Mining Reddit for Stock Valuation Signals

Matthew Mackowski
mmackowski3

Oscar Goodloe
ogoodloe3

Daniel Villaveces
dvillaveces3

Juan Antonio Garcia
jgarcia353

Xu Anne Zhang
xzhang3008

Jake Robertson
jrobertson74

1 INTRODUCTION

More people are using the internet today than ever before, and sometimes surprising outcomes are realized offline by people exchanging ideas online. In the early months of 2021, a group of people posting on the social media site Reddit encouraged each other to buy and hold shares of stock in the company GameStop, which triggered massive swings in the stock price for months. These events and others like it have their roots in people exchanging ideas online, and serve as the inspiration for our project.

2 PROBLEM DEFINITION

The vast quantity of information available online poses an interesting problem: how can a person derive insight from so many voices? We will analyze commentary from Reddit and its effects on various financial securities mentioned. Our goal is to create an interactive dashboard to allow users to explore stock price behavior in relation to current events. Precisely, Reddit provides "scores" from their community that indicate how popular a comment has been, which we aim to combine with sentiment analysis in order to examine the correlations with stock prices at a given time.

3 LITERATURE SURVEY

The idea that asset markets are influenced by social dynamics is not new, with economists examining the impact of social dynamics on investment behaviors [17]. The internet has given unprecedented volume and speed to this information, and there have been successful endeavors to use social media data to predict many events, from political elections to automobile sales [3]. The public forum postings of Reddit make it a rich source of social media data on a myriad of topics, the text of which has some expected characteristics (such as text data written in an informal style) as well as some that are more particular to Reddit (e.g., forums for specific topics are denoted as r/topic_name) [19]. Sentiment analysis is a fundamental component of successfully predicting the aforementioned events [3], and we plan to incorporate this type of analysis in conjunction with financial data to predict market trends.

Studies of Reddit comments point to the importance of how the comments are aligned within the social hierarchy of the site as well as the broader social context [8]. Reddit is organized into "subreddits" for specific interests, and Horne et al. found that similar communities may display key differences in how they process a new event as a collective, and what factors towards which they gravitate. A common algorithm for determining topics in text is latent dirichlet allocation (LDA) that has been successfully able to model topics for further refinement in multiple studies and will likely be employed in our research [6] [12] [10]. This notion of different topics gaining prevalence in similar communities is reinforced by further research into the language used in subreddits [6]. Ferrer et al. note that having a human interpreter in their process is crucial for understanding the data, as they were able to computationally model topics but not produce the explanatory social context. Thus, in our project we will carefully consider which subreddits we choose to include in our data set and their unique community features.

Sentiment analysis will be a critical component of our project. This type of analysis can be conducted on virtually any text, from hotel reviews to social media posts similar to those on Reddit [4]. Research by Renault points to finding success with data sets of 100,000 to 250,000 messages which we aim to collect from Reddit, and also indicates that the inclusion of emojis and punctuation increases accuracy [16]. Importantly, Renault also found that more complex models do not always produce better results: Naïve Bayes and Maximum Entropy were both more accurate and faster to train than more complex methods like Random Forests and Multilayer Perceptrons in his analysis. Jelodar et al. took a different approach by using a recurrent neural network with long short-term memory to achieve better testing accuracy (81.15% for their research model) compared to a logistic regression model (78.72%) or support vector machine model (77.78%), however the required training time for the recurrent neural network was not reported [10]. Modeling sentiment throughout time to examine a current state of affairs has been done [12], however we intend to take this concept further by adding in a predictive step. This type of prediction has been done with cryptocurrency and cryptocurrency-focused subreddits, with researchers achieving over 70% accuracy when forecasting directionality of price changes [20], and we hope to improve on this by applying it to more conventional market instruments.

Reddit is not the only social media site whose users have been the subject of such study. There have also been successful endeavors to use posts on Twitter to predict whether a stock would rise or fall the following day which we hope to refine by using more specific community-based discussions on Reddit rather than the broader, free-form posting that occurs on Twitter [15]. Moreover, Twitter data has been used to model public opinion by gauging sentiment on topics that are the being discussed, and responses from users can also be used to predict further details if the users have been labelled (e.g., an article can be inferred to be conservative if conservative users post about it) [7]. Even events such as which films would win Oscar awards have been successfully modeled using data from Twitter or the Internet Movie Database [2]. This gives a basis for our research as we look to identify completely new trends; given the success of using a wide variety of texts, we can expect that we will enjoy some success.

Another major component of our project is an analysis of the stock market and how it is influenced by outside events. Markets, including the stock market, are arenas of social interaction and do not exist outside the world where other events may have material impact on them [1]. News describing relevant events, augmented with more traditional stock market indicators such as momentum and rate of change, have been able to achieve better prediction accuracy than models that did not use news data [14]. This alternative (i.e., non-financial) data has been used before to help large institutional investors understand risks [13]. Our project was inspired by the short-squeeze strategy involving the company GameStop and the posts associated with it on Reddit; more positively-toned discussions were observed to correlate with higher stock returns [9]. Research has proposed that Reddit has the power to herd investors to certain stocks or scenarios, and how the impact of one such event (e.g., GameStop gaining attention) may affect unrelated stocks (e.g., AMC) [18]. Individual investors have been observed to behave differently from institutional investors, and when they do so in a great enough block they can have real impacts on the stock market [11]. These individuals who may not have the funds to wait for a market to recover have historically been prone to overreacting to short-term news [5], and we aim to find these instances before they occur.

4 PROPOSED METHOD

Our proposed product is a multi-visual dashboard that will allow users to filter for stocks, time frames, and possibly topics in order to discover price signals upon which they can potentially capitalize. The visuals we plan to include are graphs displaying the volume of Reddit posts about a specific stock, a plot comparing stock price to weighted sentiment over time, a breakdown of sentiment by subreddit, and a table of quick facts like correlation for the selected stock and subreddit(s).

We believe our approach is intuitively better than the current state of the art because:

- Real-time calculations are built into the dashboard and will allow users to interact with components of the dashboard to drill-down and filter other components in an interactive and easy-to-use manner such that those without a technical background can easily benefit
- Combination of predictions from logistic regression, sentiment analysis scores, topics derived from LDA, and a coherent visualization of these components in one place will allow users to rely on the models, the visualizations, or both to support decisions
- Prior research has focused on individual pieces, and we are assembling them to bring together analysis of text data from Reddit with stock prices
- Our visualization will provide additional social context to investors, particularly individuals who

are buying securities on their own and do not have access to teams of analysts like hedge funds and mutual funds
- Use of advanced machine learning techniques to automatically extract topics from social media comments, significantly reducing the time required to read large sets of comments that might not be of interest

Our main methods include *Sentiment Analysis*, *Topic Modeling*, and *Logistic Regression*. Below we have a detailed description of these methods along with the data collection process:

4.1 Data Collection

A significant portion of this project is the collection of data, which has been facilitated by using the Reddit application-program interface. A major hurdle to our project was a limitation imposed by Reddit: only 100 parent posts can be collected at a time. This was overcome by collecting the child comments recursively; while the limit on parent posts may still have been enforced, the discussion of interest was located within the children. Thus we were able to successfully retrieve hundreds of thousands of comments over the course of two weeks from hundreds of parent posts.

4.2 Sentiment Analysis

We have used the Python TextBlob package to derive a score for the sentiment of a given comment. We can produce both a rating of sentiment (termed polarity) as well as subjectivity for any given comment. Despite the ease of procuring such ratings, we still needed to evaluate the results. One of our weekly group meetings had time dedicated to reviewing the output together, and we found that the output was satisfactory for our purposes but had flaws. Further refinement by removing words that do not have a large impact on the meaning of a sentence (stop words) served to improve the accuracy of these scores.

Our final sentiment analysis model incorporated the TextBlob package's sentiment scoring with the score data from the Reddit posts. By multiplying these values together we infer a community sentiment about a particular idea: a negative comment with negative votes implies that the Reddit community generally supports a given topic, and likewise a negative comment with positive votes implies that the community does not support it. This heuristic for each post is then summed for all posts in a given day to indicate the overall sentiment for the day.

4.3 Topic Modeling

We have employed Latent Dirichlet Allocation with our comments in order to examine the proposed topics. The text was pre-processed to remove punctuation, and the results of the LDA analysis were reviewed in a weekly group meeting. We found that the identified topics produced by the algorithm were in a very raw form, essentially a list of words rather than a sentence. Repeating words made it difficult to find a true difference between topics.

We were able to obtain consistently more useful topic models from LDA by including n-grams in our work. These n-grams are n-length sets of words that are passed as single entries to the LDA model in order to have the model consider phrases rather than singular words. Without supplying n-grams to the model, the results were not very descriptive or coherent, often consisting of a set of words that had no obvious connection and mostly focused on entities like "Russia" or "(Tesla CEO) Elon Musk". The n-grams augmented our results and caused the result sets to have a more clear set of outputs, such as grouping "puts" and "calls" together in a single topic that indicate that stock options were a heavily-discussed topic.

4.4 Logistic Regression

Enabling users to compare this information is a great step forward, however we felt it would also be helpful to include predictions indicating if the stock is likely to increase during the next trading day based on the information at hand. We modelled the rises and falls in stock prices as they correspond to the Reddit post data combined with the sentiment analysis, in order to provide users a prediction of whether a stock is likely to rise or fall. To evaluate these predictions, we will compare them to real stock prices at later points in time. The logistic regression model is trained on the weighted sentiment score described above, the total number of Reddit awards received, and the number of posts to predict an increase in price the next day.

We have chosen to use a strategy that involves retraining the model with each filtering. This ensures that the model is using the best-available data and is also able to respond to shifting trends as sentiment and stock prices change. Importantly, we provide some easy to interpret evaluation metrics, including overall accuracy, precision, and recall so that users can make a quick determination of whether or not the model is providing meaningful results. Because the models are retrained in response to any filtering, the accuracy of a given model may be very different from any similar models. By putting the tools to evaluate the current iteration of the model in the users' hands, we empower them to make informed choices as to how good of a prediction has been produced rather than blindly trusting a black-box prediction.

4.5 Dashboard Visualization

A key component of our vision is to allow users to explore the information and find their own insights upon which they may be able to act. To enable this, we experimented with options for visualization that would support filtering and retrieving predictions based on those filters. We evaluated two distinct options: Tableau Desktop software, and a combination of Python-based solutions using matplotlib plots combined in a Jupyter notebook and arranged as a dashboard using a package called jupyter_dashboards that enables users to create a tiled arrangement of graphs and outputs from the cells of a standard Jupyter notebook.

Both tools enable end-users to apply filters to multiple fields of data and show the results, which is a critical component for our dashboards. Tableau features a native tool tip function that automatically displays details about a graph as a user hovers their mouse over it which is a nice feature to have in the data exploration. However, it is unable to retrain the machine learning models we intend to incorporate as users filter the data which would require us to pre-compute and store all possible results in order to provide the topic modeling and linear regression predictions. This is infeasible to do continuously, however the jupyter dashboard's solution would allow us to combine graphs made from matplotlib along with the results of retrained machine learning models whenever the users change the filter. This flexibility makes the jupyter dashboards option much more desirable as it can quickly filter and recalculate its results rather than requiring constant maintenance to grow an ever-expanding database of possible results.

5 EXPERIMENTS/EVALUATION

The evaluation centers on how well our final product addresses the objectives of our project. We want to see how well we meet our goal of extracting stock valuation signals from Reddit comments, while making that signal accessible to an end user, even one with relatively little technical expertise.

Our main experimental questions are:

- How accurate is our dashboard at making stock predictions?
- How applicable is it in usability regarding real world events and trends?
- How user-friendly is the visualization and user interface?

During the process we wanted to also answer the following questions regarding our approach:

- How can we apply sentiment analysis within the social constructs of Reddit, where users can not only post text but also vote in favor (or against) others' posts?
- How can we determine the major topic(s) of discussion in a given timeframe to further inform the type of securities strategies employed?
- What signals to price rises are in the comments that are left by the Reddit community? How can we know to act on those signals?
- What is the best way to provide this information in a coherent, easy-to-read fashion?

Our experiments served to direct our analysis and the development of our dashboard so that we can create an informative, interactive final product. Our experimental process has largely sought to determine how well an approach works in achieving each objective in terms of scalability, accuracy, and usability.

We evaluated how well each computational method performed and experimented by changing parameters within these methods as we went along to improve their ability to meet our objectives. In particular, we made substantial adjustments in order to make sure that computations were performed in a reasonable amount of time, allowing a user to switch between different tickers without a prolonged wait. Some of the changes we made to this end included altering the LDA analysis to only examine comments made within the last three days, filtering data for our predictive model to only comments that mentioned specific tickers, and selecting logistic regression for our predictive component as it is widely known to be highly efficient, training quickly and scaling well. We evaluate the end accuracy of a model by measuring f1-score.

For our data, we initially planned on using an API wrapper and possibly even an externally hosted database of historical Reddit comments with its own distinct API in order to overcome the restrictions the native Reddit API puts on queries. However, using a wrapper or external data repository introduces potential issues with maintainability. Further, we determined in our discussions that long-range historical data was likely to contribute fairly little, if any, signal to the resolution of real-time stock valuation signal, which was the main goal of our visualization. For these reasons, we opted to query the Reddit API directly and work within the limitations that imposed.

Lastly, we experimented with different visualization tools for our final product. Initially we had planned on using Tableau for the visual component and created our first prototype of the dashboard in Tableau, but ultimately decided that with its lack of ability to do more sophisticated calculations natively (such as LDA, sentiment analysis, and logistic regression), it would be far too involved to implement the real-time functionality we wanted, especially with one of our main goals being accessibility and ease of use. With that in mind, we switched to a Jupyter notebook with matplotlib and the jupyter_dashboards extension instead so we could package the entire visualization in one component.

The ultimate result of all this experimentation is an end product that almost anyone, regardless of background, can be easily trained to use to explore ongoing investment discourse on Reddit and make informed decisions from that data.

5.1 Results and Observations

The dashboard that we have created provides a series of clear graphs that allow users to quickly toggle different filter settings to find insights in sections of the data we have collected. The features of the dashboard can be used to give users insights about specific stocks; for the sake of brevity we have limited our results to the twenty most prominently mentioned ticker symbols. The power of this dashboard is the combination of the various components in a single place, and exploration of it is encouraged. Some interesting results that our group found are discussed here.

As a whole, we evaluated the general utility of making decisions based on a forecast using sentiment through the accuracy of the individual models fit to the top 30 ticker mentions. A.5 We notice poor f-1 scores that indicate sentiment being a lagging predictor of what will happen in the future. This is inline with much literature suggesting the nature of retail investors being reactive and emotional which shines even more on the internet. We also notice more positive sentiments in less emotional subreddits (ex: r/investing vs r/wallstreetbets). Instances with higher f-1 scores can largely be due to imbalanced datasets over the horizon we have gathered data as well as extremely poor performance of the individual stock.

Filtering for the NVIDIA (NVDA) stock ticker while keeping the full timeframe, shown in image A.1 selected yielded a logistic regression model with particularly high accuracy ($\approx$ 77%). Many of the other accuracy ratings were considerably closer to 50%, making this a noteworthy finding. The company is a major producer of graphics cards which are in high demand due to their usefulness in multiple areas of computing, and so the sentiment on Reddit may be more tightly linked to the company's fortunes than that of other companies that are the target of public figures.

Examining Google (GOOG) over the period of 3/14 - 4/12 shown in image A.2 showed opposing weighted sentiment scores for two subreddits: the stocks subreddit had a weighted sentiment score slightly above 1, whereas wallstreetbets had a weighted sentiment score of about -0.6 in the same time period. This exemplifies the notion that each subreddit has a personality of sorts, attracting different users which results in differing opinions. It is difficult to say why these two subreddits had such diverging opinions, but what is clear is that there is a clear reason to be able to differentiate between these groups' sentiments.

Facebook (FB) showed an interesting pattern for the dates 2/13 to 3/14 in image A.3. There was particularly high sentiment from the daytrading subreddit, whose name suggests that people are particularly interested in profiting from buying and reselling stocks quickly rather than holding them for a long-term investment. There was also an interesting juxtaposition between the log returns and weighted sentiment on 3/9/2022, where the graph shows a spike in the log returns of Facebook's stock but also a very steep decline in the weighted sentiment score. Presumably, something happened or an announcement was made that excited some investors but perhaps made other skeptical or hostile towards the company. Within the next few days, there was a dip in the log returns with the weighted sentiment score returning to approximately neutral (0). These types of events show that the sentiment of the posts on Reddit do not necessarily line up perfectly with stock market results, although they may be leading or lagging indicators.

Our final example, Amazon (AMZN) earned a notably high sentiment score over the period of 2/13 - 3/14 shown in image A.4. The investing subreddit in particular displayed an overwhelmingly positive sentiment for the company during this time, with a weighted sentiment score greater than 8. Most of the weighted sentiment scores observed are typically between the values of -1 and 1, making this score particularly interesting. Such outliers are not necessarily due to a substantial reason, and even with the outstandingly positive sentiment there did not appear to be an outstandingly positive return on the stock until later in the month.

6 CONCLUSIONS/DISCUSSION

This project aims to holistically blend topic modeling, social media sentiment analysis, and predictive modeling of security performance to identify the impact that world events and investor sentiment have on investment markets, and explore how well such an approach can predict price movements. Such information is useful to virtually anyone interested in investing, from individual investors to hedge funds to institutional investors; beating the market is the dream, and this is potentially a way to achieve it that we can measure by comparing predictions to market results.

We have implemented topic modeling and sentiment analysis using TextBlob models, post data, and a variety of classification models in our analysis to identify patterns. Because Reddit and financial data are free, this project was completed with no cost beyond our individual effort. As with any analytics tool, there is always the risk of identifying erroneous correlations which could mislead investors, with the corresponding payoff being the possibility of above-average monetary gain from those investments.

We included an analysis of the capabilities of the final dashboard in the results section. We believe that this product is an improvement compared to current methodologies by combining the strengths of multiple forms of analysis and empowering individuals to find lucrative investments that are undervalued. The sample dashboard results are just a few of the possibilities that can be realized through this dashboard, and we would encourage any potential users to explore to find their own insights.

We discuss that sentiment may be a leading indicator in some cases, however the sentiment positivity does not necessarily manifest itself immediately, perhaps taking a few days or weeks to be realized. The effectiveness of this measure will undoubtedly vary from one stock to another, but this dashboard enables users to explore the text data in a way that they may not be able to otherwise do. We hope that adding our dashboard analysis to an investor's toolkit will help them make smarter decisions faster to achieve their financial goals.

All team members contributed equal effort throughout the project. This was truly a group effort from start to finish, with all members attending most (if not all) weekly meetings, participating in group discussions, and contributing to the final product.

REFERENCES

[1] Jens Beckert. 2009. The social order of markets. *Theory and society* 38, 3 (2009), 245–269.

[2] Efthimios Bothos, Dimitris Apostolou, and Gregoris Mentzas. 2010. Using social media to predict future events with agent-based markets. *IEEE Intelligent Systems* 25, 06 (2010), 50–58.

[3] Jaroslav Bukovina. 2016. Social media big data and capital markets—An overview. *Journal of Behavioral and Experimental Finance* 11 (2016), 18–26.

[4] G Devi and Kamalakkannan Somasundaram. 2020. Literature Review on Sentiment Analysis in Social Media: Open Challenges toward Applications. 29 (01 2020), 1462–1471.

[5] Daniel Elkind, Kathryn Kaminski, Andrew W Lo, Kien Wei Siah, and Chi Heem Wong. 2022. When Do Investors Freak Out? Machine Learning Predictions of Panic Selling. *The Journal of Financial Data Science* 4, 1 (2022), 11–39.

[6] Xavier Ferrer, Tom van Nuenen, Jose M Such, and Natalia Criado. 2020. Discovering and categorising language biases in reddit. In *International AAAI Conference on Web and Social Media (ICWSM 2021)(forthcoming)*.

[7] Barrie Gunter, Nelya Koteyko, and Dimitrinka Atanasova. 2014. Sentiment analysis: A market-relevant and reliable measure of public feeling? *International Journal of Market Research* 56, 2 (2014), 231–247.

[8] Benjamin D. Horne, Sibel Adali, and Sujoy Sikdar. 2017. Identifying the Social Signals That Drive Online Discussions: A Case

Study of Reddit Communities. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*. 1–9. https://doi.org/10.1109/ICCCN.2017.8038388

[9] Danqi Hu, Charles M Jones, Valerie Zhang, and Xiaoyan Zhang. 2021. The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. *Available at SSRN 3807655* (2021).

[10] Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang. 2020. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics* 24, 10 (2020), 2733–2742.

[11] Cheng Luo, Enrichetta Ravina, Marco Sammon, and Luis M Viceira. 2020. Retail investors' contrarian behavior around news and the momentum effect. *Available at SSRN 3544949* (2020).

[12] Chad A. Melton, Olufunto A. Olusanya, Nariman Ammar, and Arash Shaban-Nejad. 2021. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health* 14, 10 (2021), 1505–1512. https://doi.org/10.1016/j.jiph.2021.08.010 Special Issue on COVID-19 – Vaccine, Variants and New Waves.

[13] Ashby Monk, Marcel Prins, and Dane Rook. 2019. Rethinking alternative data in institutional investment. *The Journal of Financial Data Science* 1, 1 (2019), 14–31.

[14] Pisut Oncharoen and Peerapon Vateekul. 2018. Deep learning for stock market prediction using event embedding and technical indicators. In *2018 5th international conference on advanced informatics: concept theory and applications (ICAICTA)*. IEEE, 19–24.

[15] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of Twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*. IEEE, 1345–1350.

[16] Thomas Renault. 2020. Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance* 2, 1 (2020), 1–13.

[17] Robert J Shiller, Stanley Fischer, and Benjamin M Friedman. 1984. Stock prices and social dynamics. *Brookings papers on economic activity* 1984, 2 (1984), 457–510.

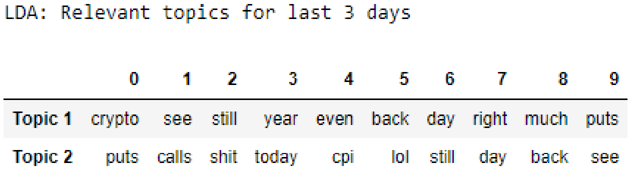
[18] Zaghum Umar, Imran Yousaf, and Adam Zaremba. 2021. Comovements between heavily shorted stocks during a market squeeze: Lessons from the GameStop trading frenzy. *Research in International Business and Finance* 58 (2021), 101453. https://doi.org/10.1016/j.ribaf.2021.101453

[19] S. Vajjala, B. Majumder, A. Gupta, and H. Surana. 2020. *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media, Chapter Chapter 8: Social Media. https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/ch08.htm

[20] Stephen Wooley, Andrew Edmonds, Arunkumar Bagavathi, and Siddharth Krishnan. 2019. Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 500–505. https://doi.org/10.1109/ICMLA.2019.00093

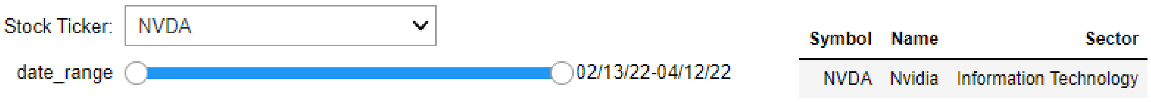
A APPENDIX

A.1 NVIDIA



```
LDA: Relevant topics for last 3 days
```

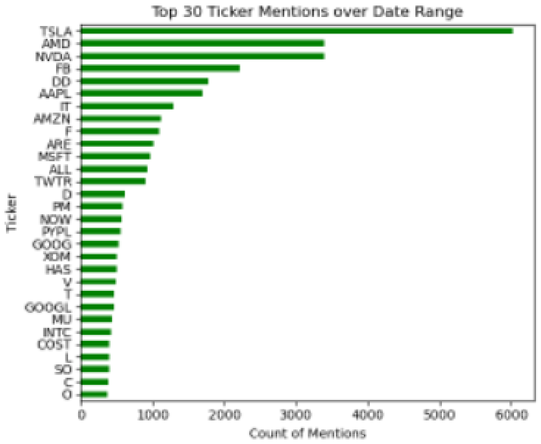
	0	1	2	3	4	5	6	7	8	9
Topic 1	crypto	see	still	year	even	back	day	right	much	puts
Topic 2	puts	calls	shit	today	cpi	lol	still	day	back	see

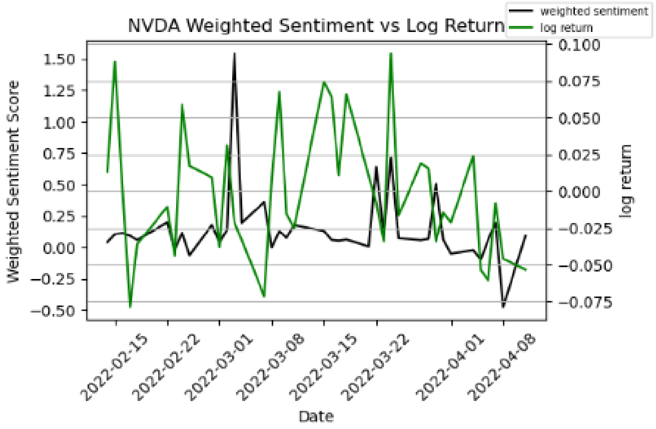


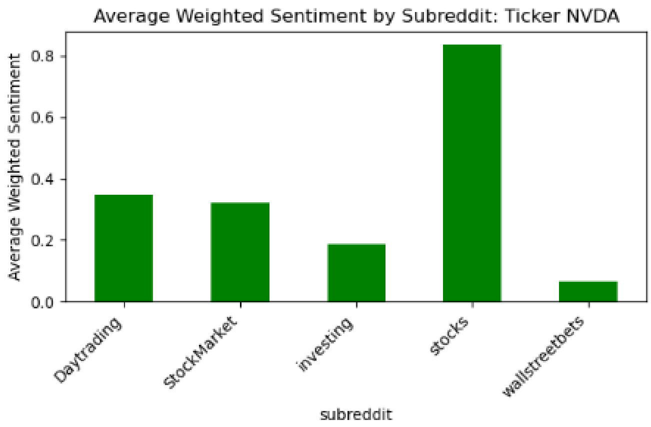
Stock Ticker: NVDA

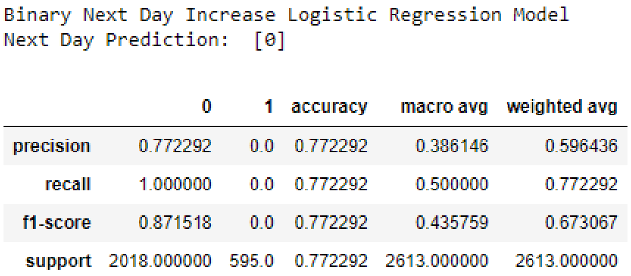
date_range 02/13/22-04/12/22

Symbol	Name	Sector
NVDA	Nvidia	Information Technology





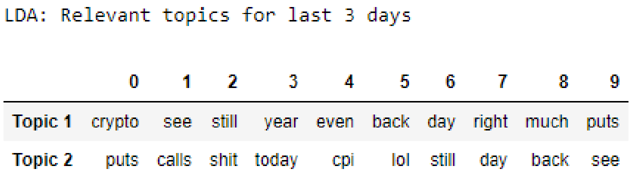




```
Binary Next Day Increase Logistic Regression Model
Next Day Prediction:  [0]
```

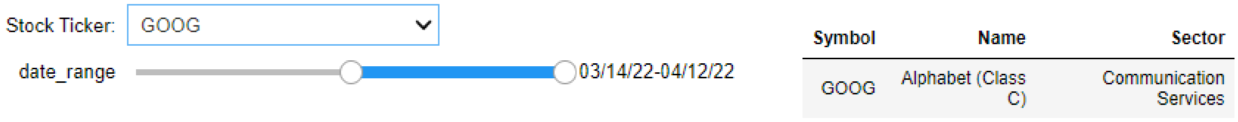
	0	1	accuracy	macro avg	weighted avg
precision	0.772292	0.0	0.772292	0.386146	0.596436
recall	1.000000	0.0	0.772292	0.500000	0.772292
f1-score	0.871518	0.0	0.772292	0.435759	0.673067
support	2018.000000	595.0	0.772292	2613.000000	2613.000000

A.2 Google



```
LDA: Relevant topics for last 3 days
```

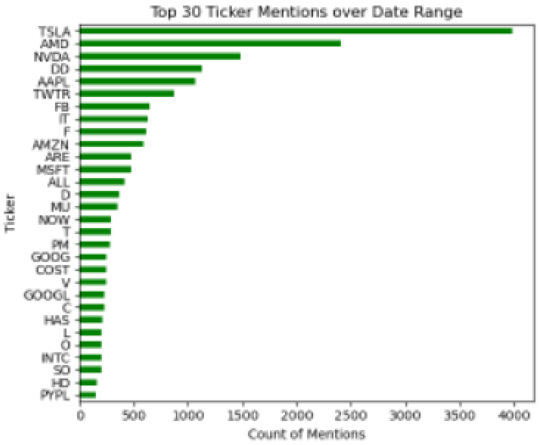
	0	1	2	3	4	5	6	7	8	9
Topic 1	crypto	see	still	year	even	back	day	right	much	puts
Topic 2	puts	calls	shit	today	cpi	lol	still	day	back	see

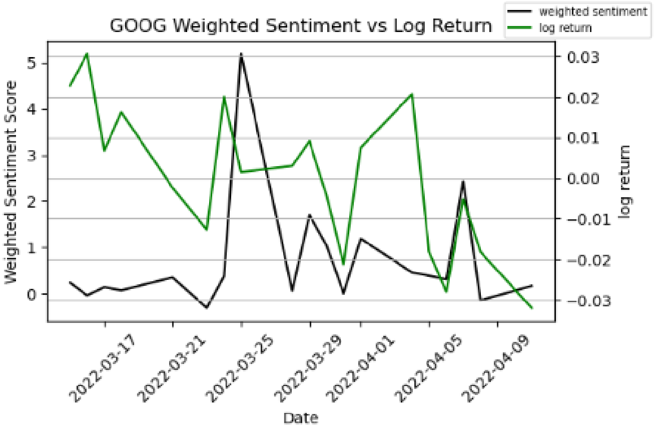


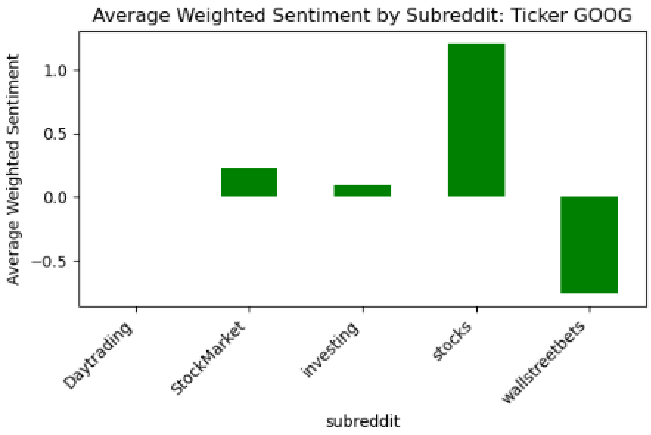
Stock Ticker: GOOG

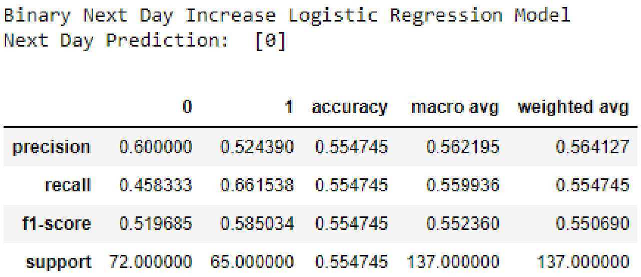
date_range 03/14/22-04/12/22

Symbol	Name	Sector
GOOG	Alphabet (Class C)	Communication Services





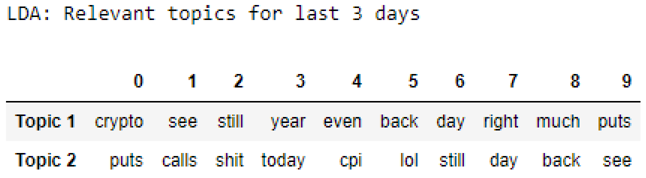




```
Binary Next Day Increase Logistic Regression Model
Next Day Prediction:  [0]
```

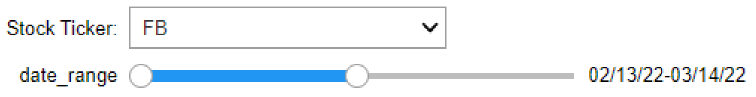
	0	1	accuracy	macro avg	weighted avg
precision	0.600000	0.524390	0.554745	0.562195	0.564127
recall	0.458333	0.661538	0.554745	0.559936	0.554745
f1-score	0.519685	0.585034	0.554745	0.552360	0.550690
support	72.000000	65.000000	0.554745	137.000000	137.000000

A.3 Facebook



```
LDA: Relevant topics for last 3 days
```

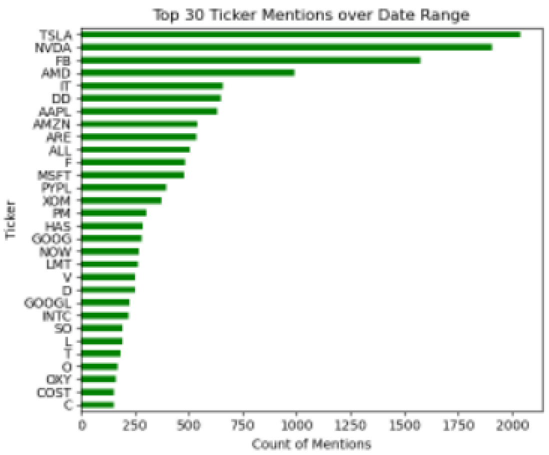
	0	1	2	3	4	5	6	7	8	9
Topic 1	crypto	see	still	year	even	back	day	right	much	puts
Topic 2	puts	calls	shit	today	cpi	lol	still	day	back	see

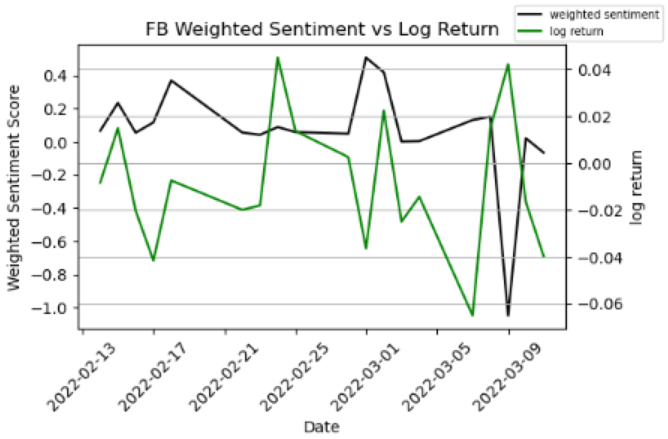


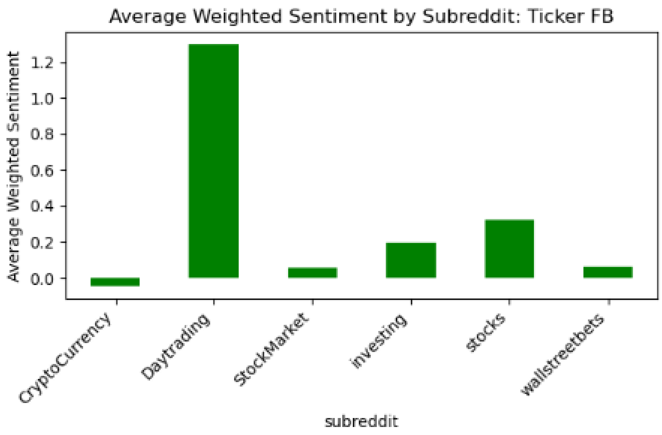
Stock Ticker: FB

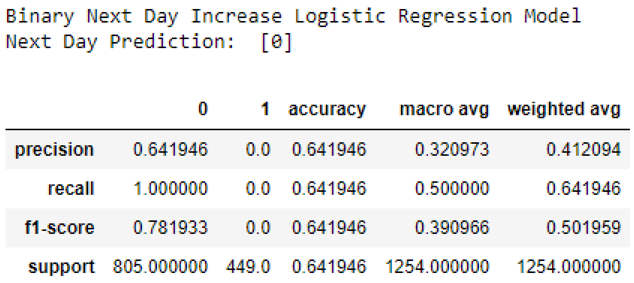
date_range 02/13/22-03/14/22

Symbol	Name	Sector
FB	Facebook	Communication Services





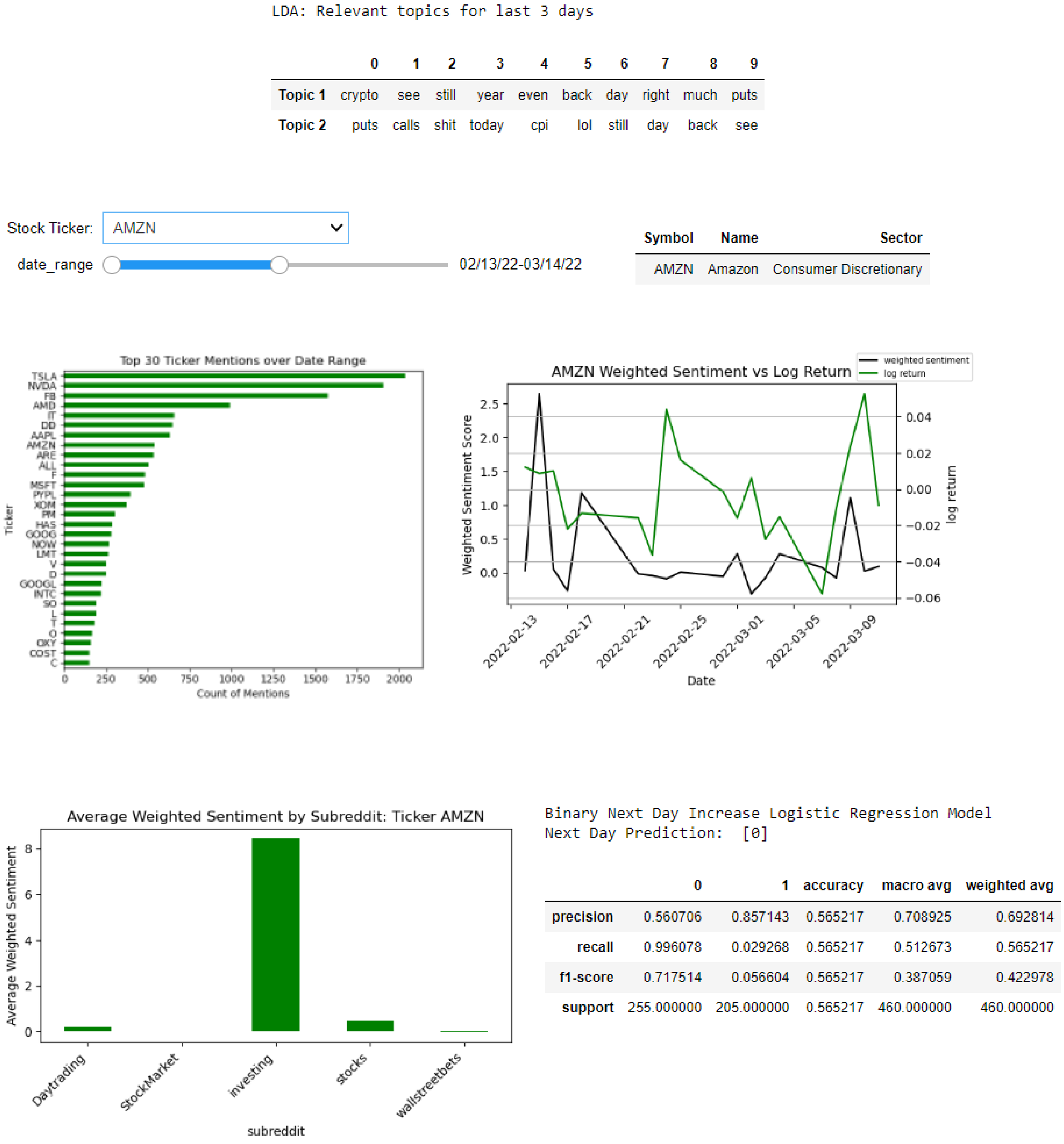




```
Binary Next Day Increase Logistic Regression Model
Next Day Prediction:  [0]
```

	0	1	accuracy	macro avg	weighted avg
precision	0.641946	0.0	0.641946	0.320973	0.412094
recall	1.000000	0.0	0.641946	0.500000	0.641946
f1-score	0.781933	0.0	0.641946	0.390966	0.501959
support	805.000000	449.0	0.641946	1254.000000	1254.000000

A.4 Amazon



```
LDA: Relevant topics for last 3 days
```

	0	1	2	3	4	5	6	7	8	9
Topic 1	crypto	see	still	year	even	back	day	right	much	puts
Topic 2	puts	calls	shit	today	cpi	lol	still	day	back	see

Stock Ticker: AMZN

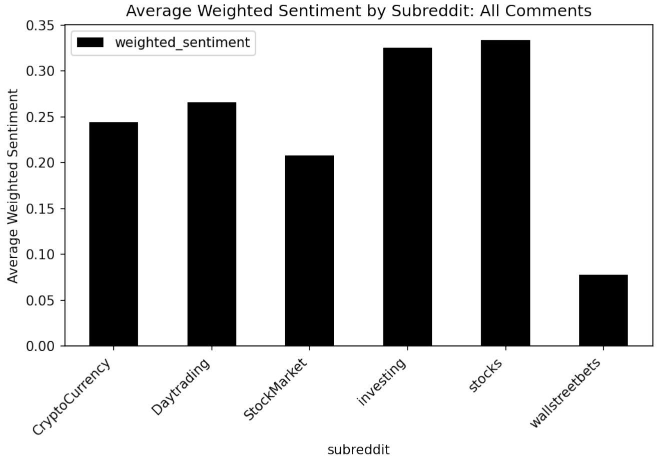
date_range 02/13/22-03/14/22

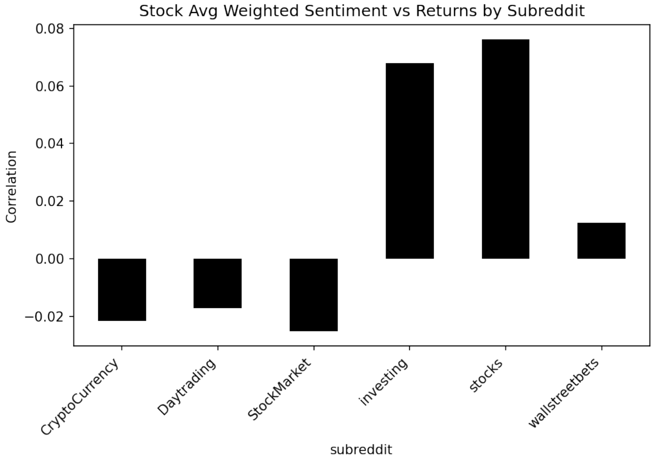
Symbol	Name	Sector
AMZN	Amazon	Consumer Discretionary

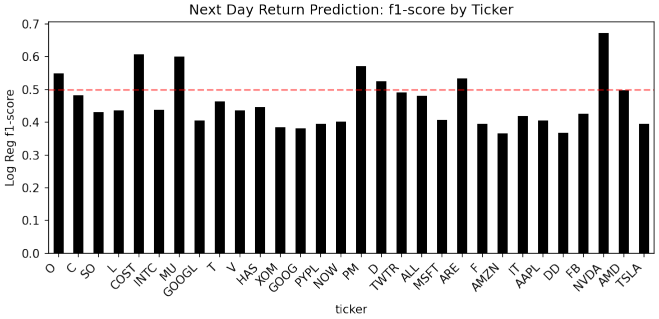
```
Binary Next Day Increase Logistic Regression Model
Next Day Prediction:  [0]
```

	0	1	accuracy	macro avg	weighted avg
precision	0.560706	0.857143	0.565217	0.708925	0.692814
recall	0.996078	0.029268	0.565217	0.512673	0.565217
f1-score	0.717514	0.056604	0.565217	0.387059	0.422978
support	255.000000	205.000000	0.565217	460.000000	460.000000

A.5 Sentiment Model Accuracy


Average Weighted Sentiment by Subreddit: All Comments


Stock Avg Weighted Sentiment vs Returns by Subreddit


Next Day Return Prediction: f1-score by Ticker