Meme and Traditional Stock Sentiment Analysis

CMPT 353

Professor: Steven Bergner

April 12th, 2025

By: Sukhman Virk, Harsh Sidhu, Aniyah Bohnen

Introduction

Today, almost everyone has access to technology, making individuals more connected than ever. The internet has not only transformed the way we communicate, but it has also opened up an endless amount of ways to build wealth and increase income. One example is the stock market, allowing users to invest money into companies for future profit. However, many media platforms exist, such as StockTwit and other sites, generating ”hype” and influencing users to invest their hard-earned money without understanding the risks. This phenomenon has caused users to invest impulsively without conducting traditional financial analysis. As a result, many investors are losing thousands of dollars. Meme stocks such as Gamestop and BlackBerry show how online influence can seriously dictate stock market spikes and drops. The purpose of this project is to identify the connection between online posts through media platforms and stock fluctuation. Further, this report helps investors understand whether positive or negative posts signal a good time to act.

Data collection and Cleansing

The data set we used for this project comes in two CSV files. One of the CSV files includes stock market data and post headlines. These are in the file StockandSentiment.csv. The data that we have collected comes from an API Alpha Vantage. These include historical data from three meme stocks GameStop (GME), BlackBerry(BB), and American Movie Classics (AMC). Similarly, we have data on three more traditional stocks. These include Apple (APL), Microsoft, (MFT), and Johnson and Johnson (JNJ). Using StockTwits, we were able to get several posts about our stocks. These include positive and negative posts. Similarly, we have another data collection file news_data.py. Similarly, news_data.py does the same. However, it fetches news headlines through another API. We will use these news Headlines to see how the media sentiment can also influence stock prices.

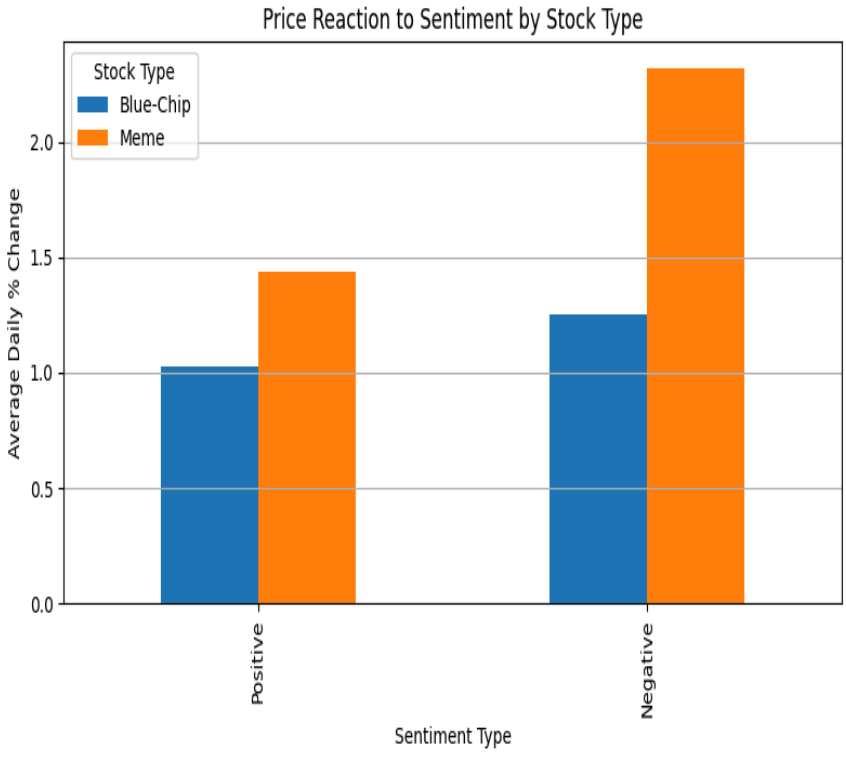
Within these data collection files we made the following changes to clean them:

1. Added columns such as Percent Change Category, which indicate whether there was a spike or drop in data. The headline column, which we grouped the stock with a news headline within the same time frame, and more.
2. Renaming multiple columns that make more sense in our data set.
3. In our analysis, when creating a dataframe for sentiment analysis, we dropped data that had empty values. Also, fill in missing values in data frames by filling in 0 in empty value.
4. Reading in and converting json files into csv and more.

Analysis

Part 1: Sentiment Analysis

Before diving into sentiment analysis, our team compared meme and blue chip stocks to create a baseline for better observing their natural behavior. For this process, we used the groupBy function and the agg function to find averages of stock type and stock prices. Our results showed meme stocks were more volatile. For example, the analysis showed that meme stocks had an average volatile rate of 6.15, compared to 5.36 of the traditional stocks. We also decided to do statistical tests, such as the Mann-Whitney, as we were unsure if the data distribution was normal. The statistical test confirmed a significant difference between the two, which means their fluctuations differ. Next, we looked at sentiment. To create a sentiment rating, we used Textblob, which assigned a sentiment rating to each post. These values



allowed us to understand whether the headlines were positive, negative, or neutral. After comparing the two, we noticed that meme stocks were influenced more by media posts. For example, positive posts increased meme stocks by 1.43 percent. Compared to traditional stocks, with an increase of 1.03 percent. An interesting insight was that even negative headlines lead to positive spikes. For example, Meme stocks rose 2.32 percent, while traditional stocks increased 1.25 percent. To visualize these results, we created a bar graph comparing the average percentage change by sentiment and stock type.Similarly, 30.77 percent of days with a negative sentiment lead to a spike. However, only 28.77 percent of positive sentiment had spike days.

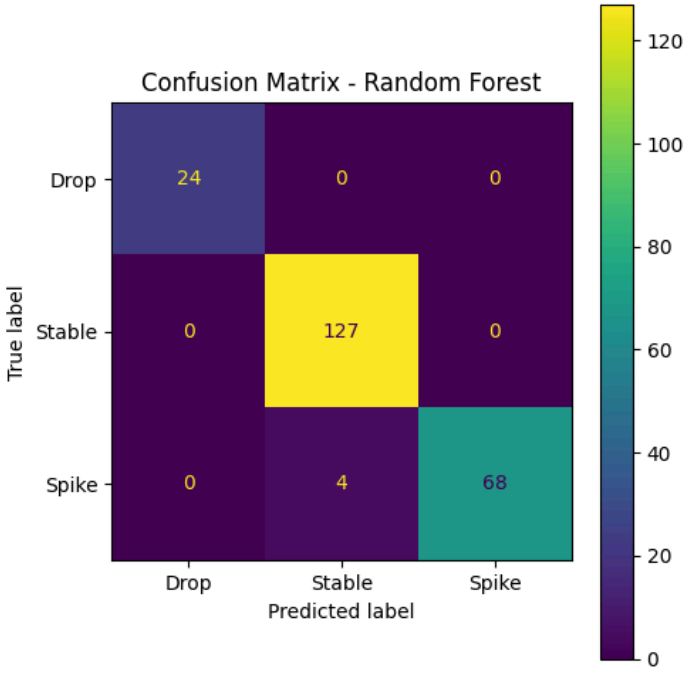
To understand the relationship between sentiment and stock prices, we performed a correlation test. The r value of our test was 0.0368. The result indicates that there is no linear relation between sentiment and price. We also applied another Man-Whitney, which concluded that there was no significant difference in sentiment on spike and drop days. Moving forward, our group did natural language processing. Through nltk, we cleaned and removed common stopwords. This way, we could generate a list of frequent words that are not a part of daily language. Further, we generated a figure through the word cloud, which shows keywords by size that appear most frequently in our data. Our team was able to showcase words such as "Trump" and "tariffs", which led to spikes. Also, our group decided to check which stocks had the highest positive and negative posts. Surprisingly, Blackberry had the highest positive post at 47. Microsoft had the most negative post at 23. Although BlackBerry is not a strong financial performer, other factors, such as meme hype and nostalgia, can help drive pessimistic sentiment. Whereas the much more traditional stocks receive more backlash and negative posts as they are much more stable and well-established firms that are expected to perform well. After looking at the most appearing words, we wanted to see which words appeared on the most spike and drop days. This analysis could highlight what language dictates market shifts. Our analysis revealed that words such as "immediately" and "expiration" lead to positive spikes the next day. These words may create a sense of urgency to invest in particular stocks.

Part 2: Predictive Modeling

To further investigate the influence of sentiment and volatility on stock behaviour, we built predictive models to classify daily stock movements into three categories: Spike, Stable, or Drop by using features derived from sentiment and price data. We implemented three classification models: Random Forest, K-Nearest Neighbors, and Gaussian Naive Bayes. The key features we used include daily sentiment score, headline count, 7-day volatility (standard deviation of past 7 days returns), and a boolean indicator of high volatility. These features capture the sentiment of the news headlines and recent price variability.

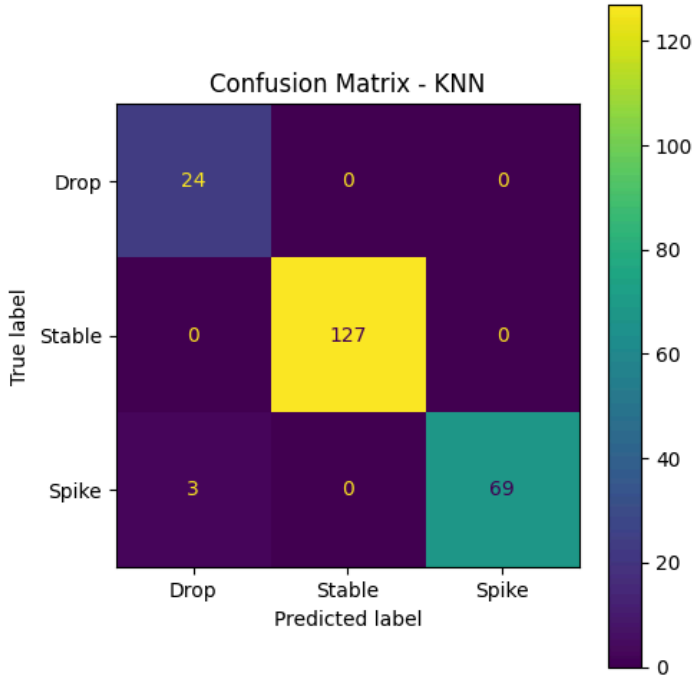
After training the models on our data, we evaluated each model's performance using standard classification metrics. The results for each classifier is summarized below:

Random Forest:
This ensemble model predicted the movement category with accuracy around 97%. It maintained strong balance across all categories suggesting that this model captured the underlying patterns. Due to this model's ability to handle nonlinear interactions, it allowed it to utilize these features fully. The Confusion Matrix below shows that the model identified 24 Drop days, 127 Stable days, and 68 Spike days. Only misclassified 4 Spike days as Stable.



K-Nearest Neighbors:

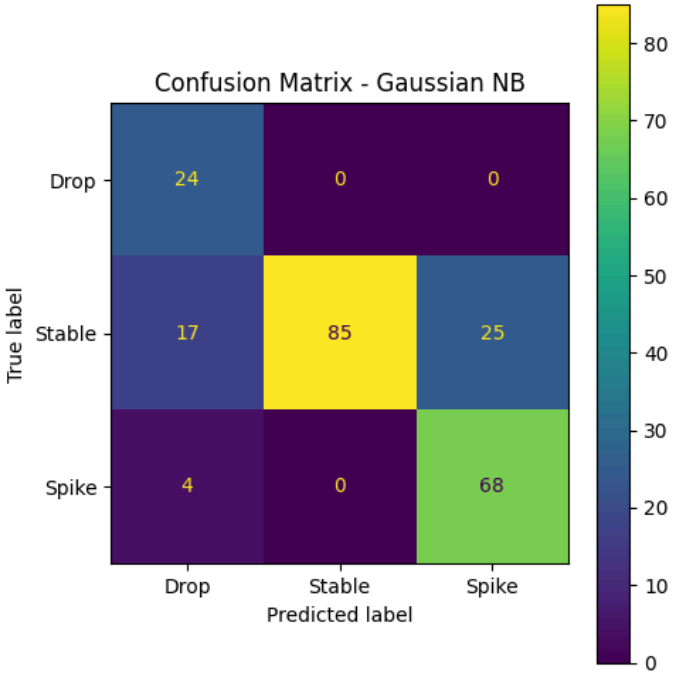
The KNN classified also performed well with accuracy near 98% on the test data. Similarly to the Random Forest model, this model was able to distinguish Spike, Stable, and Drop days with high reliability. The high performance from this model suggests that the clustering of feature values for different outcomes was very distinct in the data. In other words, Spike days and Drop days were clearly in separated regions in the feature space making them easy to classify with this approach. The Confusion Matrix illustrates the similar performance with the previous model but it misclassified 3 Spike days as Drop.



Gaussian Naive Bayes:

The Gaussian Naive Bayes model had lower performance compared to the previous two models. Its accuracy was around 80%. The precision and recall for certain classes were weaker which indicates that the model confused Spike days with Stable or Drop days. The reduced performance is likely due to Naive Bayes simplifying assumptions: it assumes feature independence and normally distributed features. However, in our dataset, sentiment, headline, count, and volatility were not independent. For example, high volatility days are correlated with high sentiment. These violations mean the model could not accurately capture the

complex relationships in the data. Thus, this model was not as well-suited to our stock movement data.



Part 3: Headlines vs Stock Behaviour

In this section, we explore the relationship between media coverage (as measured by the number of news headlines) and the behavioral categories of several popular stocks, that is, its spikes, drops, and stable days based on daily percent changes.

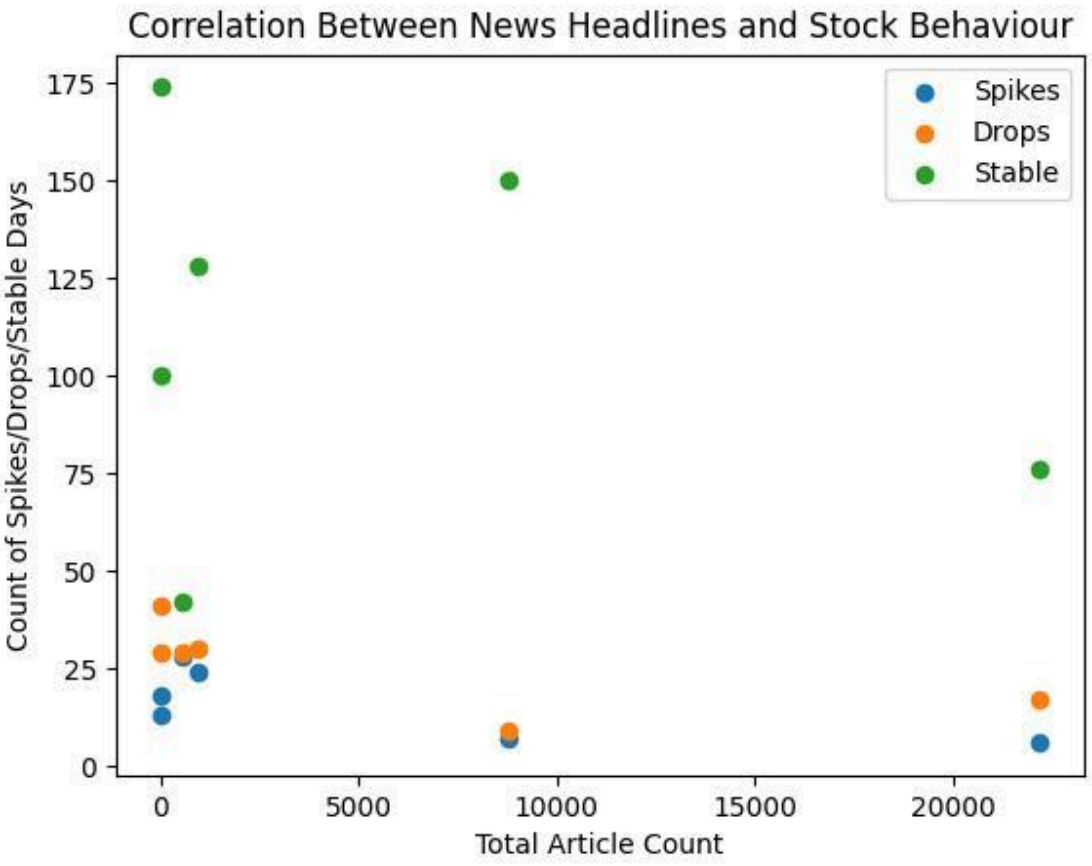
Figure 1 visualizes the relationship between the total article count for each stock and the number of days categorized as spikes, drops, or stable. A key observation is that there is no clear linear relationship across all behaviors. Some stocks with minimal news coverage (e.g., <500 articles) still experienced a relatively high number of spikes and drops, while others with a large number of headlines showed low volatility.

Using linear regression analysis from scipy.stats.linregress, we calculated the following correlation coefficients (r-values) between the number of headlines and stock behavior:

- Spikes: $r \approx -0.720$
- Drops: $r \approx -0.679$
- Stable: $r \approx -0.220$

The low r-values across all three categories suggest that media coverage is not a strong predictor of stock behavior in the short term. This aligns with the understanding that while news can influence investor sentiment, there are other market forces, such as earnings reports and economic data that often play a more direct role in driving rapid stock movements. As discussed further in the limitations section, if we had access to more data that could provide long term insights, we may be able to see a stronger relationship between media coverage and stock behaviours.

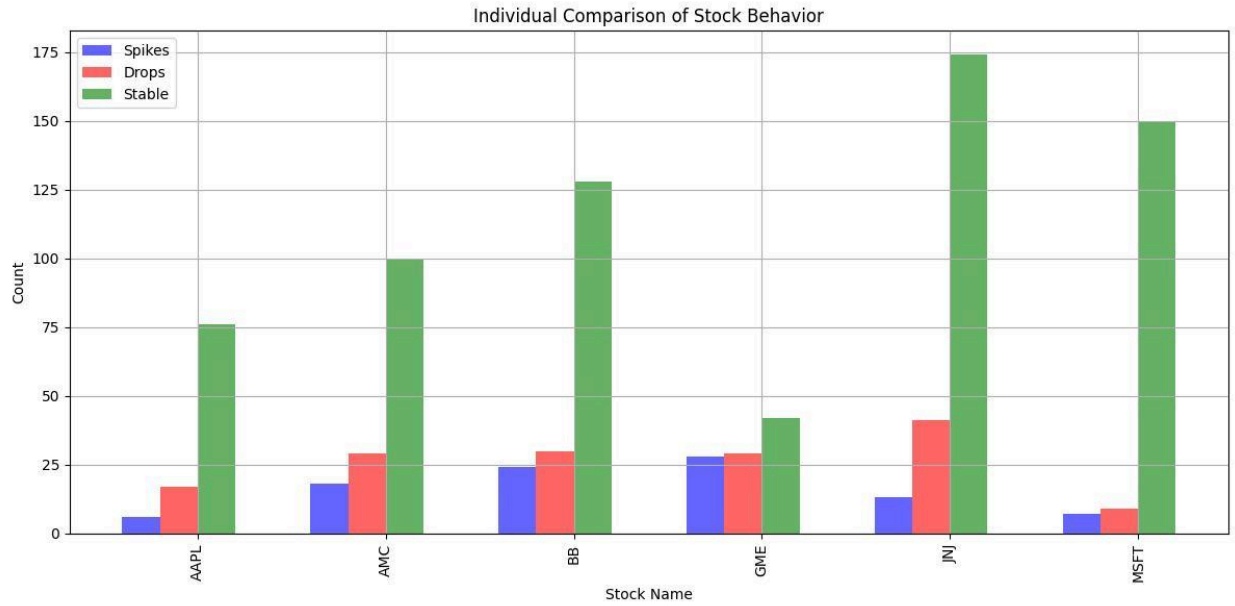
Figure 1:



The second visualization (Figure 2) shows a grouped bar chart comparing the count of spikes, drops, and stable days for each stock. Some observations include that MSFT and JNJ show overwhelmingly high stability, with minimal spikes or drops, despite having relatively high news coverage. It could be considered that the more stable stocks receive more news coverage rather than stocks that see a lot of changes in behaviour as it may not be as appealing to the general consumer to potentially invest in a volatile stock. BB, AMC, and GME, known for their high volatility, have a more balanced mix of spikes, drops, and stable days. AAPL, despite being a highly covered stock in the media, exhibited moderate levels of stability and relatively few extreme movements in the market.

This suggests that stock behavior is highly stock-specific and that coverage volume alone does not directly translate into price volatility. For instance, MSFT may receive news coverage that is more technical or business-related, which may not drive immediate market reactions, whereas GME or AMC may react more sensitively to even low volumes of speculative news.

Figure 2:



From these observations, we can see that stable behaviour is a dominating factor for major stocks which reinforces the idea that that volatility is more common in speculative or meme-driven stocks. Additionally, we discovered there is no strong linear correlation between the number of news articles and the frequency of spikes or drops in stock prices through finding the correlation coefficient and visually from the scatterplot.

Limitations

We had several limitations for our project, For example, we used a free API. Thus, for much of our data we were limited in terms of the amount of data we could get. For example Alpha Vantage limited us to 25 requests per day. Thus, we could not compare more stocks because that would limit the amount of times we could run our project. Another limitation we dealt with was our knowledge of natural language processing. Although we were able to find interesting insights, we did not learn how to perform natural language processing. Therefore, we had to learn how to do this through online resources. Furthermore, certain companies had more coverage than others, This made an unbalance between the amount of data we had. The News API and StockTwits only allowed a limited amount of posts or news headlines.

Retrospective

Without time constraints, we would have liked to analyze more media platforms. In particular, we would have preferred to examine how social media influence differs from StockTwits to Yahoo. Also, if we had more time we would have liked to create our model for natural language processing. Although we are unsure how to do this, it would be nice to do more research about this in the future. Furthermore, more time spent on researching APIs would have been helpful to get more quantity and better quality data.

Conclusion

From our insights, we learned that meme stocks are more volatile than traditional stocks. This was further shown through the Mann-Whitney test, which had a p-value of ($p < 0.00001$). Although we thought that sentiment could have an large influence on stock prices, our analysis shows little correlation. Before this project, we thought negative posts and headlines would lead to drops. Despite this, we discovered that even negative sentiment could lead to positive spikes. Our analysis shows that 30 percent of stocks that had a negative sentiment had positive spikes the next day. Only 27 percent of stocks with a positive sentiment had spikes. Therefore, our findings suggest that sentiment cannot be the only factor influencing stock prices. We also used NLP to observe the words that have positive

or negative spikes. This way, we could understand by looking at data to see if it was possible to know if there would be a spike or drop. We created a visual of the word cloud, which shows the most common words in posts. For example, most people had "tariffs" and "Trump" in their posts. These words show that even though the sentiments themselves cannot accurately predict when stocks fall or spike, they can help analyze other external factors. For example, Trump has increased his tariffs on other countries, which can increase company costs, resulting in lower performance. The low performance can impact the price of various stocks. Also, we saw "China" appear, which may be engaging in a trade war with the United States. Thus, our findings suggest the words in posts can help identify external factors influencing the stock market. We believe individuals should not use media posts alone to influence their decisions. Other factors are involved, as explained above. Positive and negative headlines do not signify a drop or a spike, as we explained, even negative headlines have led to spikes. Thus, individuals should not worry if they have stocks that are being talked about negatively on media platforms. However, they should not invest in stocks talked about positively either.

Project Experience Summary

Sukhman Virk:

- Created the “StockTwits_data.py” file, which uses Pandas, Python to clean, collect, and merge tables with over 2000 rows of data based on six stocks for further analysis.
  Applied sentiment scores to each post through textblob, so we can check for correlation.
- Used natural language processing to find words related to spike and drops, to see if there is any correlation.
- Performed statistical analysis to check correlation between six stocks.
- Reported insights to the report to share them with others

Aniyah Bohnen:

- Created the news_data.py file, this connects to a news API to collect the headlines data into a CSV (“newsapi_last_30_days_combined.csv”).
- Applied Linear Regression and Data Filtering with numpy and pandas in “news_analysis.py” to create visualizations depicting correlations between headlines and stock behaviour.
- Added Readme.md containing project description and instructions.
- Evaluated relationship between stock behaviour and news headlines.

Harsh Sidhu:

- Analyzed price and sentiment data across meme and traditional stocks to predict future price movements
- Applied machine learning classification models and statistical testing to evaluate market behaviour and volatility
- Interpreted model outputs from machine learning classifiers and summarized key insights from the analysis on the report
- Created data visualizations, including confusion matrices to clearly communicate model performance and support conclusions