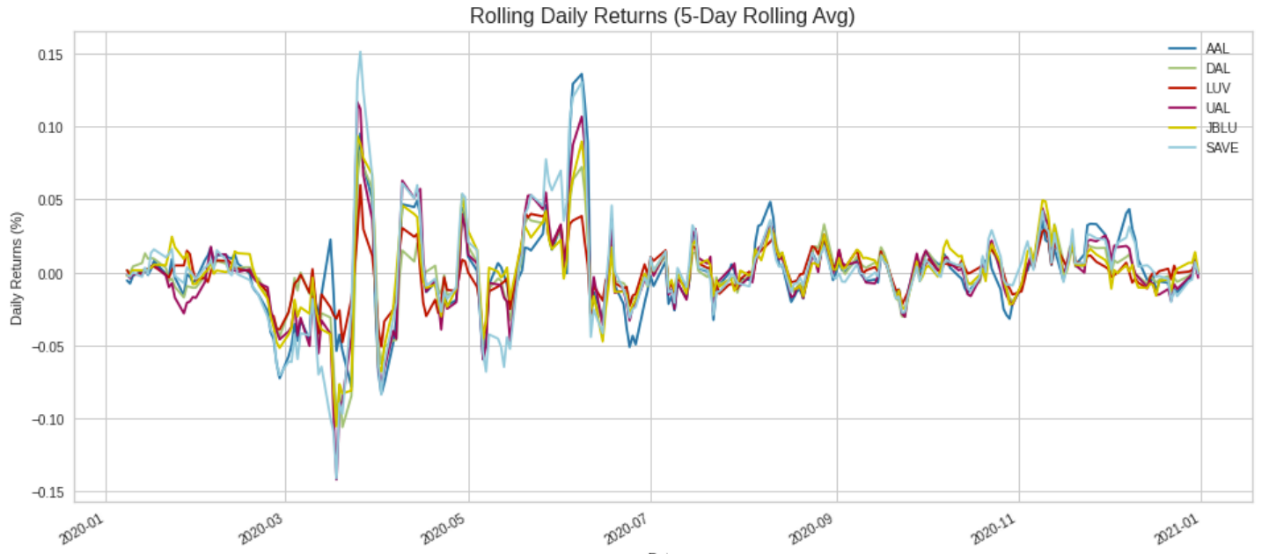
We plot out the change in sentiment each day to gauge trends over time. Ideally, we would have more data so that we could take the rolling average of the calculated sentiment scores (this would reduce the noise we see here). The first graph uses the average sentiment score per day (between [-1, 1]), and the second graph displays the average sentiment's daily percent change.





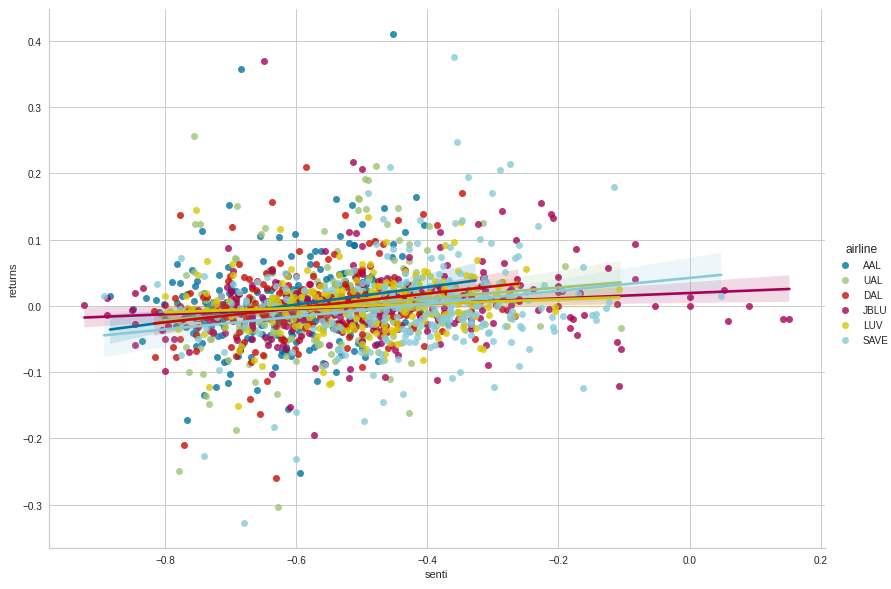
To see how the stocks have reacted over the same period, we calculate daily returns and cumulative returns. We use the 5-day rolling average of each to eliminate some of the market noise. Since both of these graphs are represented by percentages, we can conveniently relate price performance to the sentiment graph.

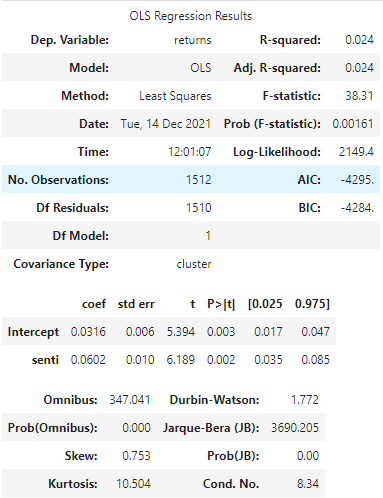


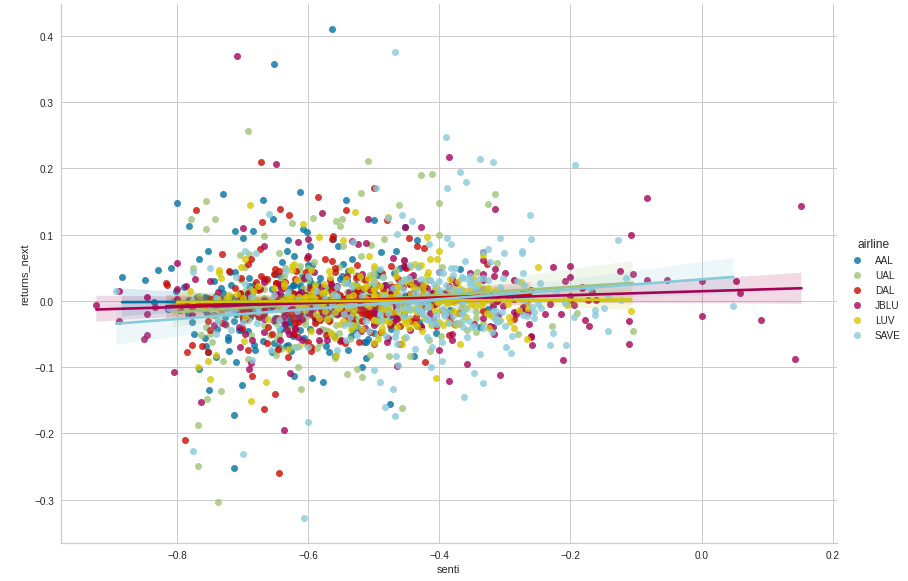


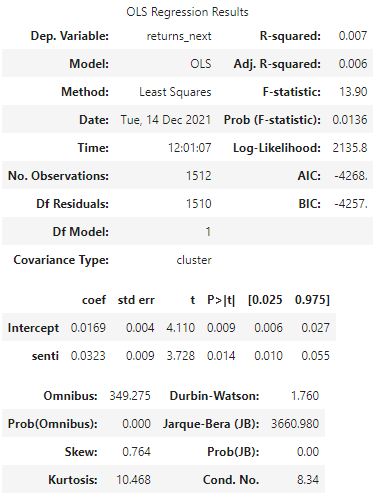
But to see a more detailed relation between sentiment and stock price, we need to do regression.

# Results

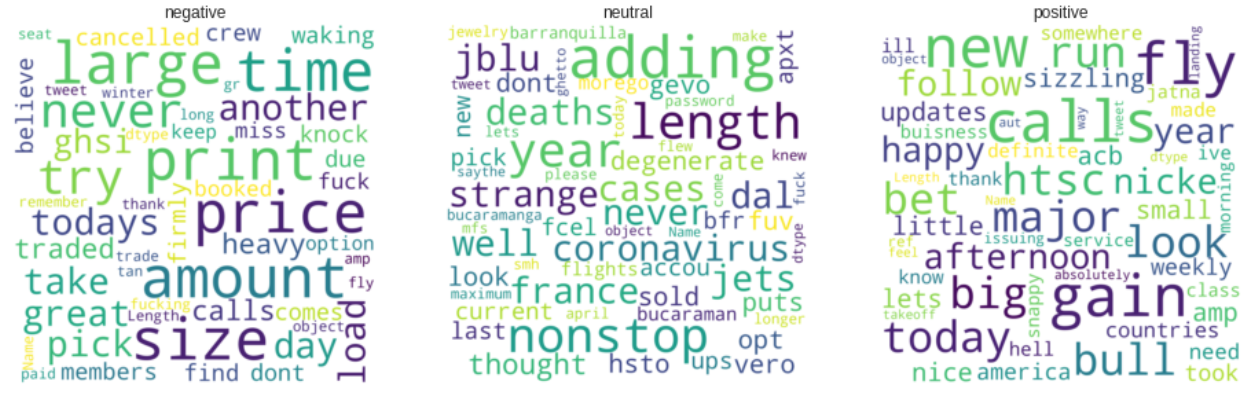
Looking at the graph for mean sentiment score per day, we can see that American Airlines consistently had the worst sentiment scores while Spirit Airlines seem to have the best sentiment scores overall. Meanwhile, Twitter sentiment for JetBlue appears to be the most volatile. The 5 day rolling average of cumulative returns for the airline stock prices shows that Southwest Airlines performed better than the others in the same period. Notably, we can see that there is a significant dip in stock prices which coincides with the lowest average sentiment scores during the period from March to June of 2020. Although this might give some indication that Twitter sentiment affected the stock prices of these airlines, we cannot be completely certain due to the lack of data. In an ideal situation, we would have years worth of data to generate rolling averages for both sentiment and stock price, which would eliminate the noise and result in a clearer association. 

The line plot and scatter plot generated by the OLS regression model indicates that there is a positive correlation between sentiment and rolling cumulative returns of airline stock price. From the model summary we can see that the t-statistic is around 6.19 which is relatively large and the corresponding p-value is 0.002 which is less than 0.005. Therefore, we can be 98% confident in rejecting the null hypothesis that there is no correlation between Twitter sentiment and airline stock prices. 





Similarly, the regression plot of the rolling daily returns and sentiment also shows a positive correlation, albeit less obvious than the previous model. Summary statistics show that a t-statistic of around 3.73 with a corresponding p-value of 0.014. From these models we can conclude that positive and negative sentiment on Twitter has a minor but tangible effect on the stock performance of airlines during COVID-19.

Additionally, we have created word clouds generated from sentiment data separated by “positive”, “negative”, and “neutral”. Most frequently used words that are associated with negative emotions include “price”, “time”, “miss”, “never”, and “cancelled”, which are self-explanatory and expected from typical complaints from airline passengers. On the other hand, words associated with positive emotions include “new”, “gain”, “thank”, “bull”, which can be interpreted as passengers being satisfied with the airline services and being optimistic about stock prices.

# Conclusion

The dramatic increase in the use of social media these past few years had a significant impact on the capital market. Firms use social media to communicate with their investor base and, increasingly, individual investors use social media to share information and insights about the prospects of stocks. Using firm-specific Twitter data and supervised-learning NLP models, we investigate the correlation between Twitter sentiment and stock market reactions, and test the predictability of Twitter sentiment for future returns. Our results show that Twitter sentiment is highly correlated with daily stock returns. Moreover, although most Twitter users are believed to be uninformed, our findings show that the daily firm-level Twitter sentiment contains information that is useful for predicting next day stock returns.

However, it is worth noting that there are several caveats in our analyses. Firstly, our dataset does not split financial information and customer reviews, though they are usually correlated at some level. The Twitter reviews only reflect a subsample of market participants and overestimate the sentiment of investors who use Twitter more frequently. Secondly, our NLP models rely on the labeled training sample of Apple, which might be systematically different from the Tweets of airline companies. The sample bias may decrease the accuracy of the sentiment calculation. Finally, our current study does not split official Twitter accounts (e.g., financial media) and personal accounts. There might be some heterogeneity among different accounts, and we are interested in what kind of users are more influential in influencing the stock prices. We believe financial Twitter accounts may indicate previous market movements, but consumer data may be more indicative of future price movements. For example, consumers fall in love or out of love with a product or service before markets can price this shift in consumer demand accordingly. All these remain areas for future research.

Our findings have important implications for market participants. The predictive power of Twitter sentiment for stock returns gives traders a strong incentive to analyze social media content carefully. Such analysis may be useful in making security selection decisions. Our results are also consistent with the notion that business entities may be able to improve transparency and market efficiency by using social media.

# References

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. Journal of political Economy, 98(4), 703-738.

Duz Tan, S., & Tas, O. (2021). Social media sentiment in international stock returns and trading activity. Journal of Behavioral Finance, 22(2), 221-234.

Gu, C., & Kurov, A. (2020). Informational role of social media: Evidence from Twitter sentiment. Journal of Banking & Finance, 121, 105969.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10‐Ks. The Journal of finance, 66(1), 35-65.