# Twitter Sentiment Analysis

Paris Vu
Springboard
February 2023

# Introduction

Stock has always been one of the main assets among institutions and it has grown even more popular among Millenials and Gen Z in recent times due to the redefinition of the money term and its price fluctuation. Moreover, scientists and analysts have gradually recognized social media's predictive power for the financial market, especially Twitter where key opinion leaders feel comfortable expressing their opinions.

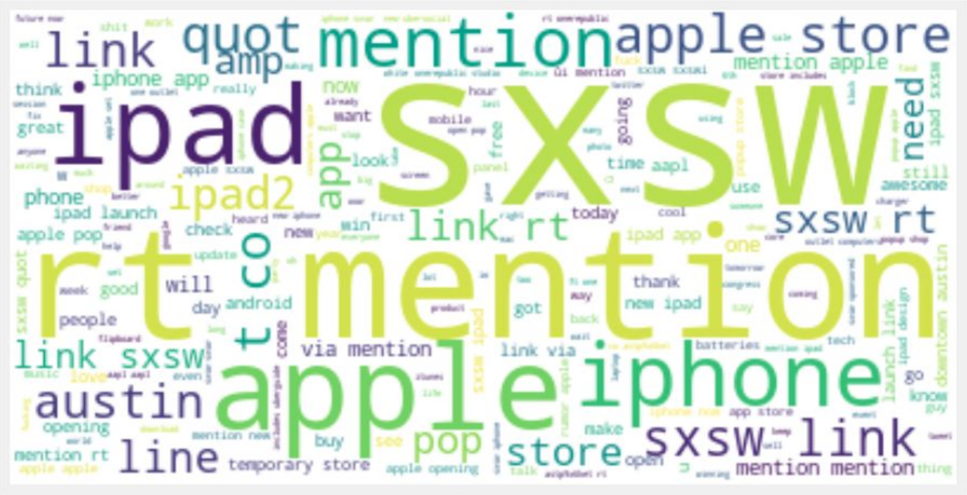
There is evidence to suggest that there can be a correlation between tweets and stock prices, particularly for publicly traded companies. The idea behind this correlation is that tweets can influence public perception, which in turn can impact stock prices.

**What opportunities exist for retail investors to develop a machine learning model that could measure sentiments of a stock using recently published Twitter's public opinion?**

# Historical Data: December 2014

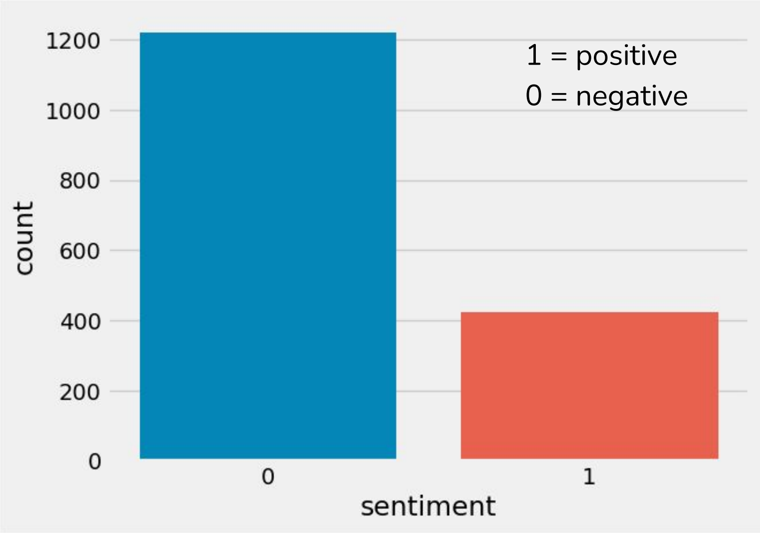
Before diving into live tweets scraping, I first explored two pre-scraped dataset (December 2014). Both dataset combined together produced a word cloud that suggests tweets are about 3 months ahead of real-life events.

One common 'big' word that appeared in both of these dataset is **SXSW**. After doing a simple google search, SXSW is a music festival that is usually held in mid-March. It is surprising to see that both the first dataset (dates in December 2014) and second dataset (unknown dates) mentioned this festival. This implies that **tweets might be 3 months ahead of real life events!**



By plotting the values count of the sentiments on these tweets from 2014, we can see that there were three times the amount of negative tweets to the amount of positive tweets. This result was not surprising since historical records show that there were a mix of positive and negative indicators in the global economy. For instances, some economic highlights at the time were:

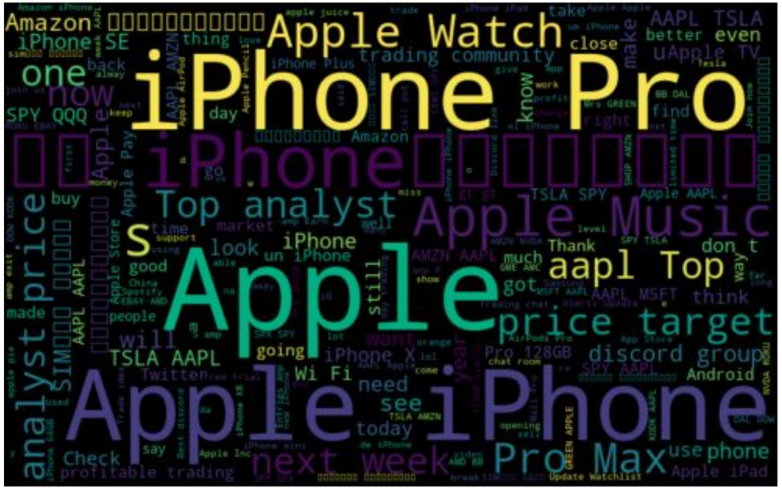
1. The US economy continued to show signs of improvement, with the unemployment rate falling to 5.6%, its lowest level since 2008. This was seen as a positive sign for the US economy.
2. The European Central Bank (ECB) announced a new stimulus program to help boost the Eurozone economy, which had been struggling with low inflation and weak growth. This program included buying government bonds, known as quantitative easing.
3. Falling oil prices continued to put pressure on oil-producing countries, such as Russia and Venezuela, which experienced economic difficulties as a result.
4. The Chinese economy showed signs of slowing down, which led to concerns about the impact on global growth.



# Live tweets from December 2022

Scraped live Tweets with specific hashtags (#appl, #apple, #ipad, & #iphone) using the SNScrape method and we found that **December 2022 was a positive month in terms of Twitter public opinion!**

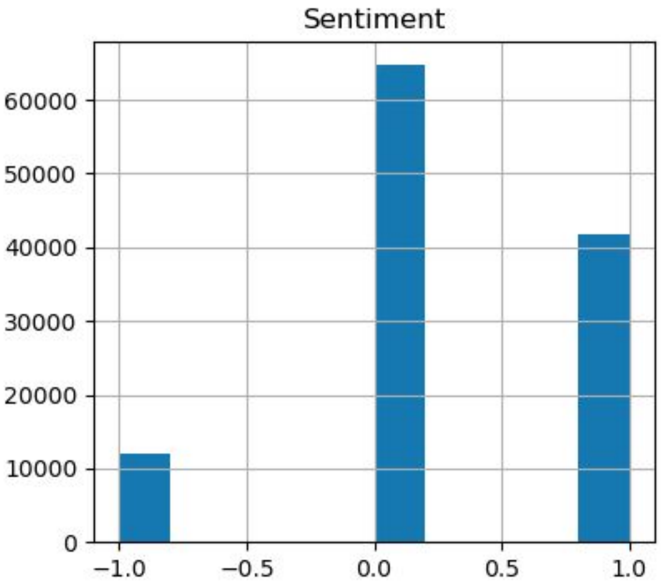
Firstly, our assumption that Tweets were 3 months ahead of real life events was false. Word cloud for December 2022 tweets did not return similar result to December 2014 word cloud (SXSW music festival).

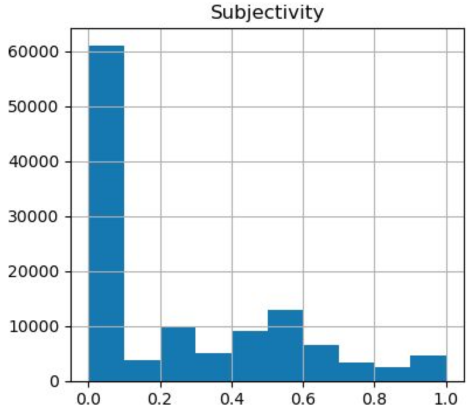


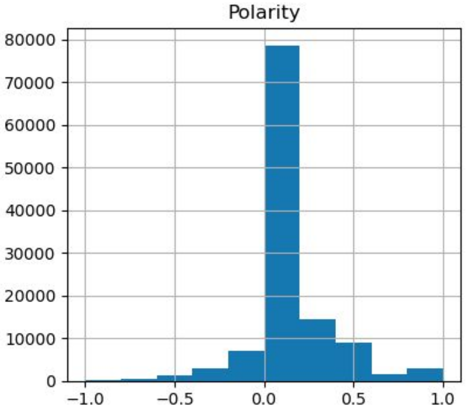
Extracted polarity & subjectivity scores.

- Subjectivity detection and polarity detection are subtasks under sentiment analysis. Subjectivity detection aims to remove 'factual' or 'neutral' content, i.e., objective text that does not contain any opinion.
- Polarity detection aims to differentiate the opinion into 'positive' and 'negative'.

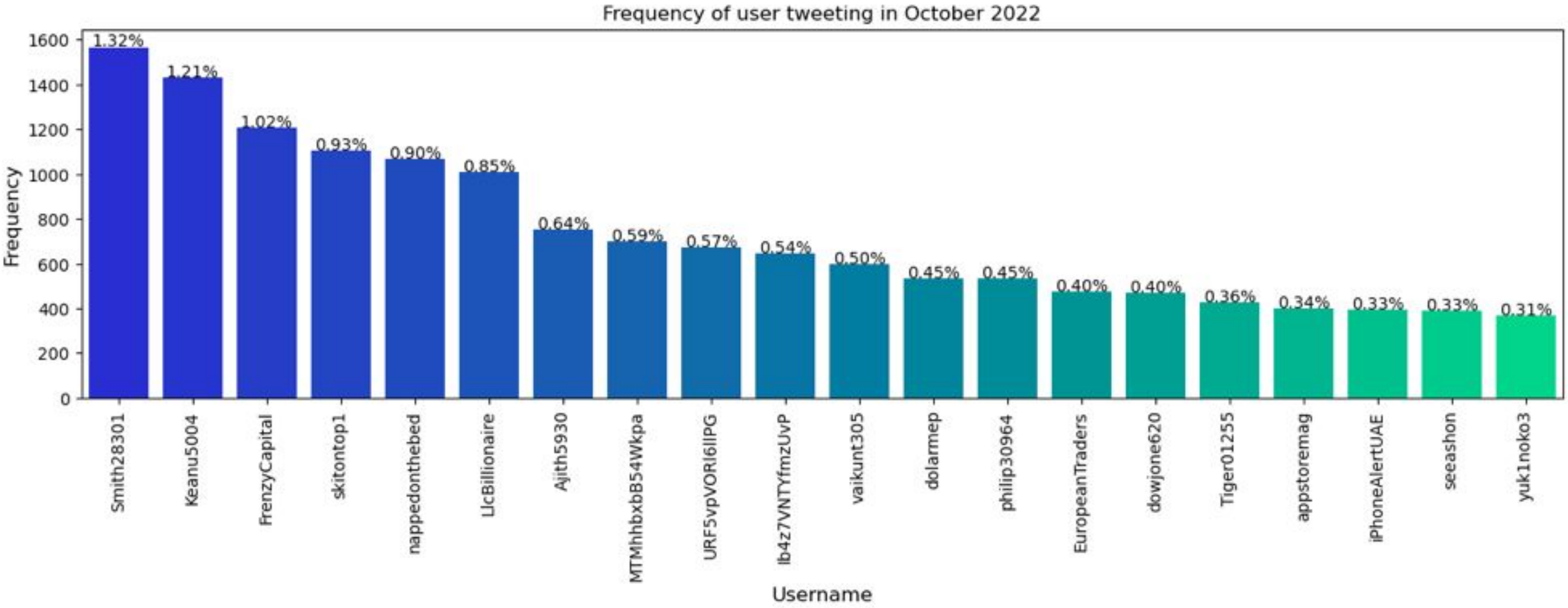
After having their polarity scores, I simplified it into numbers, called Sentiment numbers, representing the following: -1 = Negative (if polarity <0); 0 = Neutral (if polarity = 0); 1 = Positive (if polarity > 0). The plots below represent the ranges of Subjectivity, Polarity, and Sentiment scores.





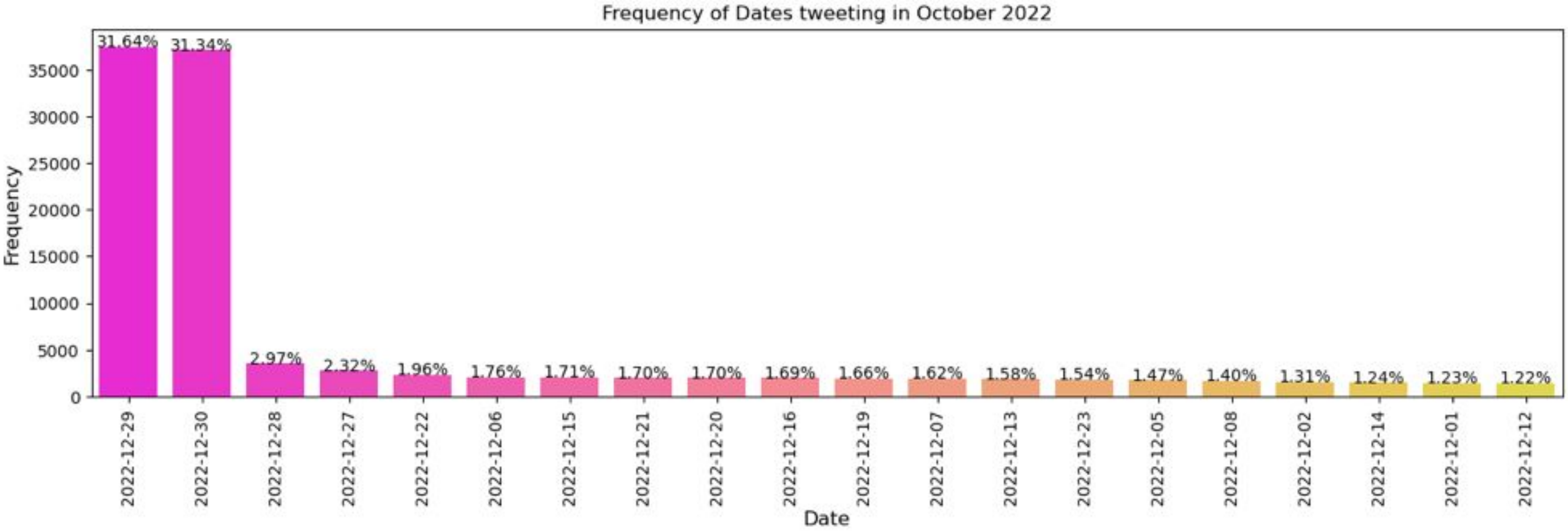


Both “Polarity” and “Sentiment” plots are in agreement that there are more positive opinions among the scraped tweets.

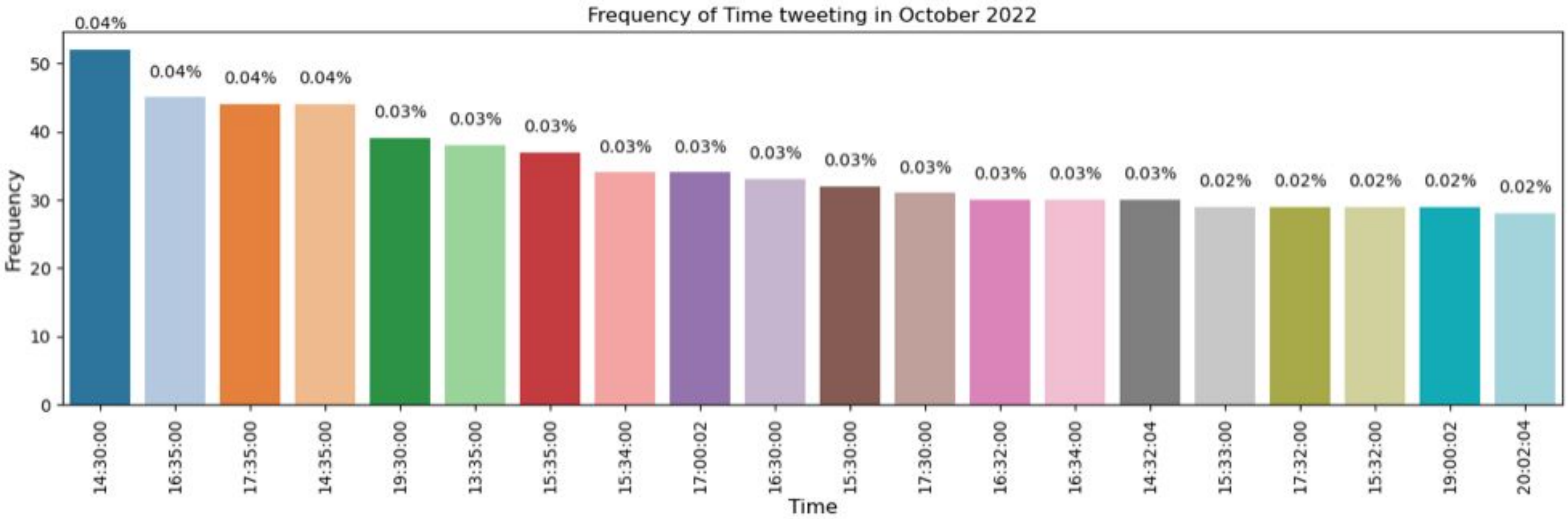


**Note:** The following illustrations have October in their titles, but they were actually for December 2022.

In December 2022, user Smith28301 posted most frequently (1.32%) using the following hashtags: apple, iphone, ipad.

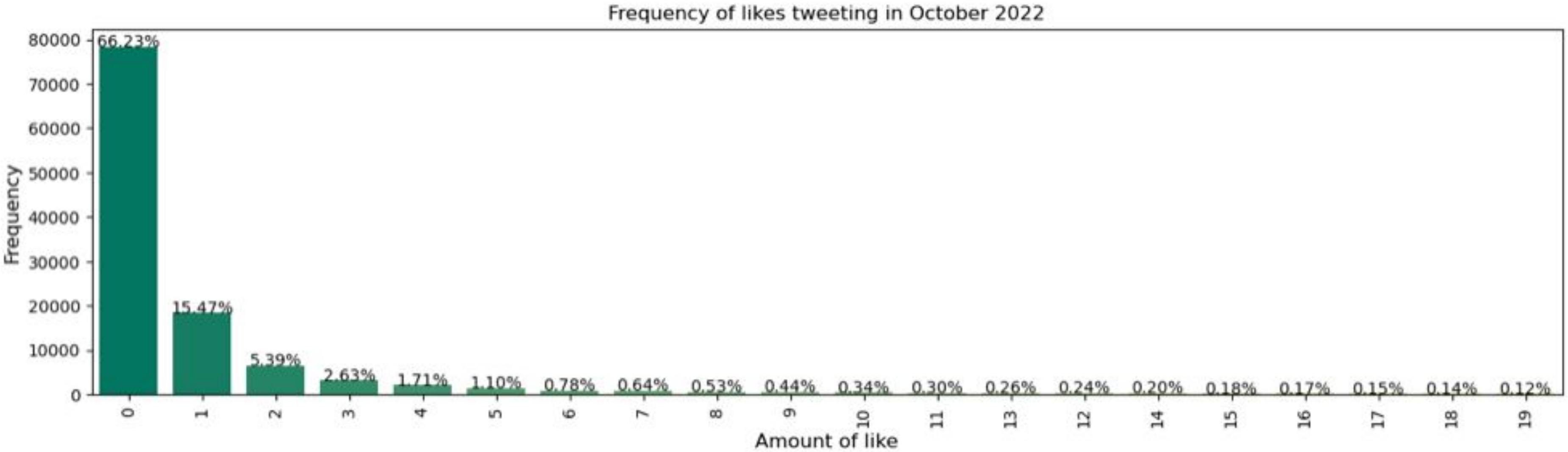


Over 60% of the tweets were posted on December 29th, and December 30th, 2022 as traders were anticipating a Santa Rally. A "Santa rally" is a term used to describe a phenomenon in which the stock market experiences a rise in stock prices in the final weeks of the calendar year, typically in the last week of December. The name "Santa rally" comes from the fact that it often coincides with the holiday season, and some traders believe that the surge in buying during this time is due to increased consumer spending and a generally positive mood.

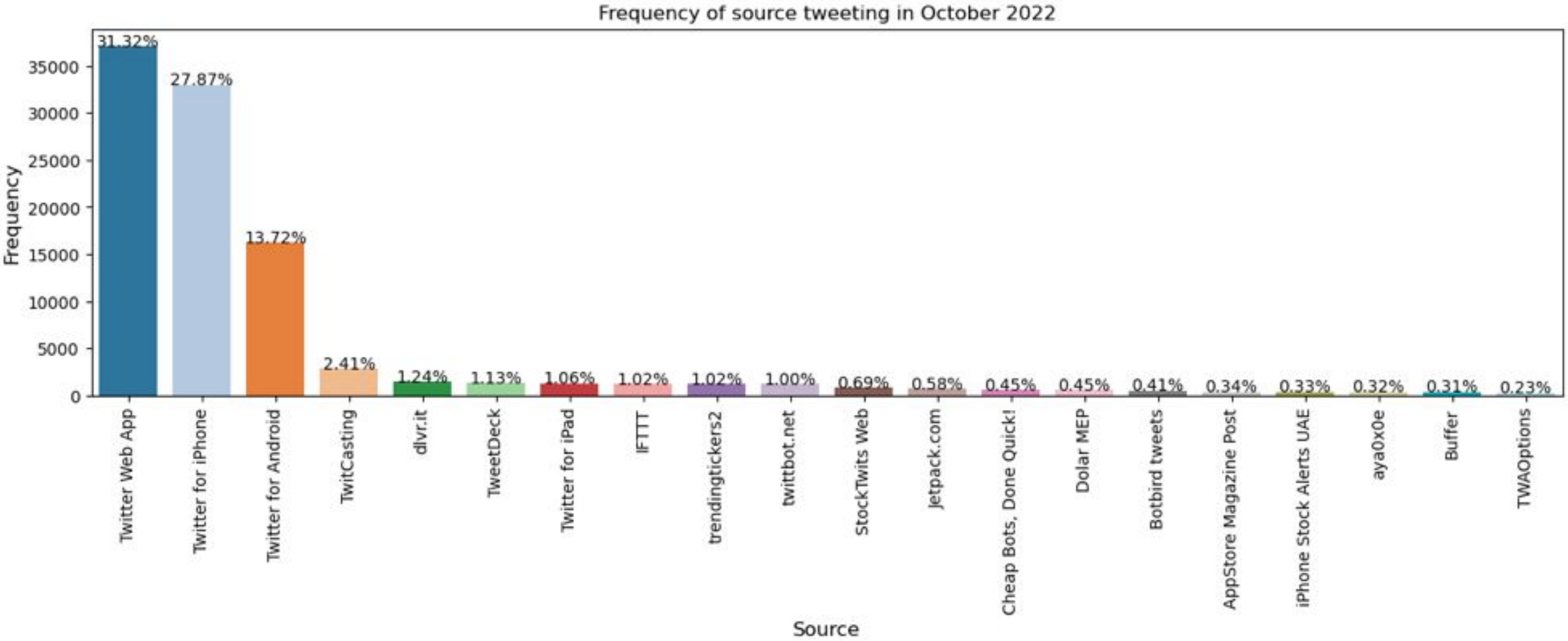


Most tweets were published around 2:30 pm to 2:35 pm UTC (6:00 am to 6:30 am PST), which is considered pre-market hours. Pre-market hours are important to retail traders because they offer the opportunity to react to news and events that may have occurred outside of regular trading hours. During pre-market hours, traders can place orders and make trades based on new information, such as earnings releases, economic data, or geopolitical events, that may affect the prices of the securities they are interested in.

By trading during pre-market hours, traders can potentially gain an advantage over other market participants who only trade during regular hours. They may be able to take advantage of price movements that occur before the official opening of the market, or they may be able to react more quickly to news and events that may impact the market.

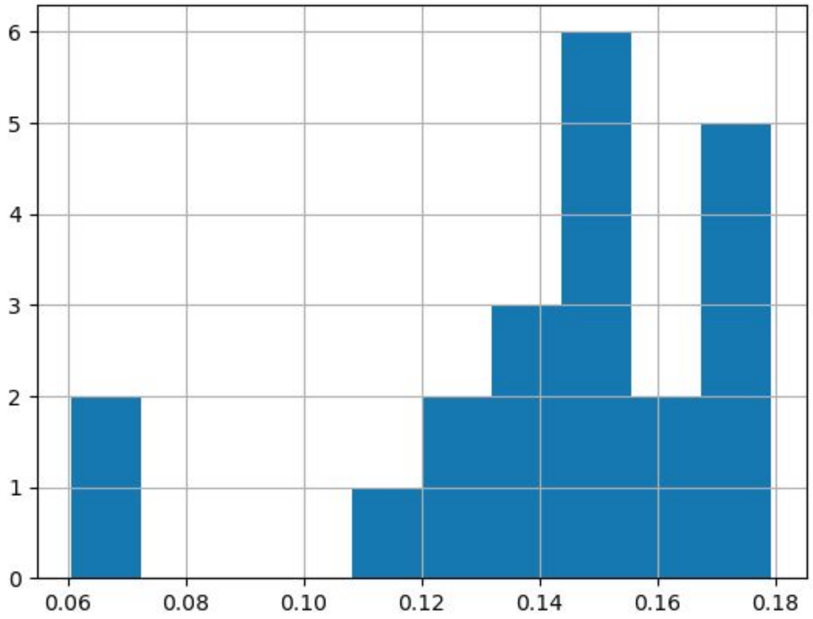


Over 65% of tweets received 0 likes and about 15% received 1 like



Over 30% of tweets were published using the Twitter app

In the Exploratory Data Analysis notebook, I merged the tweets dataset and the stock dataset together (merged_dataframe), showcasing the overall sentiment of tweets of each trading day. I plotted the polarity to see the ranges of the sentiments of these tweets based on its overall daily score. Looking at the range, we can see that the range of 0.14 to 0.16 seems to be the most common score for the month of December 2022. Due to this, we'll choose 0.15 as the "mean", or "neutral" zone for the tweets. Anything that is above 0.15 will be considered a "positive" day, and anything below 0.15 is "negative". After appointing the daily polarity scores to its corresponding label, we now have 11 positive days, and 10 negative days. This, again, shows that the month of December was a positive month.

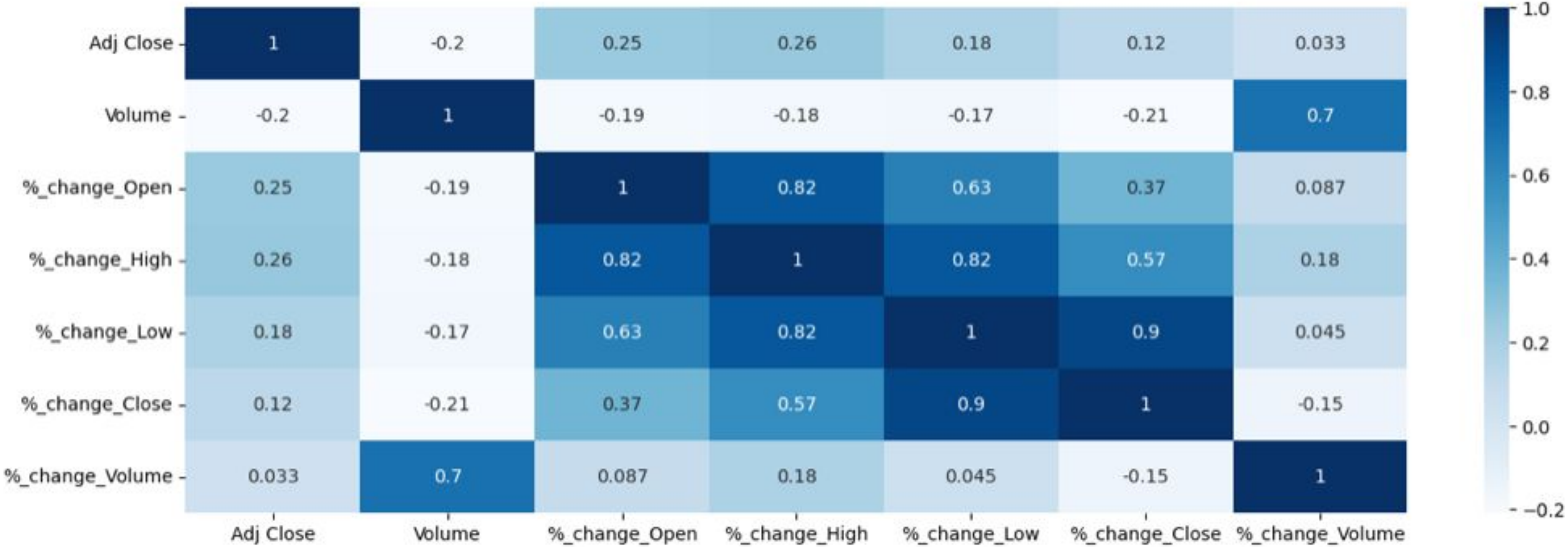


# Stock Price Analysis

Although stock price wasn’t the primary input for our model, I still analyzed Apple stock price for the month of December 2022 to get a better idea of the overall trading sentiment for the month.

The main variable we are going to focus on in this dataset is the ‘Volume’. Variables having a large correlation value with volume represent that those numbers might have intrigued a large number of buyers and sellers. Correlation between those features and the volume feature will tell us how a change in that feature impacts the number of stocks traded that day.

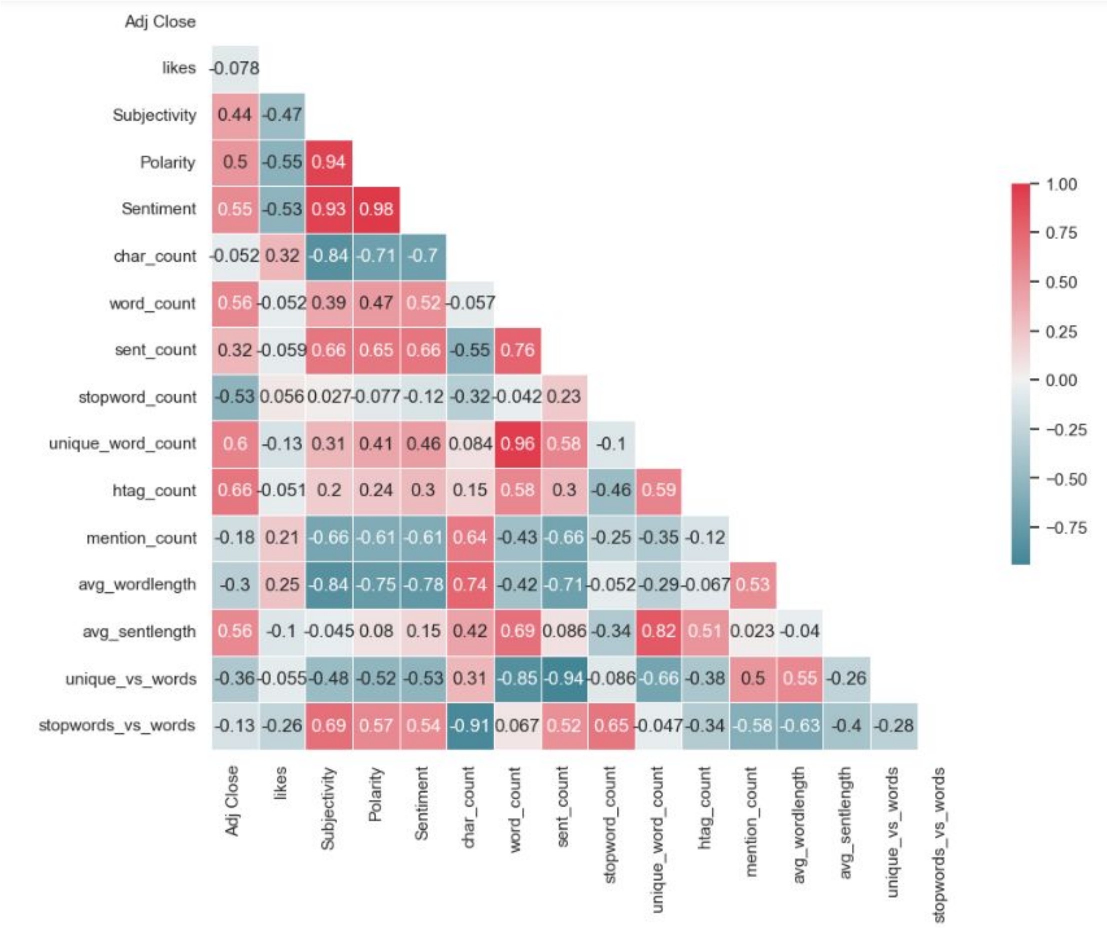
The %_change_Volume shows the most positive correlation with the volume feature. This means that the greater the difference between the volume of the stock today and the volume of the stock price of yesterday, the greater will be the stocks traded that day. As there is no information about the number of buyers and sellers, we can only guess that a high difference in volume might attract more buyers. Whereas a fewer difference may attract more sellers. With this in mind, we know to focus on days with high percentage change in trading volume.


Adj Close
Volume
%_change_Open
%_change_High
%_change_Low
%_change_Close
%_change_Volume
1 -0.2 0.25 0.26 0.18 0.12 0.033
-0.2 1 -0.19 -0.18 -0.17 -0.21 0.7
0.25 -0.19 1 0.82 0.63 0.37 0.087
0.26 -0.18 0.82 1 0.82 0.57 0.18
0.18 -0.17 0.63 0.82 1 0.9 0.045
0.12 -0.21 0.37 0.57 0.9 1 -0.15
0.033 0.7 0.087 0.18 0.045 -0.15 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

# Text Mining

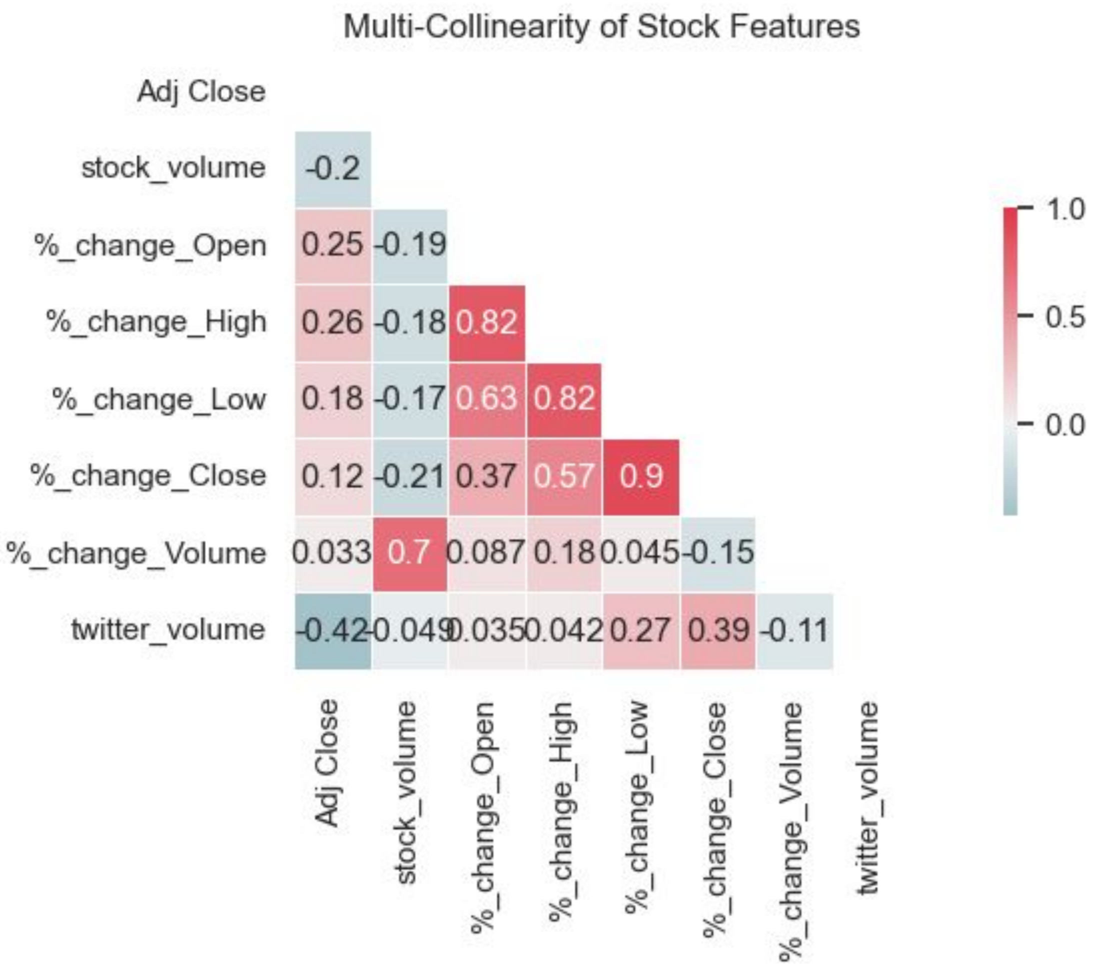
Text Mining is the process of deriving meaningful information from natural language text. I first extracted number of characters present in a tweet, number of words present in each line of tweet, number of punctuation, number of words in quotation marks, number of sentences, number of unique words, count of hashtags, count of mentions, count of stopwords, count of average word length, count of average sentence length, ratio of unique words to total word count, and ratio of stopwords to total word count. Although these feature extractions do not go into our final model, this is to help me understand which features correlate with the tweets sentiment.

Highly correlated variables should be avoided when creating models because they can skew the output. If there are two independent variables that are representing the same occurrence (i.e SqFt of a house vs bedrooms in a house) it can create “noise” or inaccuracy in the model. Models rely solely on outside information in order to create a useful output and having collinear (correlated) variables can create an inflated variance in at least one of the regression outputs.



| | Adj Close | likes | Subjectivity | Polarity | Sentiment | char_count | word_count | sent_count | stopword_count | unique_word_count | htag_count | mention_count | avg_wordlength | avg_sentlength | unique_vs_words | stopwords_vs_words |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Adj Close | | | | | | | | | | | | | | | | |
| likes | -0.078 | | | | | | | | | | | | | | | |
| Subjectivity | 0.44 | -0.47 | | | | | | | | | | | | | | |
| Polarity | 0.5 | -0.55 | 0.94 | | | | | | | | | | | | | |
| Sentiment | 0.55 | -0.53 | 0.93 | 0.98 | | | | | | | | | | | | |
| char_count | -0.052 | 0.32 | -0.84 | -0.71 | -0.7 | | | | | | | | | | | |
| word_count | 0.56 | -0.052 | 0.39 | 0.47 | 0.52 | -0.057 | | | | | | | | | | |
| sent_count | 0.32 | -0.059 | 0.66 | 0.65 | 0.66 | -0.55 | 0.76 | | | | | | | | | |
| stopword_count | -0.53 | 0.056 | 0.027 | -0.077 | -0.12 | -0.32 | -0.042 | 0.23 | | | | | | | | |
| unique_word_count | 0.6 | -0.13 | 0.31 | 0.41 | 0.46 | 0.084 | 0.96 | 0.58 | -0.1 | | | | | | | |
| htag_count | 0.66 | -0.051 | 0.2 | 0.24 | 0.3 | 0.15 | 0.58 | 0.3 | -0.46 | 0.59 | | | | | | |
| mention_count | -0.18 | 0.21 | -0.66 | -0.61 | -0.61 | 0.64 | -0.43 | -0.66 | -0.25 | -0.35 | -0.12 | | | | | |
| avg_wordlength | -0.3 | 0.25 | -0.84 | -0.75 | -0.78 | 0.74 | -0.42 | -0.71 | -0.052 | -0.29 | -0.067 | 0.53 | | | | |
| avg_sentlength | 0.56 | -0.1 | -0.045 | 0.08 | 0.15 | 0.42 | 0.69 | 0.086 | -0.34 | 0.82 | 0.51 | 0.023 | -0.04 | | | |
| unique_vs_words | -0.36 | -0.055 | -0.48 | -0.52 | -0.53 | 0.31 | -0.85 | -0.94 | -0.086 | -0.66 | -0.38 | 0.5 | 0.55 | -0.26 | | |
| stopwords_vs_words | -0.13 | -0.26 | 0.69 | 0.57 | 0.54 | -0.91 | 0.067 | 0.52 | 0.65 | -0.047 | -0.34 | -0.58 | -0.63 | -0.4 | -0.28 | |

According to the multicollinearity for the tweets dataset, as the color becomes darker in either direction (red or blue), meaning that those variables are more highly correlated and should not be paired together in the same model. With that said, since our y is Adj. Close, we need to eliminate any features that are highly correlated to Adj. Close.



No block has significant correlation for stock data.

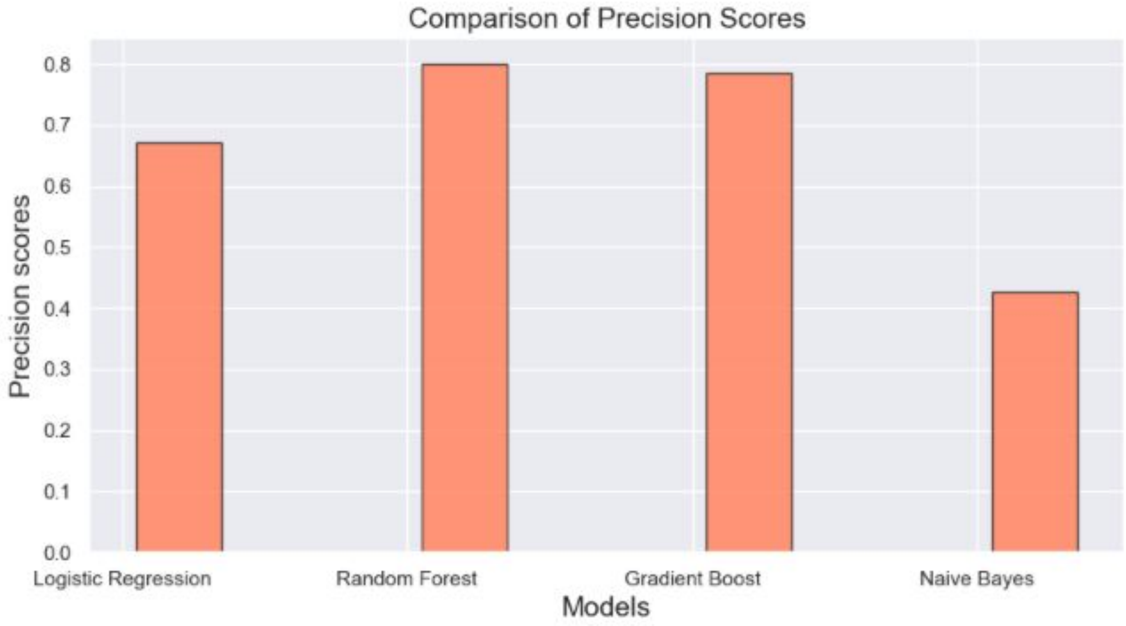
# Modeling Overview

Multi-class Classification in Supervised Learning

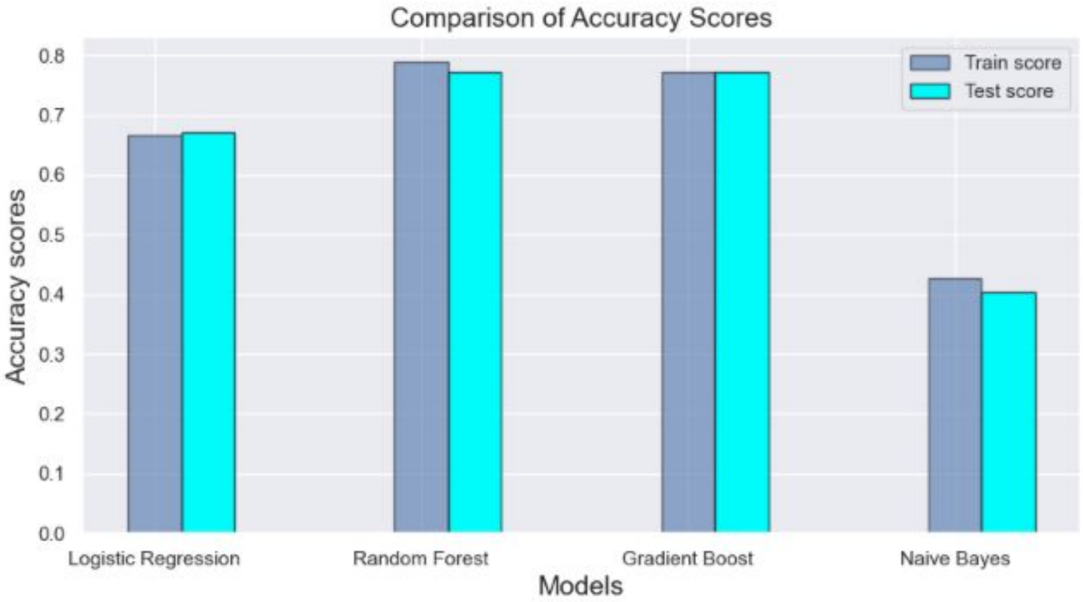
Model Evaluation: Precision

In general, if the goal of the sentiment analysis model is to identify all instances of negative sentiment in a dataset, regardless of false positive predictions, then recall would be the priority. However, if the goal is to ensure a high level of accuracy in sentiment predictions, such that false positive predictions are minimized, then precision would be the priority.In our capstone, we value high level of accuracy, hence precision is chosen.

# Precision Score Comparison



# Accuracy Score Comparison



# Bias/Variance Tradeoff

Due to our high precision scores in Random Forest and Gradient Boost models, our models are very prone to overfitting. Both models are nonlinear and systematically have low bias/high variance.

In order words, if we pass one training dataset into different models, they are more likely to yield different outputs. Results are not consistently and difficult to generalize.

# Conclusion

In order to predict the sentiment of each tweet, we have vectorized the tweets and applied classification machine learning models. Here we have used the following classification models:

- Logistic Regression
- Random Forest
- Gradient Boost
- Naive Bayes

The model evaluation is based on splitting the dataset into train and validation set.

Cross-Validation procedure is used under the k-fold CV approach

We have evaluated each model in terms of model accuracy score, and 'precision' score for both the training and test data, and plotted them. The two best performing models are the Random forest and the Gradient boost. Both are the ensemble model, based on decision trees.

# Conclusion (continued)

Next, we have carried out the Randomized Search CV for the hyperparameter tuning for both the models separately. This step was the most time consuming one in terms of computation. (The RF model took much longer time). I originally attempted the exhaustive Grid Search CV, but with the result of the optimized hyperparameters, we have again fitted the two models, and got the predictions separately. The model precision did not improve much with the optimized parameters, increasing from 0.800 to 0.801 and accuracy staying the same as 0.785, for RF and GB, respectively.

# Future Research

There are possible alternatives to this project that could lead to potential improvements:

- Include engineered features from the stock data obtained from the Exploratory Data Analysis notebook as the values for X in our models
- Include text mining features from the tweets data obtained from the Preprocessing and Training Data notebook as the values for X in our models
- Use a whole year worth of tweets instead of just the month of December 2022
- Use different hashtags