Stock Scrapers

Nathan DePiero
John Ryan Byers
Carlos Betancur
Hunter Adrian

Background

The influence of social media has grown significantly in recent years, making it a potentially powerful tool for predicting trends and influencing various sectors of society, including the financial markets. Social media platforms such as Twitter and Reddit have emerged as popular forums for discussions and speculations about the stock market, potentially influencing trading patterns and volatility. In our project we aimed to explore this association and investigate the possible impacts of social media mentions on stock price volatility.

Hypotheses

- There is a statistical difference between the volatility levels one day after a low number of stock mentions and volatility levels one day after a high number of stock mentions.
- There is a significant difference in volatility between one day before and one day after an abnormally high number of stock mentions.
- There is a statistical difference between the volatility following a higher number of Twitter stock mentions and the volatility following a high number of Reddit stock mentions.

Data & Sources

We collected data from three separate sources that we merged as tables into one SQL Database.

- **WallStreetBets Dataset:** Reddit posts from 1/28/21 to 8/16/21 containing users' posts mentioning certain stocks. From the various fields of information the original source contained, we extracted the number of posts each day during the time period for every stock.
- **Stock Market Tweets Data:** Twitter posts mentioning stock tickers from 4/9/2020 to 7/16/2020. Similar to the Reddit dataset, we aggregated the number of posts per day for each stock.
- **Yahoo Finance API:** extracted stock price volatility for each of the mentioned stocks and associated time frames.

We chose not to store the contents of posts, authors, or other personal information in our database because of the sensitive nature of these fields. In addition, appropriate measures were taken to clean our data and remove duplicates.

Methodology

For our hypothesis testing we ran Two Sample T-tests and one Paired T-test. We used the Two Sample T-test for our first and third hypotheses because we wanted to measure the difference in means between two samples. For our second test we ran a Paired T-test because we were comparing matched pairs of volatilities which were not independent variables.

In the machine learning part of this project, we conducted a linear regression seeking to identify a possible trend between the number of mentions and stock volatility one day later. Finally we ran a k-means analysis to find possible correlations between a stock's attributes and market cap.

Hypothesis Testing

To test our first hypothesis, we conducted a Two-Sample T-Test, comparing mean price volatility a day after high social media mentions versus low social media mentions which did not return a significant p-value **(0.171)**.

We tested our second hypothesis with a Paired T-Test comparing price volatility the day before and after high social media mentions, which returned a significant p-value **(0.0047)**. Upon inspecting the results further, we found that the volatility decreased from one day before to one day after on average. We expected posts to increase volatility, so this surprised us. A possible explanation for our findings is after a period of high volatility, people tend to react with posts, and then the stock settles. Therefore, the stock volatility changing is not a result of the post, but the post is a result of the stock volatility.

To the right, Figures 1 and 2 display share of mentions by each stock on Twitter and Reddit respectively. We observed different stock mentions shares on each social media platform and wanted to determine whether the relationship between the number of mentions and price volatility differed between Twitter and Reddit. However, a Two-Sample T-Test comparing price volatility a day after social media mentions on Twitter versus Reddit returned an insignificant p-value **(0.657)**.

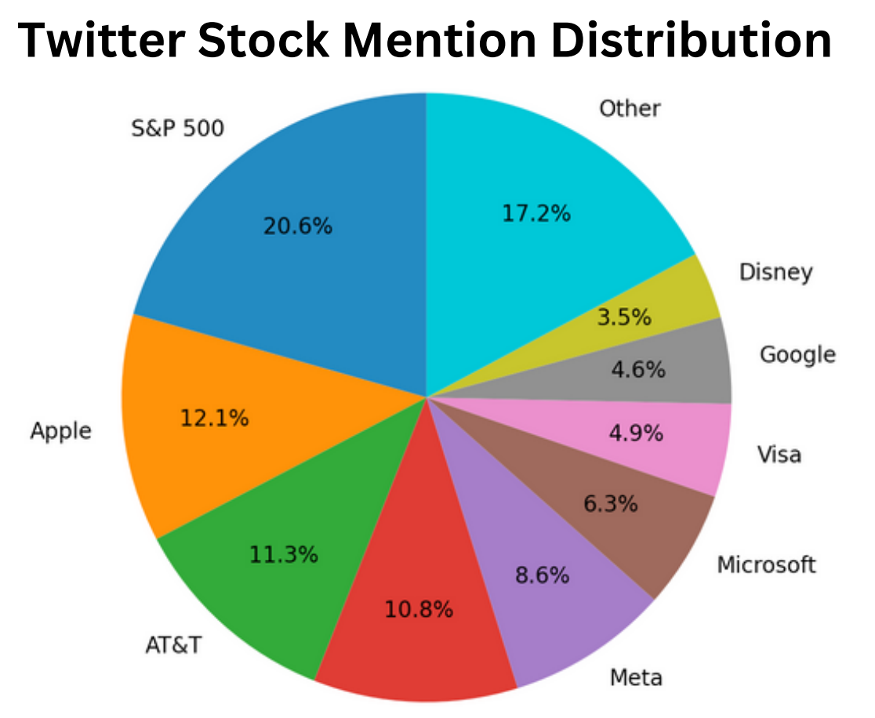
Twitter Stock Mention Distribution

Figure 1: Stock Mentions on Twitter

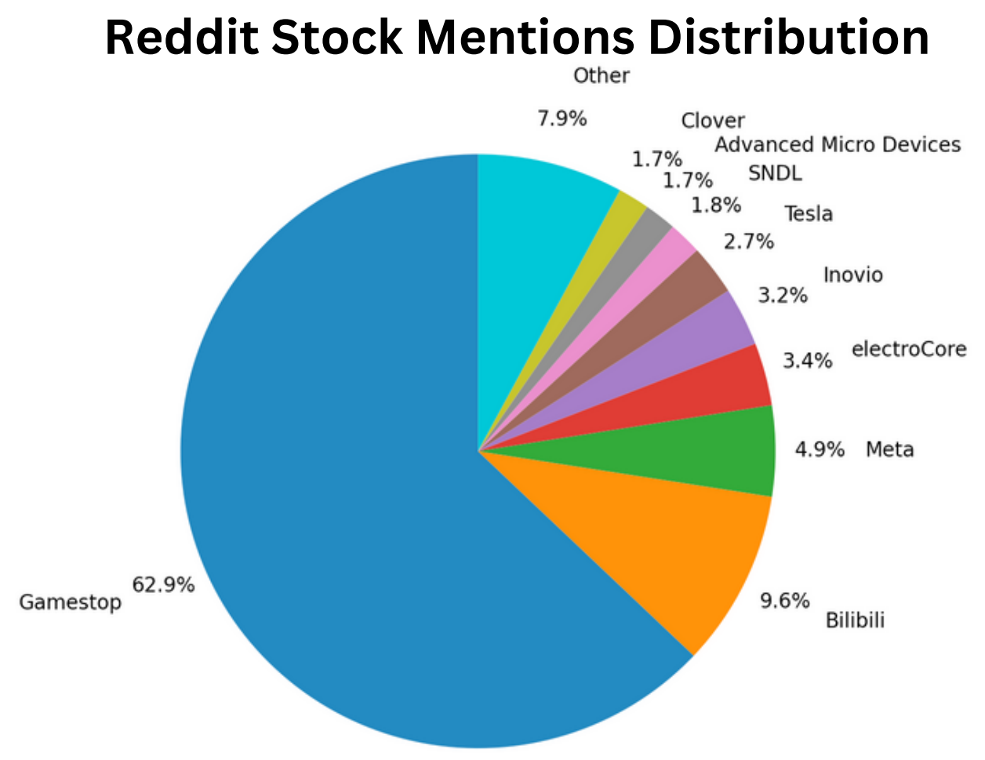
Reddit Stock Mentions Distribution

Figure 2: Stock Mentions on Reddit

Machine Learning

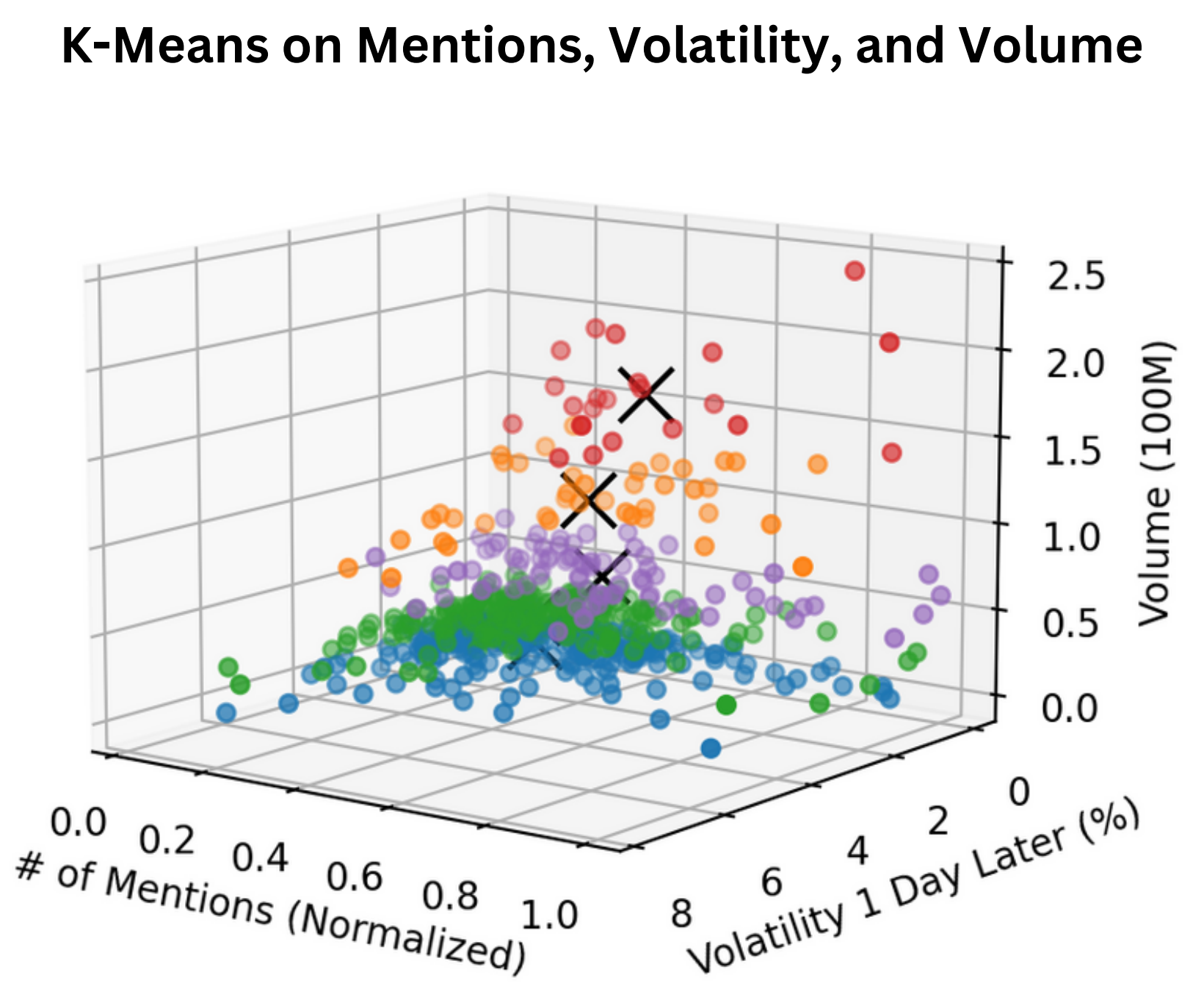
K-Means on Mentions, Volatility, and Volume

Figure 3: K-Means

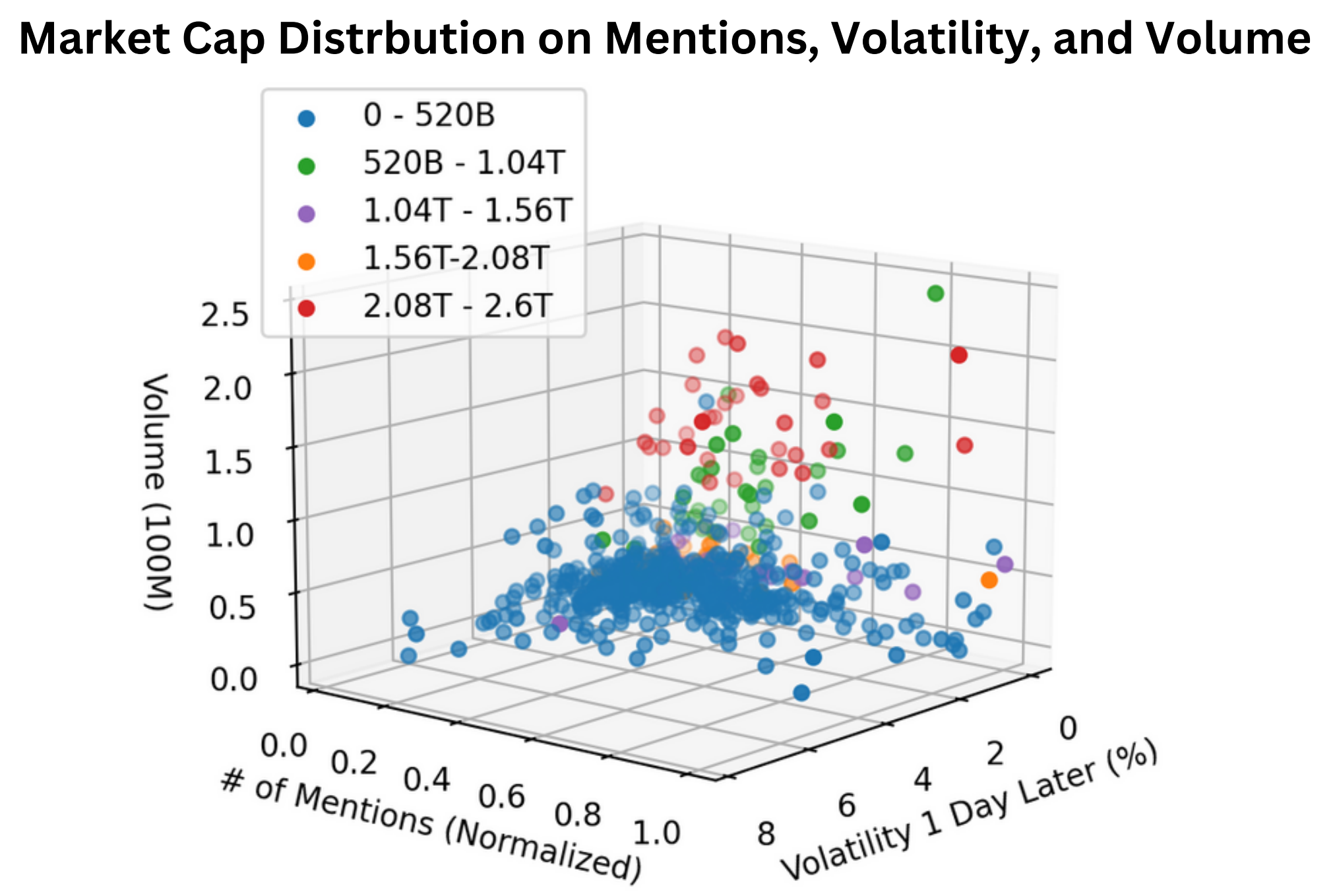
Market Cap Distrbution on Mentions, Volatility, and Volume

Figure 4: Market Cap Visualization

We ran K-means on three dimensions of the data: number of mentions, volatility, and volume (number of shares traded in a day). Our goal was to see whether the groupings generated by k-means reflected the stock's market cap (worth of company). Figures 3 and 4 depict the k-means and the market cap groupings, respectively. We found there was some relationship between the two groupings, but we recognized that the k-means groupings were mostly along the volume axis, and likewise for the market caps. This makes sense because companies become more popular as the market cap increases, and therefore, their stocks are traded more.

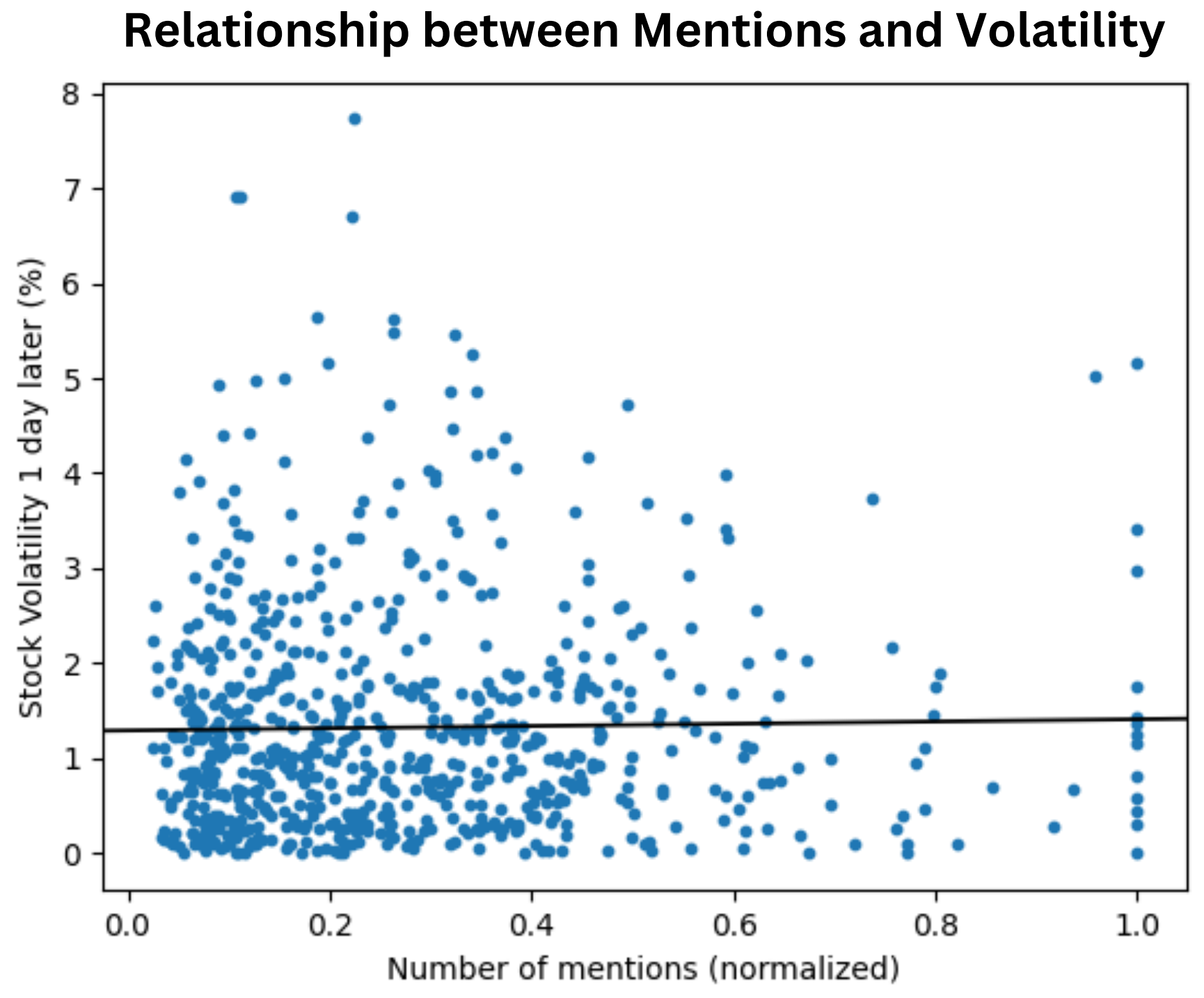
Relationship between Mentions and Volatility

Figure 5: Linear Regression on Twitter Data

In hopes of determining whether a relationship between social media mentions and price volatility existed, we wanted to run a regression. Since both variables were continuous, we did not use a logistic regression which is for discrete variables. In addition, our analysis focused on two variables, so we did not use multiple linear regression. Figure 5 depicts the linear regression we ran, with the number of mentions as the independent variable and stock price volatility one day later as the dependent.

We used an 80/20 train test split and received an MSE of 1.46 on the training set and 1.71 on the test set. The difference in errors was not significant, suggesting our regression model is generalizable. We got r-squared values of 0.001 and 0.0005 on the train and test sets, thus failing to indicate a linear relationship between mentions and volatility.

Takeaways

Hypothesis Testing

1. There is **no** statistically significant difference between stock price volatility one day after a low number of social media mentions and one day after a high number of social media mentions.
2. The difference in stock price volatility the day before high social media mentions and the day after **is** statistically significant.
3. There is **no** statistical difference between the volatility following a higher number of Twitter stock mentions and the volatility following a high number of Reddit stock mentions.

Machine Learning

- **Linear Regression:** Observed no statistical evidence suggesting a significant relationship between the number of mentions and volatility one day later.
- **K Means Algorithm:** After visual inspection, we noticed that the graphs looked fairly similar and the groupings were mostly along the z-axis. which uncovered a relationship between daily volume and market capitalization.

Conclusions

Our analysis failed to provide statistical evidence supporting our first and third hypotheses, nor were we able to draw any conclusions from our Linear Regression. We were disappointed that we were not able to uncover a predictor for stock movements, but since the market is usually efficient it was to be expected.

However, our hypothesis two testing showed that volatility tends to decrease from a day before a high number of mentions to a day after. These results surprised us initially, but made sense after further thought as explained. Unfortunately, we do not think this information is actionable in stock trading. Moreover, our k-means component highlighted a pattern between market cap and the volume of shares traded.

Overall, the inspiration for this project rested on the news that Reddit and Twitter users had made impacts on stock price by telling others to buy or sell shares. With our findings however, we determined that this was not the case.

Limitations

One limitation of our analysis was span of time our datasets provided. Our Reddit and and Twitter datasets drew posts from a year each. If we were to continue studying this relationship we would search for larger datasets covering a longer period of time. In addition, we would look for posts from other social media sources such as Instagram and TikTok.

Additionally, our approach focused primarily on stock mentions and stock price volatility. Collecting other stock qualities or even looking at sentiments of mentions during these time periods might have also enhanced the quality of our tests and machine learning components.