TO THE MOON: Hype Stock Price Prediction using Sentiment Analysis of Reddit & Google News

Tiffany Huang
Princeton University
twhuang@princeton.edu

Alexis Sursock
Princeton University
asursock@princeton.edu

Abstract

The GameStop (GME) short squeeze of January 2021, orchestrated by users of Reddit forum r/WallStreetBets, has brought to light the impact of public sentiment on stock prices. In this work, we investigate the relationship between sentiments expressed on Reddit and in Google News headlines on 21 different stocks, focusing especially on GME. Using Granger causality we find that there is a one-directional causal relationship between GME-related Reddit sentiment and GME stock price changes, although the same is not true for news headlines. We also use gathered sentiment to predict daily stock prices over several months, employing a neural network and several other classification and regression models. We find that the neural network outperforms other models, and works well in predicting hype and tech stock prices, and less well in more volatile, less-traded stocks. Future work should explore if extracting sentiment from other social mediums could improve prediction of these two stock categories.

1 Introduction

Can social media sentiment predict stock prices? This question has been explored heavily by scholars, most of which end up siding with the Efficient Market Hypothesis (EMH) - a theorem that states that asset prices already reflect all publicly available data and thus are hard or impossible to predict. However, the recent volatile and bubble-like stock performance of GameStop (GME) as a result of actions taken by users of the r/WallStreetBets forum on Reddit has underscored even further the role of public sentiment in market predictions. This volatility is shared by other ”hype stocks”, or stocks that were heavily favored by r/WallStreetBets, such as AMC Entertainment Holdings (AMC) and Nokia (NOK). This leads us to the question: how will public sentiment analysis perform when predicting such unpredictable stocks?

In this work, we extract Reddit posts from r/WallStreetBets for three types of stocks: hype (as defined above), established (well-known, stable stocks), and volatile (stocks that have performed unpredictably in our chosen time period of October 2020 to April 2021). We investigate seven stocks within each group. Recognizing the difficulty of analyzing informal language used on Reddit, we also extract Google News headlines about the same 21 stocks over the same period. We perform sentiment analysis on these extracted texts and input it into a recurrent neural network, along with the stock prices of the past five days, to predict stock price changes for each day. We also use other regression models along with the neural network to see which has optimal accuracy. We evaluate our models based on a self-created accuracy metric that assesses stock movement predictions. Our findings support our hypothesis that Reddit sentiments are predictive of GME stock outcomes; we see some success in predicting other hype stocks as well.

2 Related Work & Data

2.1 Literature Review

Stock market price prediction is a widely studied research area, often leading to a conclusion supporting the Efficient Market Hypothesis (EMH): that external stimuli are immediately incorporated into the price and arbitrage opportunities are rare in the ensuing random walk pattern; as a result, future stock prices are near impossible to accurately predict [12]. Some researchers, however, have demonstrated the effect of public sentiment (mostly via Twitter) on the stock market with relative success, identifying some market inefficiencies and opportunities for price fluctuation prediction [12]. However, most of them employ simplifying assumptions or unrealistic constraints: for example, instead of predicting a specific stock's price, many predict the DJIA (Dow Jones Industrial Average), a stock market index, instead [1]. Additionally, most prune their data to avoid increased variability and remove external forces [12]. We, on the other hand, find that external forces are part of the everyday stock market and hope to explore their impact on prediction.

In light of the spectacular GameStop short squeeze which was planned days in advance by Reddit users, we set out to reexamine the EMH theorem considering hype stocks. Upon starting this paper in April 2021, no research had been published specifically examining the predictability of hype stocks through Reddit sentiment. Since then, a small number of papers have done so with remarkable results, showing that, counter to the EMH theorem, Reddit activity can be predictive of future stock prices[11] [20] [21]. Upon finding these newer papers, we decided to expand our focus to remain novel by incorporating other hype stocks and contrast Reddit and Google News predictivity. Additionally, as there currently is no large and publicly available data set of labelled Reddit stock posts[1], let alone for 21 different stocks, we decided to publish our data sets to the data sharing platform Kaggle for future research[18].

2.2 Data Collection

A significant portion of our work relates to the initial data collection and preprocessing. Upon being approved, we utilized the Reddit[15], Google News [2], and Yahoo Finance APIs [3] to create our own data set. We gathered all Reddit posts on the r/WallStreetBets SubReddit in a 6 month period (October 1st 2020 - April 29th 2021) that mentioned either the company name (e.g. GameStop) or their NYSE ticker symbol (e.g. GME), the stock's unique identifier on the stock exchange, in the title.

We chose 21 representative stocks from three distinct stock groups:

1. **Hype:** Stocks whose value has increased significantly without much change in fundamental value; mostly ”hyped up” by online forums such as r/WallStreetBets on Reddit. [GameStop, AMC, Nokia, Tesla, BlackBerry, Express, KOSS]
2. **Established Tech:** Relatively stable technology firms with low corporate bond rates and high trading volumes in the S&P500 or S&PChina [Microsoft, Google, Facebook, Amazon, Apple, Alibaba, Netflix]
3. **High-Volatility:** A variety of companies that have either been hit by or benefited from the pandemic and whose stock prices have fluctuated tremendously over the last year [Moderna, Royal Caribbean, Carnival Cruise Line, Zoom, Beyond Meat, American Airlines, Boeing]

We pooled data from the 21 chosen stocks and compiled over 43,300 Reddit posts, labelled with their title, body text, upvote ratio, overall score, and the date on which it was shared[2]. Similarly, we searched all articles on the Google News platform for mentions of our stocks and extracted 86,600 article headlines over the 6 month period. Lastly, we extracted daily opening, closing, and trading

[1] Although there are data sets of Reddit posts none are stock-specific and allow only for general forum sentiment rather than sentiment towards a certain stock.

[2] Note: We initially extracted over 105,000 Reddit posts which included stock mentions in both the title and the body of the post. We found that doing so congests our data set with irrelevant posts and thus decided to solely include title mentions for a more accurate sentiment analysis.

volume data for all stocks on the NYSE (New York Stock Exchange) and throughout the interval from the Yahoo Finance database.

2.3 Data Preprocessing

The extracted data in its raw form is incompatible between data sets (Reddit, Google News, and Yahoo Finance). To ensure reliable results we set out to clean and match the data sets:

1. We initially preprocess our data set of Reddit posts and news headlines by transforming the time zone used from UTC to New York time (either EDT or EST depending on date) to adequately match our data to the New York Stock Exchange's opening and closing times.

2. Next, even though social media postings and news headlines occur every single day, stock exchanges regularly go on break for holidays and weekends, leaving us without opening and closing prices for these periods. Given the breaks, stock prices tend to jump more significantly when reopening as market sentiment builds up and changes. For reopening days, we thus do not only use the previous day's sentiment, but instead gather a weighted average of the sentiment during the closed interval for more accurate predictions.

3. Lastly, we ensure that our data is easily readable and identifiable by both our Bag of Words and Sentiment Analysis models. We remove special characters, transcribe Reddit vernacular, and convert emojis into words before using WordNet tokenization and lemmatization from the NLTK library. Our data set largely avoids spam Reddit posts by only extracting posts with an upvote-ratio of above 0.5 and a score of at least 1.

2.4 Expectations

We expect to find a causal relationship between GME-related Reddit sentiment and stock price movement. We also expect that prediction of GME and other hype stocks will be more accurate using Reddit sentiments, while prediction of tech stocks and volatile stocks will be more accurate using Google News sentiments. We will assess prediction accuracy using a self-made metric described in the Results section and plan on using Granger Causality to explore any causal relationships between our data sets.

3 Methods

3.1 Sentiment Analysis

We set out to conduct sentiment analysis on both the Reddit and Google News data sets to create a daily sentiment score for each stock. We evaluated two different natural language processing (NLP) libraries – VADER (Valence Aware Dictionary for Sentiment Reasoning) and Flair – to classify posts and headlines. VADER, which is built on a standardized sentiment lexicon for social media analysis, was chosen for both its efficiency, allowing us to avoid a speed-performance trade-off, and its outperformance when used with well-known NLP libraries, such as SentiWordNet[5]. Furthermore, VADER provides us with both a text's sentiment polarity and intensity in the form of four continuous scores (”compound”, ”positivity”, ”negativity”, and ”neutrality”), whereas Flair only provides a ”positive” or ”negative” label. We felt that using VADER would better capture the overall sentiment of a given string, leading to more accurate results when performing prediction. This was confirmed we found the Granger Causality scores (which quantify whether sentiment scores predict stock prices) for both models, and VADER significantly outperformed Flair.

To employ VADER for the Reddit data set, we calculate and store the compound, positivity, negativity, and neutrality scores for each r/WallStreetBets post's title and body text. For the Google News data set, we do the same for each headline. In order to get the overall sentiment, we find all of the posts or headlines for a given day, and sum their scores so that each day in our data set has four corresponding VADER sums. We also keep track of the number of posts or headlines for a day so we can calculate the average of each of these sums.

3.2 Bag of Words

In creating our Reddit data set, we noticed that a lot of the extracted posts included vocabulary and phrases that may not be easily understood by traditional sentiment analysis algorithms. As a result, we decided we also wanted to represent the posts using a bag-of-words model, in an effort to capture as much information about sentiment as possible. Our thought process was that by not performing any sort of analysis on text before inputting it in our prediction model, we might be able to obtain more ”pure” and accurate results. For consistency’s sake, we also created a bag-of-words representation of our Google News data set.

We chose the bag-of-words model because it is the simplest way to numerically represent text by turning it into fixed-length vectors[22]. We begin by appending each day’s preprocessed posts or headlines to create one long string of words for every day in our investigation period. We then employ the TfidfVectorizer library to create the bag-of-words representation[10]. This first uses CountVectorizer[8] to create a sparse matrix representing the presence of tokens in a string, and then employs TfidfTransformer[9] (tfidf stands for term-frequency times inverse document-frequency) to weight token importance by scaling down the weight of tokens that appear very frequently. We input both unigrams and bigrams (two-word phrases) and perform feature selection on the matrix by including only the 1000 most common tokens.

For the Reddit data set, we do this process for both the titles and bodies of r/WallStreetBets posts so that our end matrix has 2000 columns. For the Google News dataset, we only perform this for the headlines, so our matrix has 1000 columns.

3.3 Algorithms

We use a total of 5 different models, including three from the SciKitLearn Python libraries , and two deep learning algorithm (one trained on sentiment and one on BOW (bag of words)).

1. **Long Short Term Memory Network (LSTM) for Sentiment:** optimized for every stock based on number of hidden layers, hidden layer size, and the activation function. Utilized due to their strong performance on time series analysis (reasoning in 3.4 Spotlight)[16]
2. **Long Short Term Memory Network (LSTM) for Bag of Words:** optimized for every stock based on number of hidden layers, hidden layer size, and the activation function. Used to avoid abstraction when quantifying textual sentiment.
3. **Support Vector Classification (SVC):** using rbf or linear kernel, C = 1.0 - we employ them because support vector machines find frequent application in stock price direction forecasting, one of the goals of this paper.
4. **Logistic Regression:** In addition to SVC, we also employ logistic regression as a means of binary classification to see if classification works well for stock prediction.
5. **Linear Regression:** In contrast to Logistic Regression, we use Linear Regression to directly compare a standard regression model to a binary classification one for this research question.

3.4 Spotlight Method: Long Short-Term Memory Networks

Time dependent relationships are often investigated through recurrent neural networks, or neural networks that are able to take in prior inputs to achieve better learning. This is necessary in the context of stock price prediction because we want to use prior days’ patterns to inform our prediction of current stock prices. However, recurrent neural networks often run into the issue of short-term memory loss; they are unable to retain information for long periods of time.

A solution to this issue is to use a specific type of recurrent neural network: a Long Short Term Memory network (LSTM). A LSTM can ”dynamically learn and determine whether a certain output should be the next recursive input”[6]. The below figure demonstrates the structure of a LSTM. Note how the repeating module (the middle green cell) has four neural network layers (represented by the yellow boxes); a standard recurrent neural network would only have one. These additional layers allow the LSTM to prioritize keeping ”important” information while discarding ”unimportant” knowledge.

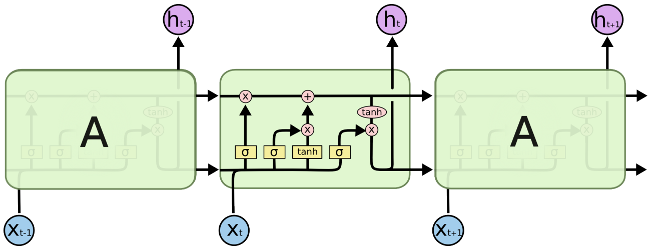

Figure 1: A visual representation of the LSTM, courtesy of [13]

A key component of a LSTM is the the cell state, represented in the figure by the black arrow near the top; this can be thought of as a ”highway that transfers relative information all the way down the sequence chain” [14]. The cell state is calculated by some of the layers, or gates, discussed earlier. In the figure, the left-most layer is a sigmoid layer called the ”forget gate layer”[13]. It takes in the x_t, the input vector at time t, and h_{t-1}, the output vector from time $t-1$, and outputs a 0 or 1 for each value in cell state C_{t-1}. Here, 0 indicates that the neural network should forget this piece of information, while 1 indicates that it should remember it [13]. In other words, at this gate, the model decides what information it should retain from its earlier steps. Then, the model decides what new information to add to the cell state. It does this through the next two layers in the figure: a sigmoid layer called the ”input gate layer” and a tanh layer (which condenses values to be between -1 and 1 to regulate the network) that ”creates a vector of new candidate values, $\tilde{C}_t$, that could be added to the state”[13].

Once we have calculated the new cell stat C_t, we arrive at our final layer: the output gate. This layer decides what the hidden state should be by taking in the prior hidden state, the current input, and the current cell state [14]. We run a sigmoid layer on the cell state to determine which parts to output, and then run the cell state through a tanh layer (again, this helps regulate the values to be between -1 amd 1) to output a hidden state h_t. This hidden state h_t is the output of the cell and is passed to the next cell, along with the cell state C_t [14].

A key assumption of LSTMs, and recurrent neural networks in general, is that the state at a current time step is dependent on a previous time step[19]. While we contend that this is a safe assumption for our specific use of LSTM, we recognize that there are other situations in which time steps also require information from the ones following. It is likely that our results would be improved through implementation of an advanced LSTM (A-LSTM), which has been shown to provide slightly better predictions compared to standard LSTM[19]. LSTMs require significant computational bandwidth due to their four linear layers, which means that they are not suitable for all research problems.

In our work, we employ the LSTM within our recurrent neural network that we use to predict stock price changes. We employ 5-fold GridSearch Cross-Validation to optimize the parameters of the model by training it on the first 40 days of input data (which includes either sentiment analysis scores or bag of words representations of text for each day, and prior days’ stock price changes). We then use this optimized model and train and test it on our data (using the aforementioned train-test split).

4 Results

4.1 Exploration of GME Sentiment and Stock Outcomes

We begin by visually exploring the relationship between Reddit sentiment about GME and the stock’s price and trading volume fluctuations, with the hope of finding some correlation between the two. Plotting the two trends on the same axis shows some correlation, though with a time lag (see appendix: 4). Next, we find that there is some (albeit weak) direct positive correlation between a day’s overall positive compound score (that is, the sum of all the positive compound scores derived from Reddit posts for that day) and its GME stock trading volumes (see appendix: 5). We explain this correlation as being due to the fact that positive sentiments expressed in the r/WallStreetBets forum likely encouraged users to buy more GME stock, increasing trading volumes.

We also plotted Reddit sentiment scores with GME stock price changes and found a similar visual correlation. On the other hand, our plots for Google News sentiment scores with GME stock outcomes did not show any significant correlation. We turn next to calculate Granger Causality scores to better quantify and understand the relationships between sentiments and stock outcomes.

4.2 Granger Causality

Granger causality is a ”statistical concept of causality that is based on prediction” [17][4]; it revolves around the concept that causality can be shown by measuring the ability of past values of one time series to predict future values of another [17]. It depends upon the assumptions that a) the two time series have a linear relationship with one another (though non-linear Granger causality tests do exist) and b) that the two time series are stationary, meaning they have the same mean over time.

We utilize Granger causality scores to establish whether our sentiment scores can be used to predict future stock developments, and if so, which sentiment scores are most closely related to price movements. We first shift our data set to be stationary around a deterministic trend. We validate that stationarity has been achieved using the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, which tests that a given time series is univariate and stationary[7].

The Granger causality package calculates four test metrics for the given time series: two F-test values and two Chi-Squared test values. We choose to prioritize the F-test values because F-tests are more accurate for smaller sample sizes, and also because Chi-Squared is generally used for categorical values (whereas both of our time series are continuous values). While the two F-test values (SSR-based and parameter-based) are essentially the same metric, we believe the SSR-based metric (where SSR stands for sum of squared residuals) will provide a better sense of the accuracy of predicted values compared to actual. Granger causality calculates these metrics over the course of several time lags; we choose to look at metrics over the course of one-day, two-day, and three-day lags.

The null hypothesis here is that no Granger causality exists between the two time series (specifically, that the second time series does not cause the first). The below table demonstrates Granger causality scores between different VADER sentiment scores derived from Reddit and Google News, and GME stock price changes.

	Reddit			News		
Lag	Positivity	Negativity	Neutrality	Positivity	Negativity	Neutrality
1	0.0354	0.0290	0.0377	0.2470	0.2544	0.4520
2	0.1027	0.1204	0.0518	0.3509	0.2509	0.679
3	0.2659	0.2241	0.1016	0.2417	0.0912	0.1689

Table 1: p-values obtained using Granger causality analysis on different sentiment scores and with different lags (in days)

With a p-value below 0.05 for a one day lag, we can reject the null hypothesis and find that Reddit sentiment has a one-directional causal relationship with GameStop prices in that period. This is to be expected given Reddit's strong influence on the stock's price in our time window. Interestingly, we do not see statistically significant p-values for two-day and three-day lags, which implies that sentiments expressed on r/WallStreetBets may only have an immediate, short-term effect on GME stock price changes. We are unable to reject the null hypothesis for sentiments derived from Google News, so we cannot say that Google News sentiments is Granger causal of GME stock price changes.

In performing Granger causality for the reverse direction (whether stock price changes are causal of sentiment), we do not find causality for the Reddit sentiments, proving a solely one-directional, causal relationship between sentiment and price. For the Google News sentiments, we find that we can reject the null hypothesis (p=0.0358) for GME stock outcomes being Granger causal of a given day's summed compound sentiment scores on a one-day lag. This indicates that stock price outcome might be causal of Google News sentiment, which makes intuitive sense given that headlines report on stock outcomes.

4.3 Stock Price Prediction

Even though Granger causality provides us with positive results, identifying a causal relationship between the GME stock price and the previous day's Reddit sentiment, we hope to further evaluate the correlation between both, especially because Granger causality assumes a linear relationship between both time series, which may not necessarily be assumed here. Thus, we set out to run 5 different learning algorithms, including Linear Regression, Logarithmic Regression, SVC, and two neural networks. All models have been optimized using hyper-parameter tuning and cross validation.

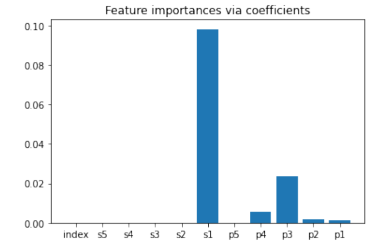

Figure 2: Sentiment score (s) importance with different lags (e.g. s1 - one day lag and previous stock price changes in percent (p). Clearly showcasing that sentiment (in this case positivity) with a one day lag is most important in predicting stock prices.

Prior to model fitting, we employ LASSO feature selection on the data set to identify which features (i.e. types of sentiments and their associated lags) are most indicative of stock price fluctuations. We chose Lasso feature selection due to its advantages over step-wise selection, mainly its penalization of the l1-norm which forces increased sparsity during variable selection and leads to more accurate results. Our feature selection backs up our Granger Causality findings and reaffirms that the ”positivity score” with a one day lag is significantly predictive of price changes. A larger lag of 2 - 5 days gradually reduces the accuracy and other sentiment scores, such as neutrality and negativity, have a lower weighted importance with similar lag trends. These findings convince us to prioritize the positivity sentiment with a short lag of 1 - 3 days to fit our model.

Input	Algorithm	Hype	Tech	Volatile	GameStop
Reddit	Logistic Regression	54.7	49.2	50	65.7
	Linear Regression	52.2	49.3	50.8	62.9
	SVC	52.4	50.1	49.2	71.4
	Neural Network	**62.5**	**59.2**	**52.4**	71.4
	Neural Network (BOW)	56.1	57.2	49.1	**73.6**
Google News	Logistic Regression	54.7	49.2	50	65.5
	SVC	55.3	54.1	53.7	56.9
	Linear Regression	48.3	50.3	54.1	47.9
	Neural Network	**59.7**	**57.5**	**56.7**	**65.7**
	Neural Network (BOW)	57.6	54.3	50.2	64.7

Table 2: A comparison of model performance based on algorithm, stock type, and data set.

We chose to focus on directional accuracy, which is to see if our model can predict the directional change (up or down) in price of the next day. We also used MAPE (mean absolute percentage error) evaluation, but like other scholars [12], we found that more discernible results can be found with a binary directional metric. In the context of our findings, the Efficient Market Hypothesis (EMH) would predict a 50% directional accuracy in the long-run, as prices should always be all-encompassing. We find that logistic regression and SVC (Support Vector Classifier) are not able to consistently predict significantly above the random state. This indicates that using a binary classification model may not be perform best for predicting stock price outcomes. Similarly, linear regression under-performs, which is expected given the unrealistic assumption of a linear relationship between stock price developments and market sentiment.

Our LSTM neural network however provides impressively accurate and significant directional accuracy (when trained on continuous variables and then converted to a binary classification). It outperforms all other regression models across predictions for all types of stocks. In addition, we find that Reddit sentiment is strongly predictive of next-day stock price movements for GameStop (GME) and other ”Hype” stocks with a predictivity of 73.6% and 62.5% respectively. Reddit strongly outperforms news headlines' in these categories, indicating that the r/WallStreetBets subreddit has a more accurate read or potentially a causal effect on hype stock price developments. This can in part be traced back to the fact that hype stocks often move at price levels far above the fundamental value of a stock, something that might concern (negative sentiment) journalists but excite (positive sentiment) Reddit users.

Both Reddit and Google News headlines were both also able to predict established tech firms' stock prices with a 59.2% and 57.5% accuracy, which although lower than GME's and other hype stocks' predictivity levels is still significantly above the random state. As expected, volatile, non-hype stocks were most difficult to predict, as many of the firms have extremely high variance and little sentiment data which lead to a lower accuracy. Interestingly, this is the only stock type where news headlines (56.7%) outperform Reddit data (52.4%), possibly due to the more analytic and finance-based approach of journalists when compared to the average Reddit user.

Separately, we also explore stock price movement prediction with a bag-of-words representation of our data set instead of using sentiment scores. We only input the bag-of-words model into our neural network, and not other regression models, because of the large size of the bag-of-words matrix. We find that the bag-of-words input leads to higher accuracy when predicting GME stock movement from Reddit, confirming our hypothesis that VADER sentiment analysis may perform poorly on Reddit ”slang” and lead to less accurate predictions. Using bag-of-words does not lead to more accurate predictions for any other type of stock, which leads us to believe that VADER sentiment may perform better in contexts where difficult-to-comprehend slang is less commonly used.

Overall, we find that Reddit sentiment, in comparison to Google News sentiment, is more predictive of hype stocks, slightly more predictive of tech stocks, and less predictive of volatile stocks. This may be attributed to the fact that volatile stocks were not as discussed in the r/WallStreetBets forum.

5 Discussion and Conclusion

In this work, we investigate the relationship between online sentiment and stock outcomes, particularly in the context of hype stocks such as GameStop. Through analysis performed on our created data set of Reddit posts and Google News headlines, we find that there is a Granger causal relationship between VADER sentiment scores of Reddit posts and hype stock price movements, as well as one between stock outcomes and Google News sentiments. Furthermore, in using these sentiment scores to predict future stock outcomes, we find that our neural network obtains relatively high prediction accuracy for all types of stocks. Assessing feature importance through LASSO shows that the prior day's sentiment is a key predictor of today's stock price movement. Interestingly, we find that a bag-of-words representation of Reddit posts leads to better prediction accuracy for GME (73.6 percent) than sentiment analysis scores (71.4 percent), perhaps indicating that sentiment scoring performs less well for social media platforms, though this would necessitate further exploration.

Our results support our hypothesis that r/WallStreetBets sentiments influence GME stock price movements. Almost all of our regression models performed best when predicting GME stocks, likely due to the fact that GME was over-represented in both Reddit and Google News data (28 percent of our data set was GME-specific). Future work should explore additional types of stocks and other sources of public sentiment (other Reddit forums such as r/Daytrading, other social media, blog posts, etc.) to understand how public sentiment performs in the context of stock price prediction. We also encourage exploration of other sentiment analysis algorithms such as Bidirectional Encoder Representations from Transformers (BERT) to potentially increase accuracy levels. Lastly, we believe that a potential ensemble model, combining both a bag of words and sentiment analysis neural network with additional algorithms can further improve our accuracy levels. We encourage future researchers to utilize the data set we published [18] or extract their own data using the methods found in our attached notebook.

Acknowledgments

We want to thank the whole COS424 team for a fantastic learning experience and for all the hard work put into supporting us this semester! We want to thank Dr. Li for supporting us on Ed and for organizing the Gathertown presentation, which helped us significantly improve our approach. We want to acknowledge Professor Engelhardt, Deniz Oktay, Siena Dumas Ang, and Andy Jones for coming to our Gathertown presentation and providing us with extremely helpful feedback and tips. Lastly, we want to acknowledge Yaniv Ovadia and Sulin Liu for helping us in Office Hours (and often staying overtime to do so!).

References

[1] Dell Zhang Andrius Mudinas and Mark Levene. Market trend prediction using sentiment analysis: Lessons learned and paths forward. https://arxiv.org/pdf/1903.05440.pdf.

[2] Python Software Foundation. Googlenews 1.5.8 api documentation. https://pypi.org/project/GoogleNews/.

[3] Python Software Foundation. Yahoo-finance 1.4.0 api documentation. https://pypi.org/project/yahoo-finance/.

[4] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[5] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf.

[6] Ching-Ru Ko and Hsien-Tsung Chang. Lstm-based sentiment analysis for stock price forecast. *PeerJ Computer Science*, 7, 2021.

[7] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178, 1992.

[8] Scikit Learn. Count vectorizer - sklearn. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

[9] Scikit Learn. Tfidf transformer - sklearn. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html.

[10] Scikit Learn. Tfidf vectorizer - sklearn. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

[11] Cheng Long, Brian M Lucey, and Larisa Yarovaya. " i just like the stock" versus" fear and loathing on main street": The role of reddit sentiment in the gamestop short squeeze.

[12] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf.

[13] Christopher Olah. Understanding lstm networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs.

[14] Michael Phi. Illustrated guide to lstm's and gru's: A step by step explanation. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21.

[15] Reddit. Reddit api documentation. https://www.reddit.com/dev/api/.

[16] Sreelekshmy Selvin, R Vinayakumar, E. A Gopalakrishnan, Vijay Krishna Menon, and K. P. Soman. Stock price prediction using lstm, rnn and cnn-sliding window model. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1643–1647, 2017.

[17] A. Seth. Granger causality. *Scholarpedia*, 2(7):1667, 2007. revision #127333.

[18] Alexis Sursock and Tiffany Huang. Reddit wallstreetbets: Hype stock posts - kaggle data set. https://www.kaggle.com/aaaaaaaaade/reddit-wallstreetbets-hype-stock-posts.

[19] Fei Tao and Gang Liu. Advanced lstm: A study about better time dependency modeling in emotion recognition, 2017.

[20] Zaghum Umar, Mariya Gubareva, Imran Yousaf, and Shoaib Ali. A tale of company fundamentals vs sentiment driven pricing: The case of gamestop. *Journal of Behavioral and Experimental Finance*, 30:100501, 2021.

[21] Štefan Lyócsa, Eduard Baumöhl, and Tomáš Výrost. Yolo trading: Riding with the herd during the gamestop episode. Technical report, Kiel, Hamburg, 2021.

[22] Victor Zhou. A simple explanation of the bag-of-words model. https://towardsdatascience.com/a-simple-explanation-of-the-bag-of-words-model-b88fc4f4971.

6 Appendix

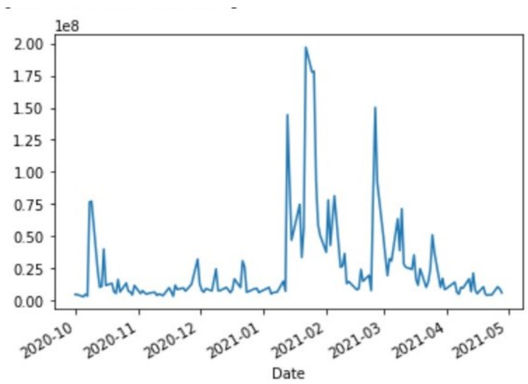

Figure 3: GME Trading volume from October 1st to April 29th with extremely strong variability due to the Reddit short squeeze

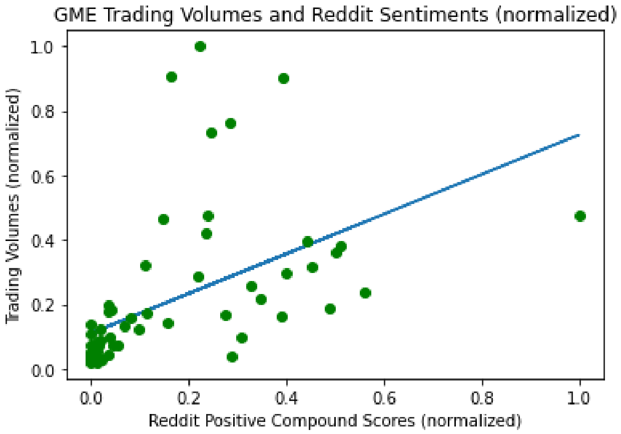

Figure 4: Plot of normalized GME-related Reddit sentiment and GME trading volumes, with line of best fit

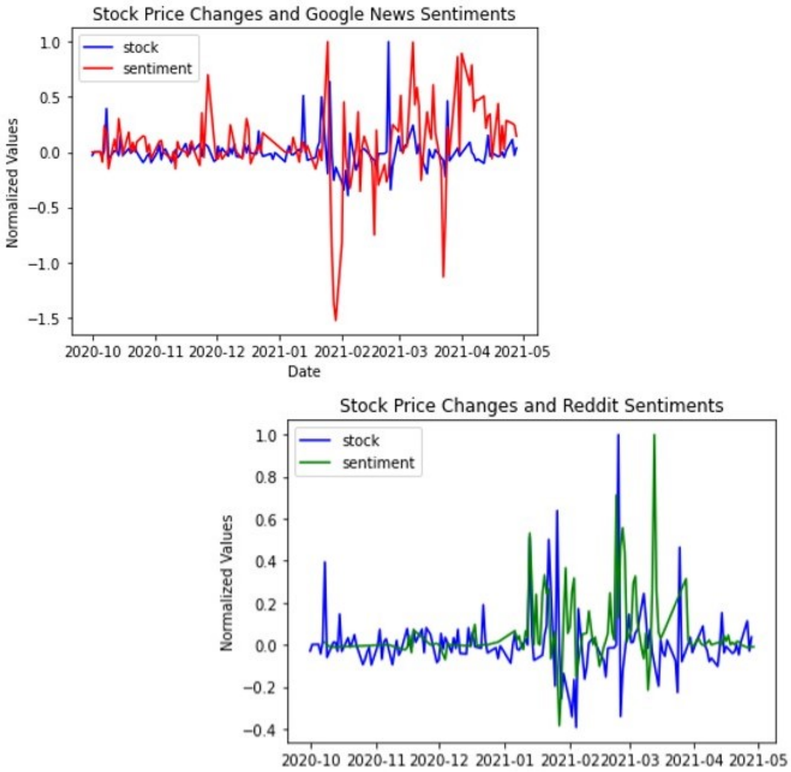

Figure 5: Minimal correlation between Google News sentiments and GME stock price changes, but significantly stronger correlation between Reddit sentiments and GME. This relationship is explored more thoroughly in the context of Granger's Causality