Predicting the Movement of “Meme” Stocks

Prabhav Pandya[1], Omkar Prabhune[2], Pritesh Pawar[3], Dr. ST Patil[4]
[1]Student, Vishwakarma Institute of Technology, Pune. (prabhav.pandya19@vit.edu)
[2]Student, Vishwakarma Institute of Technology, Pune. (omkar.prabhune19@vit.edu)
[3]Student, Vishwakarma Institute of Technology, Pune. (pritesh.pawar19@vit.edu)
[4]Professor, Vishwakarma Institute of Technology, Pune. (patil.st@vit.edu)

***Abstract*—**Predicting the stock market has been a popular topic of discussion among the investors, analysts and even researchers for a long time. But due to the number of factors affecting stock prices daily including, but not limited to, political, environmental & economical factors, it has become extremely difficult to accurately predict the market. Over the past two years, we have seen stocks like GME & AMC fail some of the most advanced models out there. These stocks were pushed by retail traders who used social media networks like Reddit & Twitter. In this paper, we find commonalities between such stocks and use sentiment analysis on Twitter data to analyze public sentiment and the correlation of “meme” stock prices with it.

Keywords*—Meme Stocks, Twitter Analysis, Machine Learning, Semantic Analysis*

I. Introduction

Earlier the forecasting of the trade market was mainly based on the previous record of the stocks. Later many researchers disproved this approach as the only way to forecast the movements of the trade market. The prices of stocks in the trade market change vigorously due to many reasons such as supply and demand, market volatility and many other dependent and independent factors. Currently, many of the existing systems are performing time series analysis, which involves the sequential plotting of a set of observations or data points at regular intervals of time. It does so by studying the previous outcomes and their progression over time.

In January 2020, a community of traders on r/wallstreetbets and twitter noticed major hedge funds heavily shorting Gamestop’s stock. 140% of Gamestop’s public float was being being shorted to be precise. Traders on these online forums decided to coordinate and buy Gamestop’s shares which triggered a short squeeze and led to Gamestop’s price going up from $17 in the beginning of the month to $500 per share. This was when the term “meme stocks” was coined, because they showed unusual price movements without any change in fundamentals of the underlying asset.

Most stock prediction models focus on technical and fundamental aspects of stocks. Though these models have improved over time with the use of more sophisticated time-series models, they still tend to miss one of the most important factors affecting stock prices daily, Public Sentiment. Studies focussing on the Quantitative correlation between financial news and stocks have shown moderate levels of statistically significant correlation between news and financial performance[1], implying a correlation between public sentiment and stock prices. Another study has shown a positive correlation between a stock’s volume and the volume of its web search queries[2].

According to Datareportal 2022 global overview, more than half of the world now uses social media and the average daily time spent on social media is around two and half hours. So the data available is potentially very large. Nowadays, social media is portrayed as a platform to share thoughts on various topics happening around the world like prices of stocks and has a notable impact on other people’s opinion. There are many platforms like Reddit, Twitter which are gaining popularity and researchers are also considering them for studies. Twitter is an example of a social media site, and its primary purpose is to connect people and allow people to share their thoughts with a big audience in real time.

In this paper, we leverage this data and find the correlation between public sentiment on Twitter and stock prices. We specifically focus on “meme” stocks, due to their high volatility and dependence on retail investors who are motivated to trade through social media[5]. “Meme stocks” is an umbrella term for stocks or cryptocurrencies that have gained attention among retail investors and show unusual stock price movements.

II. Literature Review

Various studies have been done in the field related to the correlation between stock prices and news sentiments.

In one of the most famous works published by Johan Bollen, Huina Mao states that to analyze the tweets as data input, they mainly used frame of mind analyzing systems i.e., Google Profile of Mood States(GPOMS) and Opinion Finder. The GPOMS calculates the frame of mind in 6 ways: Happiness, Kindness, Vitality, Calmness, Alert and surety. Then they cross-validate this result with significant cultural events. After that, they used Neural Networks to investigate the hypothesis that different people have made and are predictive of the changes in Dow Jones Industrial Average closing values. They gave an accuracy of 87% in forecasting the movement. Also, they achieved a reduction in mean average percentage error by greater than six percent. Along with Neural Networks, they have also used granger causality analysis. The accuracy metric used here is MAPE.

Another work published by Kevin Hu, Daniella Grimberg, Eziz Durdyev proposed the hypothesis that the emotions of public expression on social media platforms like Reddit or Twitter are correlated with the movement of stocks in the trade market. They have divided their approach into two parts. The first part consists of a classification model representing the public's emotions and expressions defined by the data on social media platforms. The second part consists of another classification model, a binary classification model comprising vector representations and information about the tweets, to forecast the movement of a stock. The dataset used here is the sentiment140 dataset. They preprocess the dataset and use tokenization to generate tokens used for predictions. For sentiment analysis, they have used a pre-trained BERT model. The first approach performs well, but the second approach, building a prediction model, was implemented due to the different challenges faced and certain limitations.

The paper published by Haizhou Qu and Dimitar Kazakov addresses the problem of prediction using news content. They have forecasted the stock price of 27 openly traded stocks and the recent news about all these organizations acquired from Yahoo Financial News Dataset. They used two metrics, one to find differences between two time series and another to amplify the distance between these two news items. To evaluate the correlation with the mantel test.

Our work differs from the rest as we focus on stocks that have shown price movements that aren't triggered by changes in their fundamentals but due to public sentiment on social media. We hypothesize that these stocks, especially cryptocurrencies, correlate highly with public opinion and not just their underlying fundamentals.

III. Data Collection and Preprocessing

This study requires us to train a sentiment analysis model for stock market related texts. It requires for the model to understand words like “bullish”, “bearish”, or even phrases like “to the moon”, which is generally used to suggest an up movement of a stock. The best way to train a model to semantically understand such tweets is to use twitter stock data itself and not general sentiment analysis text. For this study, we fetched and labeled tweets with their sentiment, where 0 represents a negative sentiment and 1 represents a positive sentiment. After a sentiment prediction model is ready, it is used with real-time twitter data, for which we used twitter API. For each respective date, selected stock's price data is also fetched.

Since twitter text in its raw form can be difficult to understand and confuse the model, it is first cleaned and preprocessed before being fed to the model. Text preprocessing is done as follows:

1. Text is converted to lowercase.
2. Stop words are removed.
3. Twitter users generally use ${ticker} (where ticker can be $GME, $AMC, etc) when talking about a particular stock. $ticker is replaced with “stockname”.
4. Special characters are removed and replaced with empty space.
5. Text is stripped and redundant white spaces are removed.

The final text is then tokenised and converted to either bag-of-words or word embeddings.

IV. Methodology

A. *Sentiment Analysis*

In this study, we used two approaches to train a sentiment prediction model – Bag of Words and a traditional Word Embedding Model

1) Bag-of-words

Bag of words is a word representation method commonly used in natural language processing. It doesn't retain the grammar or even the order of words but only the multiplicity of words in for of a vector. Each index of the vector represents a unique word in the whole corpus.

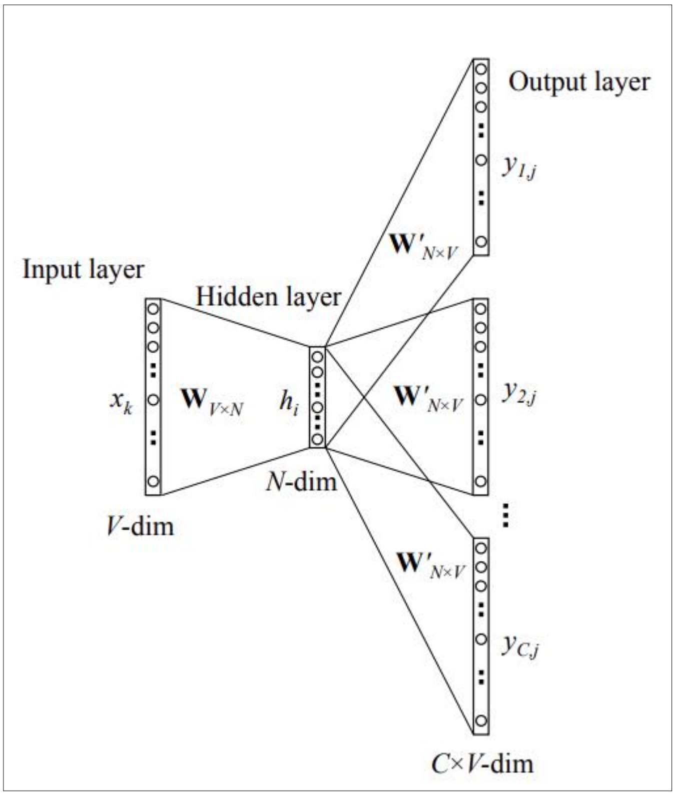

Fig. 1. Architecture of a BOW model

2) Word Embedding (Skip Gram)

Word embeddings represent words in an n-dimensional vector space. Each word has a unique coordinate in the vector space and words that are closer in the vector space are expected to have a similar meaning in a well-trained model. Word embeddings are generally used where understanding of semantics can help improve a model's performance.

Sentiment analysis can be very field specific. Using movie reviews data or Amazon reviews data won't work as well when used on stock market tweets. For word embeddings, using wikipedia based or general english based models also won't work well due to frequent use of informal words on twitter. To tackle this problem, we used twitter based word2vec models. Sentiment prediction problem can be boiled down to a classification problem where a tweet has to classified as either positive or negative for a ticker. We experimented with decision tree models, logistic regression models, and even neural networks. Each model was tested with bag-of-words and word embeddings to ensure the most optimal combination is selected for further testing.

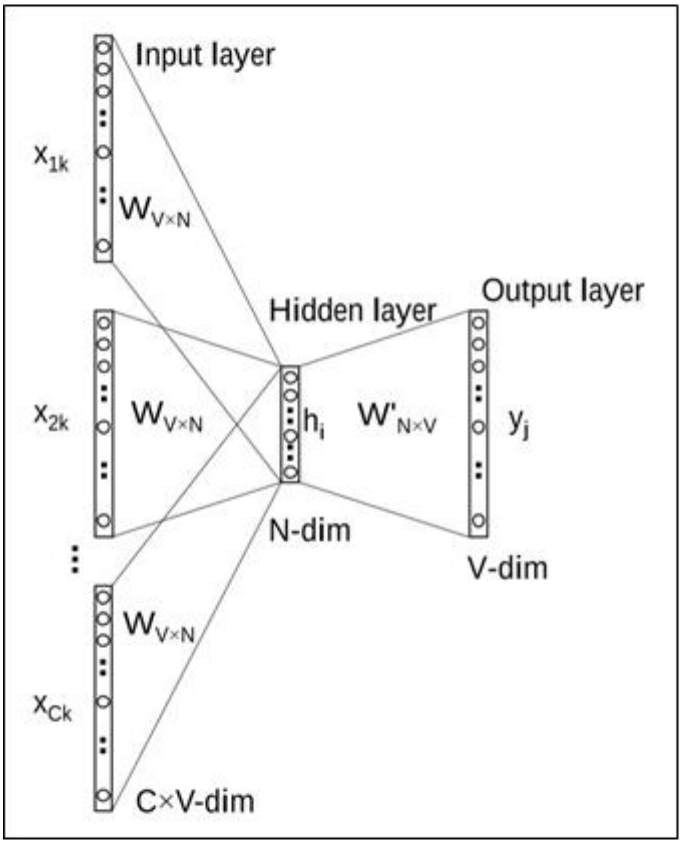

Fig. 2. Architecture of a Skip Gram model

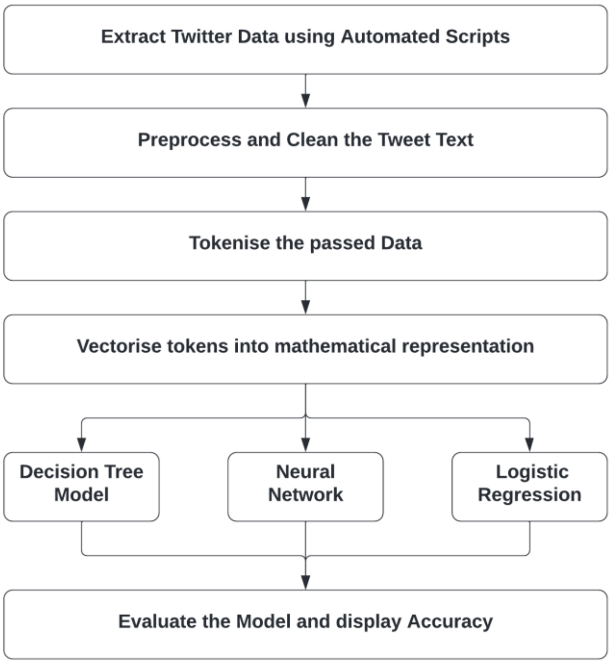
Extract Twitter Data using Automated Scripts
Preprocess and Clean the Tweet Text
Tokenise the passed Data
Vectorise tokens into mathematical representation
Decision Tree Model
Neural Network
Logistic Regression
Evaluate the Model and display Accuracy

Fig. 3. Algorithm for Sentiment Analysis

B. *Model Training*

We used scikit-learn python library to train the models, and tweaked parameters to get optimal results. The data is passed either as a bag-of-words vector or an average-of-words vector generated using gensim based pre-trained twitter-glove model. Table 1 shows the accuracy of each model:

6. MODEL PERFORMANCE

Vector Type	*Logistic Regression*	*Decision Tree*	*Simple Neural Network*
Bag of Words	0.6826	0.6608	0.6921
Word2Vec	0.7113	0.5693	0.6941

Performance shown as Accuracy

C. *Pearson's Correlation*

The Pearson's Correlation, or Bivariate Correlation is simply a measure of the linear similarity between two sets of data. This measure can have a value from -1 to +1 implying totally inverse correlation to total direct correlation. A measure, or coefficient of 0 implies no correlation between data. We use this measure to test our various models against each other.

D. *Correlation Analysis*

For analysis, Twitter data for stocks is fetched using the Twitter API. Each tweet is analyzed for its sentiment and finally, an average sentiment of all tweets is calculated for a particular day. This is done for individual days. Stock price data for a day later is fetched for the day, since retail investors' sentiment before the market starts trading will affect the prices on that day. Finally, we compute the Pearson's Correlation (as mentioned above) between the Stock Ticker Data and the prices predicted by our model for various equity or cryptocurrency-based securities.

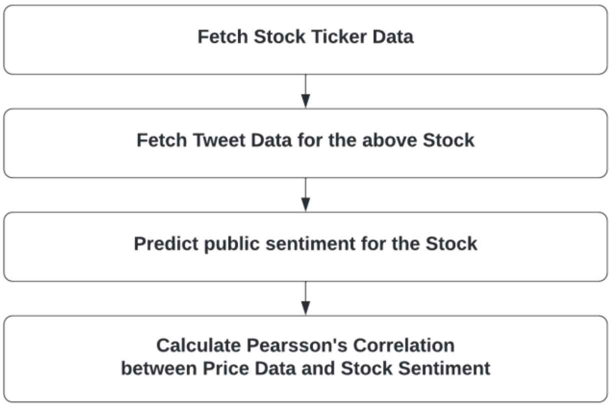
Fetch Stock Ticker Data
Fetch Tweet Data for the above Stock
Predict public sentiment for the Stock
Calculate Pearsson's Correlation
between Price Data and Stock Sentiment

Fig. 4. Algorithm of Evaluating our Results

V. RESULTS

Tickers for some of the most popular and volatile meme stocks and cryptocurrencies were analyzed. The results suggest a low-to-moderate correlation between stocks and its sentiment but a moderate-to-high correlation between cryptocurrencies and public sentiment on Twitter.

7. CORRELATION FOR VARIOUS SECURITIES

Ticker	*Pearson's Correlation*
$TSLA	0.5782
$GME	0.4933
$AMC	0.1257
$AAPL	0.2785
$AMZN	0.3672
$GOOGL	0.2381
$BTC	0.4825

Ticker	*Pearson's Correlation*
$TSLA	0.5782
$ETH	0.5661
$LUNA	0.9015
$ADA	0.6776

Correlation calculated for the past week due to API restrictions

VI. Conclusion

In this paper, we have designed and compared models to determine whether a given security belongs to the class of "Meme" Stocks i.e. whether its value is principally dependent on public sentiment, rather than fundamental value. Our main contribution is proving the effectiveness of such a model, using word2vec technologies to detect sentiment, rather than simply using word-lists.

Another interesting conclusion we can reach is that, even while considering the worst of the proven meme stocks - $AMC and $GME, we get a lower than expected percentage of value attributed to sentiment. While this might be due to the time since the r/wallstreetbets short squeeze, it is still fairly high amongst other equity stocks.

At the same time, we notice a lot of the value of cryptocurrencies being derived from their public sentiment (with the examples of $ADA and $LUNA shown above), with some deriving close to 90% of their value from their reputation alone.

VII. Future Work

This study has shown promising results and could be further extended to taking a longer time frame, including factors like following and influence of the people talking about these securities (As in 2021, Elon Musk's tweets on cryptocurrencies and stock market highly affected the prices) and the amount of attention a given security is getting.

Acknowledgment

We express our gratitude and regards to our college, Vishwakarma Institute of Technology and its faculty for the guidance, constant encouragement received throughout the course of developing this project.

Furthermore, the project would also not have been possible without the guidance and support of Prof. Dr. ST Patil who helped us while working on this project.

References

[1] H. Qu and D. Kazakov, "Quantifying correlation between Financial News and stocks," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 2016, pp. 1-6, doi: 10.1109/SSCI.2016.7850021.

[2] Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., Weber, I.: Web search queries can predict stock market volumes. PLoS ONE 7(7), e40014 (2011)

[3] Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Compu-tational Science, 2(1), 1-8 (2011)

[4] Hu, Kevin, Daniella Grimberg, and Eziz Durdyev. "Twitter Sentiment Analysis for Predicting Stock Price Movements."

[5] P.-F. Pai, and C.-S. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting, Omega, vol. 33, no. 6, pp. 497–505, 2005.

[6] D. Duong, T. Nguyen, and M. Dang, "Stock market prediction using financial news articles on ho chi minh stock exchange," in Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, ser. IMCOM '16. New York, NY, USA: ACM, 2016, pp. 71:1–71:6.

[7] Costola, Michele, Matteo Iacopini, and Carlo RMA Santagiustina. "On the "mementum" of Meme Stocks." Economics Letters 207 (2021): 110021.

[8] H. D. Huynh, L. M. Dang, and D. Duong, "A new model for stock price movements prediction using deep neural network," in Proceedings of the Eighth International Symposium on Information and Communication Technology, ser. SoICT 2017. New York, NY, USA: ACM, 2017, pp. 57–62.