COLUMBIA UNIVERSITY
School of Professional Studies

Project Report

WallStreetBets Tech Waves:

Sentiment Meets Stocks

By:

Ayushi Khasnobis (AK4825)

Ishu Jaswani (IJ2243)

Rishika Todi (RT2866)

Shivangi Mathur (SM5287)

APAN 5205: Applied Analytics Frameworks & Methods II

Professor: Dr. Vishal Lala

INTRODUCTION

The project aims to analyze the sentiment of discussions on the subreddit r/wallstreetbets, a popular online community of investors and traders, towards top ten technology stocks. By utilizing sentiment analysis & text mining tools, the project will examine the overall impact of the subreddit's discussions on different technology stocks over a specified period. The project seeks to identify trends in sentiment towards these stocks and provide insights into the potential impact of online discussions on stock market behavior. This study can help investors and traders to make more informed decisions when trading technology stocks based on the overall sentiment of the online community.

RESEARCH PROBLEM

Is there any relationship between people's sentiments and the company's stock prices in the technology industry in the long-term?

The major goal of this project is to understand whether there is a relationship between people's sentiments and stock prices. The top ten technology companies to be analyzed are:

- Apple - AAPL
- Meta - META
- Amazon - AMZN
- IBM- IBM
- Microsoft- MSFT
- Google - GOOG
- Dell Technologies- DELL
- HP- HPQ
- Uber Technologies- UBER
- Adobe- ADBE

The project will evaluate the relationship between discussion on WallstreetBets[3] and stock prices in the technology industry. WallStreetBets is a subreddit where participants discuss stock and option trading. Sentiment can be defined as the overall attitude or opinion of individuals or groups towards a particular company or industry.

It is crucial to keep in mind that sentiment is just one of many variables that can affect stock prices, and it can be challenging to anticipate how sentiment will affect prices in the near future. Stock prices may eventually be more significantly impacted by long-term trends in macro economic conditions, industry prospects, and financial performance. As a result, even while sentiment research can be a useful tool for determining how the general public feels about a company or sector, it is crucial to take into account other variables that could have an immediate effect on stock prices.

DATA

Text Scraping (Wallstreetbets, Reddit): To extract data from Wallstreetbets, Reddit, get*Textwsb()* was used. The function takes in two arguments, dates and keywords, where dates is a vector of dates which matches the predicted dates for closing stocks and keywords is the company name to search for in the comment section of wallstreetbets. The function uses the *find_thread_urls()* and *get_thread_content()* functions from the "RedditExtractoR" package to find and extract text data for the specified company and dates.

The code then combines all the text from all the companies and presents it as a data.frame which can further be analyzed.

Long Term Stocks (Quantmod Package): To analyze the impact on the long term stock prices the Rstudio package named *Quantmod* was utilized to obtain data of the ten selected technology companies from January 1, 2019 to April 27, 2023 and the frequency was kept as weekly. The data was originally obtained with columns open, high, low, close, adjusted close stock prices and volume and as such no cleaning was required. For the purpose of performing time series analysis the frequency was weekly stocks and only close columns were utilized.

ANALYTICAL TECHNIQUES

Sentiment Analysis: Sentiment analysis is the process of using natural language processing and machine learning techniques to extract subjective information from text, such as opinions, emotions, and attitudes. The application of sentiment analysis is imperative to analyze the sentiment of r/wallstreetbets users towards stock prices of top ten technology companies. Following techniques were applied to perform the sentiment analysis on the r/wallstreetbets subreddit.

1. **UPSIDE :** Regex, short for regular expression, is a sequence of characters that forms a search pattern used to match and manipulate text. To analyze the user sentiment, we matched phrases “call” & “ put” to understand who wants to buy and sell in the market. This will give us a better idea to understand the movement of stock prices. Then the upside was calculated, which is the number of “call” minus the number of “put”.

2. **AFFIN Lexicon:** The Affin lexicon is a popular sentiment lexicon in R Studio, which is used to classify text data into positive, negative, and neutral sentiment categories. A sentiment lexicon is a list of words or phrases that have been pre-labeled with a positive, negative, or neutral sentiment score. The Affin lexicon assigns a sentiment score to each word in the lexicon, ranging from -5 (extremely negative) to +5 (extremely positive). For example, the word "profit" has a sentiment score of +3, while the word "loss" has a sentiment score of -3. Based on this, we can analyze how people perceive the company. A negative score indicates a negative sentiment, while a positive score indicates a negative sentiment.

3. **JOCKER Lexicon:** The jocker lexicon assigns positive or negative sentiment scores to words based on their semantic meaning. In R, the jocker package provides access to the jocker lexicon and allows us to perform sentiment analysis on text data using this lexicon. The jocker_sentiment() function takes a vector of text documents and returns a numeric vector of sentiment scores for each document based on the jocker lexicon.

   For example, the text "I love apples" would have a positive sentiment score, while the text "I hate spiders" would have a negative sentiment score. By calculating the sentiment

score for each document, we can determine whether the overall sentiment expressed in the text is positive or negative.

Time Series Analysis

Time series analysis is a suitable analysis technique for stock prices of technology companies because it allows investors to identify patterns and trends in the historical data that can be used to predict future movements in the stock price. Time series data is a collection of observations made over time, and the patterns and trends present in the data can be used to make informed decisions about future investments.

Time series analysis was performed on stock prices which had a frequency of weekly. A total of eight time series models were performed on each of the ten technology company stocks. There were no additive or multiplicative models that were performed because the frequency of the data was too high. The aim of the time series analysis was to find a suitable model for every company individually and root mean square error was used as the key metric to determine the most suitable model for every company. Then models with the lowest RMSE for each company will be used to forecast and obtain an average return for each company. Below is the list of models that were evaluated and their descriptions:

Models Description

1. **Average Method:** The forecast for the next period is simply the average of the historical data. This is a very basic approach and is often used as a benchmark to compare with other, more sophisticated, forecasting models.
2. **Naive Model:** A simple forecasting model that uses the last observed value as the predicted value for the next time period. The naive() function seems to implement this method.
3. **Seasonal Naive Model:** A variant of the naive model that accounts for seasonality in the data. The snaive() function seems to implement this method.
4. **Drift Model:** A simple model that assumes the time series follows a random walk with drift. The rwf() function seems to implement this method.
5. **Simple Exponential Smoothing:** A basic model that uses a weighted average of past observations to forecast future values. The ses() function seems to implement this method.
6. **Holt's Method:** A more advanced method that extends the simple exponential smoothing model by also accounting for trends in the data. The holt() function seems to implement this method.
7. **Holt's Method with Damping:** A variant of Holt's method that adds a damping parameter to reduce the impact of extreme forecasts. The holt() function with the damped argument set to TRUE seems to implement this method.
8. **ARIMA:** Automatic Model Selection: A popular time series model that accounts for both autocorrelation and seasonality in the data. The auto.arima() function seems to implement this method and automatically selects the best model for the given data.

The model that gave the least train and test RMSE was with ARIMA Automatic Model Selection for nine out of the ten technology companies. The anomaly was Uber Technologies which gave a better train and test RMSE with the Drift Model instead of ARIMA Automatic Model Selection.

RESULTS

Text Analysis: These are the sentiment scores obtained from analyzing text data for each company using the three lexicons.

	Average Sentiment Score		
	Affin	**Jocker**	**Upside**
AAPL	0.3578278	0.10981758	-193
META	0.3478504	0.11095766	64
AMZN	0.3129013	0.10845743	-43
IBM	0.2241294	0.08724435	-25
MSFT	0.3637148	0.11969402	-61
GOOG	0.3359110	0.11463593	-111
DELL	0.4383658	0.13535255	-14
HPQ	0.3148297	0.07143325	-19
UBER	0.2982735	0.10843180	-8
ADBE	0.4461704	0.11819464	-27

NOTE: Upside = Calls - Puts

Based on the analysis of the text, it was determined that the sentiment expressed was mostly neutral. This implies that both the buyer and the seller have equal levels of influence and bargaining power within the market.

Time Series Analysis: These are the root mean square errors of the most suitable model based on the RMSE obtained for each company.

	RMSE & MODEL SELECTION		
	RMSE TRAIN	**RMSE TEST**	**MODEL**
AAPL	4.504397	4.504397	ARIMA
META	12.524706	12.524706	ARIMA
AMZN	7.300725	7.300725	ARIMA
IBM	4.905945	4.905945	ARIMA
MSFT	8.802332	8.802332	ARIMA
GOOG	6.917067	6.917067	ARIMA
DELL	2.087318	2.087318	ARIMA
HPQ	1.483804	1.483804	ARIMA
UBER	2.755170	2.755170	RWF
ADBE	21.485961	21.485961	ARIMA

NOTE: Upside = Calls - Puts

The ARIMA model produced the lowest RMSE values for all the tech companies except for UBER, which had the lowest RMSE value from the drift model model.

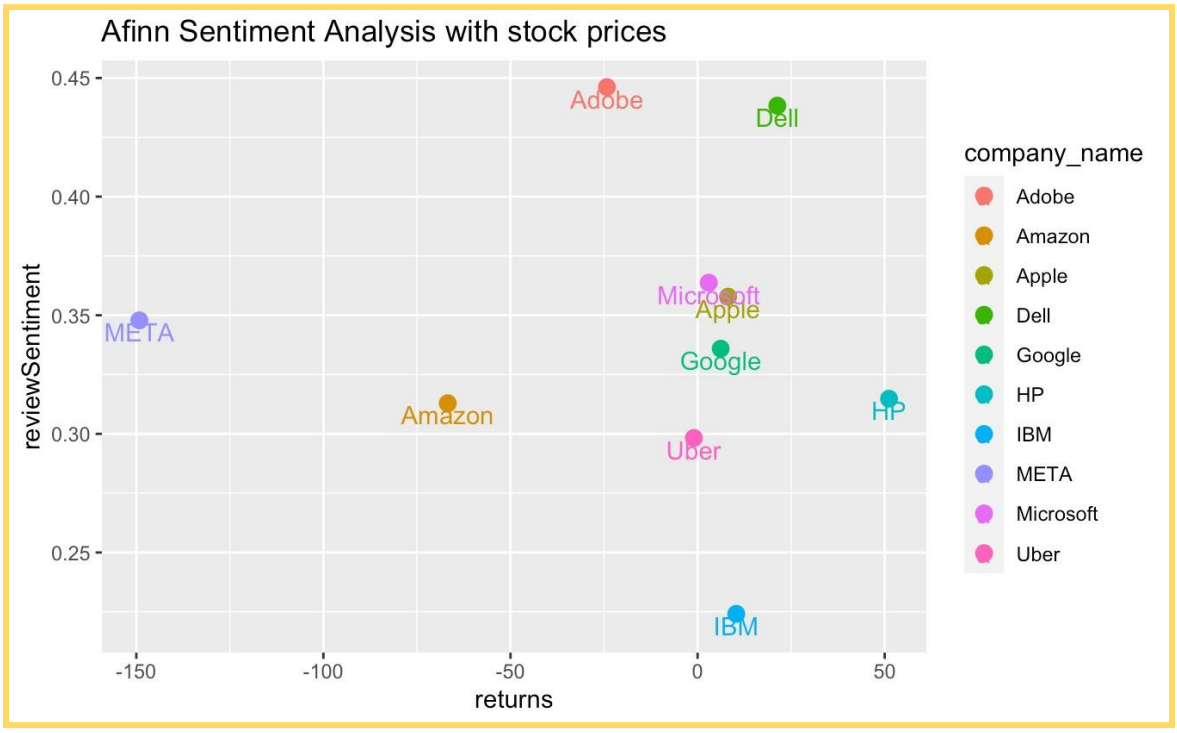
CONCLUSION & RECOMMENDATIONS

Conclusion:

After plotting the average review sentiment and average return price for all the ten technology companies, it could be concluded that there lies some impact of people's sentiment on stock prices and there needs to be more alternative data which needs to be considered eg: Economics Data and Financial Statements data

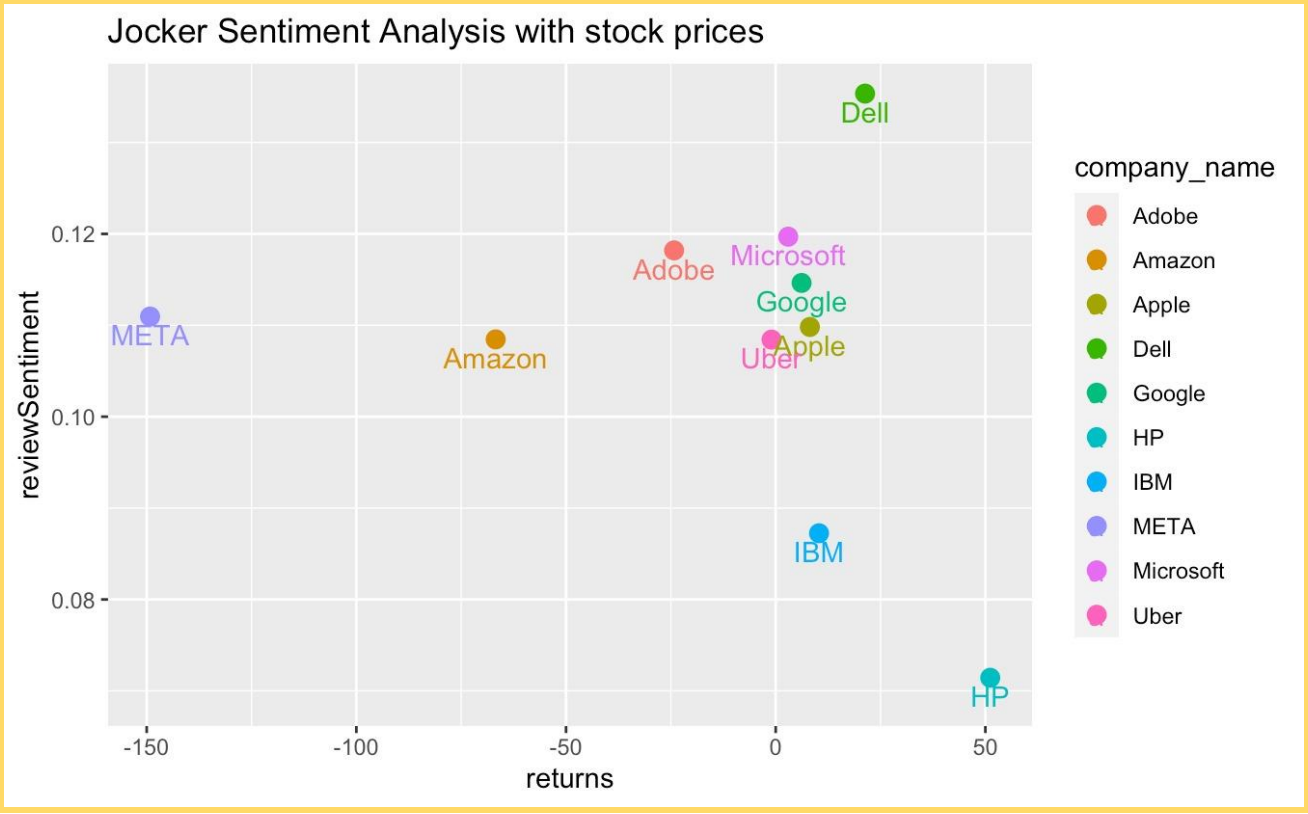
Affin Sentiment Score

In the graph, Dell has the highest review sentiment score and positive returns. IBM has the lowest sentiment score but still positive returns. This shows that there was more power in the buyers who are active on wallstreetbets and they proved to be right.



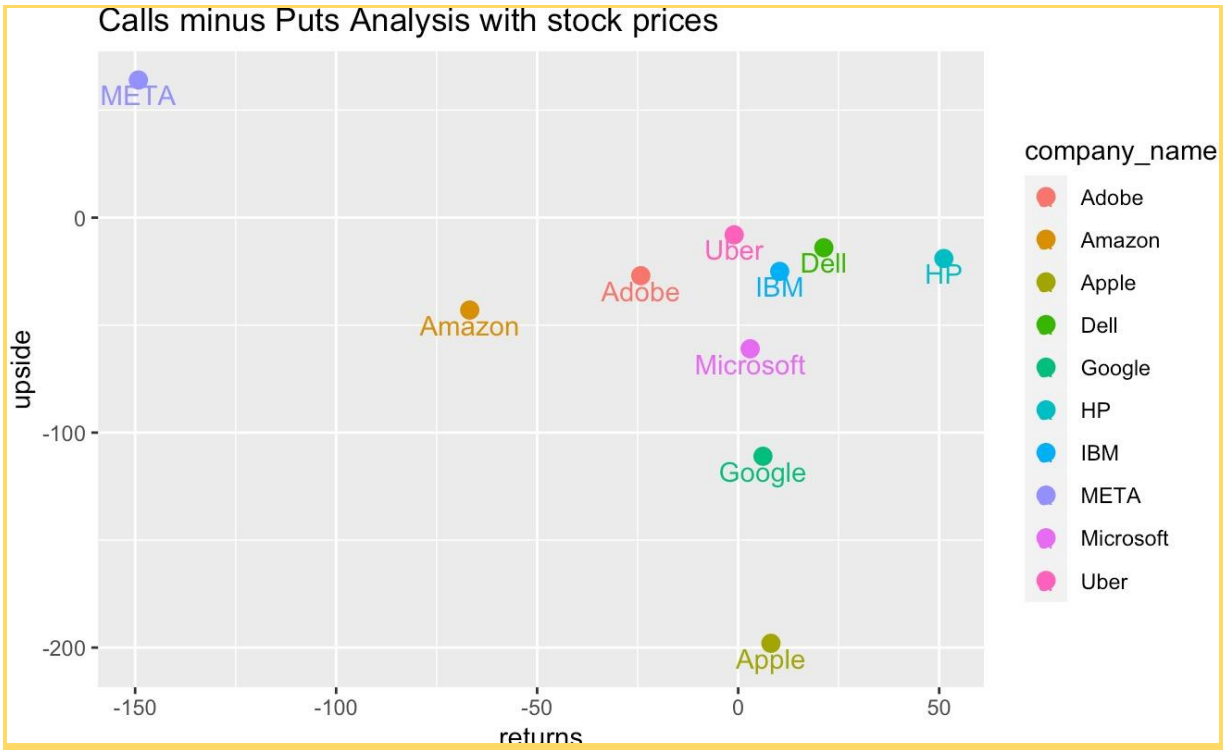
Jocker Sentiment Score

In the graph, Dell has the highest review sentiment score and positive returns. HP has the lowest sentiment score but still positive returns. This shows that there was more power in the buyers who are active on wallstreetbets and they proved to be right.



Upside Sentiment Score

In the graph, Uber has the highest upside and positive returns. Apple has the lowest upside but still positive returns.



Future Recommendations:

- Attempt to obtain weekly sentiments and juxtapose them against weekly closing prices to gain a more detailed understanding of how sentiments are impacting stock prices.

- Add financial statements data and other alternative financial data such as economic data to better understand the reason for stock prices momentum.

- Deploy data engineering techniques to optimize the speed at which we can get results so that traders can too make use of this analysis.

References

1. Jayaram, N. (2022) *"The Rising Power of the Individual Investor: How Social Media Sentiments and User Activity Impact Stock Price Volatility and Trading Volume"* CMC Senior Theses. 2873. https://scholarship.claremont.edu/cmc_theses/2873

2. *Investigation of the impact of financial information on stock prices ...* (n.d.). Retrieved April 28, 2023, from https://www.abacademies.org/articles/investigation-of-the-impact-of-financial-information-on-stock-prices-the-case-of-vietnam-7223.html

3. *Investigation of the impact of financial information on stock prices ...* (n.d.). Retrieved April 28, 2023, from https://www.abacademies.org/articles/investigation-of-the-impact-of-financial-information-on-stock-prices-the-case-of-vietnam-7223.html