Analyzing impact of twitter sentiment on stock market dynamics using spectral clustering and deep learning

Vagmin Viswanathan, Rahul Gupta
Math 76.01: Topics in Applied Math
Department of Mathematics, Dartmouth College

***Abstract*— Modern social media platforms have become influential forums for discussing and disseminating public opinion on topics like the financial markets. Previous literature has extensively demonstrated the impact of investor sentiment on stock price movements. This study seeks to further characterize the underlying structure of the relationship between time series data of equity price metrics and sentiment streams during the pandemic. We aggregated web scraped financial tweets and employed natural language processing techniques for sentiment analysis and labeling. We then constructed a network and employed spectral clustering to explore structures within the data. Then, to ascertain the predictive value of our sentiment metrics, we utilized regression techniques, random forest analysis, and LSTM neural networks. These allowed us to determine equities impacted the most by investor sentiment and enhanced our model's predictive power of future equity prices. The findings of this research provide insights into the dynamics of social media sentiment in financial forecasting, offering a new perspective on market analysis.**

***Index Terms*— sentiment analysis, spectral clustering, LSTM neural network, random forest classifier**

I. INTRODUCTION

The stock market is a complex and dynamic system, influenced by a myriad of factors ranging from economic indicators and corporate performance to geopolitical events and investor sentiment. At its core, the movement of stock prices reflects the collective actions of buyers and sellers in the market, driven by their expectations for future earnings and company growth prospects. Traditional financial theories, such as the Efficient Market Hypothesis (EMH) [1], suggest that stock prices at any given time fully reflect all available information. However, the proliferation of social media platforms and no fee retail trading apps since 2020 have allowed investor sentiment to be disseminated and represented in the financial markets to an ever greater extent.

In particular, platforms like X[1] (formerly known as Twitter) have drawn investors to its significant sources of real-time news, opinions, and sentiment, accessible to millions of users worldwide. The immediacy and breadth of information available on social media have the potential to influence public perception and investor behavior, thereby affecting stock prices. Recent advancements in Natural Language Processing (NLP) and machine learning have paved the way for sophisticated sentiment analysis of these platforms. These techniques enable the extraction of subjective information from tweets, allowing the application of traditional data science analysis to come to bear.

This paper is structured as follows: Section II summarizes the findings from the existing literature. Section III explains the data collection and prepocessing process. Section IV describes the models and methods used for our analysis. Primarily we focus on network analysis techniques to understand the structure of our data and machine learning techniques to understand feature importance. In Section V, the results of our models have been provided; followed by a discussion of their significance for the literature. In Section VI, we discuss our findings and in Section VII we conclude our paper with acknowledgements in Section VIII and references in Section IX. The Appendix provides further details on additional select companies. Our code and processed data is also documented and accessible on GitHub[2].

II. LITERATURE SURVEY

According to Giachanou and Crestani, Twitter Sentiment Analysis (TSA) has been an exciting field of research since 2009 [2] where foundational papers have established natural language processing and data mining techniques. The Twitter API (Application Programming Interface)[3] has long provided open access to millions of user generated tweets that could be scrapped and analyzed. Giachanou and Crestani describe in their survey of machine learning methods how TSA is difficult because of the 140 character limit, informal style, and unique word embedding structure [2]. Traditional sentiment analysis literature typically tries to understand sentiment through text polarity. TSA methods require twitter specific stop words, tokenizers, word embedding structures, multilingual and multi-modal approaches. Advances in deep learning models and labeled datasets around 2012 has enabled the field to mature beyond these fundamental challenges. Furthermore, Wiebe et al. demonstrates how to create sentiment time series where sentiment scores were smoothed by taking the daily positive versus negative ratio with a moving average window of the past k days [3].

[1] https://twitter.com/

[2] https://github.com/

[3] In recent years the takeover and re-branding of the platform as X has severely limited access to data mining

A natural application of these sentiment time series is found in financial market prediction. The foundational work by Bollen et al. investigated the correlation between collective mood states derived from Twitter feeds (Happy, Calm, Anxiety) and the value of the Dow Jones Industrial Average (DJIA), employing a Fuzzy neural network for prediction [4]. Their findings affirm a strong correlation between public mood states on Twitter and the DJIA's fluctuations, underscoring the potential of social media sentiment as a predictor for stock market movements. Others like Dickenson and Hu have explored the correlation between public sentiments on Twitter and stock market movements, employing the Pearson correlation coefficient to analyze stock increases and decreases [5]. In the subsequent years the extensive literature documenting correlations between tweet sentiment and stock prices over time through a variety of machine learning models has exploded [6][7][8].

More recently, much work has been done to understand the structure of the relationship between the markets and sentiment time series [9][10]. Mendoza-Uridales et al. has utilized transfer entropy to find that the direction of information transfer between the two time series is from public sentiment to stock prices. They also find using the EGARCH model that the intensity of the impact of sentiment on stock price is asymmetric, with negative sentiment being statistically different from positive sentiment [11]. On the other hand, analysis conducted by Cui and Liu on the impact of investor sentiment on financial markets, as observed in China's online lending platforms, used the BERT model to highlight how positive and negative sentiments correlate with financial indicators like borrowing interest rates and periods [12]. Subsequently, integrating investor sentiment with Long Short Term Memory (LSTM) models significantly improved the accuracy of their financial predictions.

Our research builds upon the outcomes of these preceding studies, introducing a novel approach that not only predicts the rise and fall in stock prices based on Twitter sentiments but seeks to understand the structure of the correlations through network analysis techniques. Unlike previous studies, we developed an advanced sentiment analysis model that outperforms the one used by Dickenson and Hu, incorporating spectral clustering to group similar sentiments more effectively. We also use this sentiment time series to predict future stock prices with a random forest classifier. The success of Cui and Liu inspired us to also use LSTM models to better capture the complex dynamics of market behaviors and temporal dependencies in stock price movements.

III. Data

A. Data Collection

For the construction of our sentiment time series, we started our study with the collection of Twitter data. In particular, we were interested on tweets related to the financial markets during the Covid-19 pandemic due to the high level of market volatility and significant social media activity. In addition, the rise of popular apps like Robinhood[4] democratized access to the markets during this time, further amplifying the impact of retail investor sentiment on the markets.

Due to the restrictions on web scrapping on X, we turned to aggregating existing repositories of web scraped data. A promising dataset from April 9, 2020, to July 16, 2020 by Taborda et al. comprises of around a million finance related tweets [13]. Other datasets we found included tweets scrapped using StockerBot[5] and Reddit posts from the infamous WallStreetBets[6] subreddit that were of lower quantity and quality.

To correlate the sentiment time series with actual stock market performance, we obtained historical equity pricing data from the Wharton Research Data Services (WRDS) database [14]. This data included daily closing prices, volume, SIC codes[7], and 63 other relevant financial metrics for all companies listed on the NASDAQ during the period studied.

For each company, we selected a set of features that are traditionally influential in stock price movements. These included volume (VOL), return (RET), bid and ask prices (BIDLO, ASKHI, BID, ASK), shares outstanding (SHROUT), adjustment factors (CFACPR, CFACSHR), opening price (OPENPRC), number of trades (NUMTRD), and various return measures (RETX, vwretd, vwretx, ewretd, ewretx, sprtrn). We also included the total sentiment score as a feature, aiming to capture the overall sentiment towards the company on Twitter.

B. Data Processing

Given the raw tweet and equity metric data on each date, our goal was to produce a sentiment and stock pricing time series. To this end, we began by cleaning the Twitter data. Given the voluminous and unstructured nature of the tweets, we undertook an extensive data cleaning process, including the removal of duplicates, spam, and non-English tweets. Similarly, we also cleaned the stock market data to remove any anomalies or missing values to ensure consistency and accuracy in the dataset.

We then tagged each tweet with the company of interest that was discussed (multiple mentions of companies got multiple tags) by extracting the tickers. For example, the tweet in figure 1 got tagged as discussing $WTI. From the metadata, we were also able to extract the date of the tweet.

Figure 2[8] shows the distribution of tweets across companies. As we can see, the tweets are heavily skewed to a handful of the most tweeted about companies. These tended to be tech stonks like the Magnificent Seven[9].

[4]https://www.robinhood.com/

[5]https://github.com/dwallach1/StockerBot

[6]https://www.reddit.com/r/wallstreetbets/

[7]A system used by government agencies to classify companies into major industries and sectors

[8]https://twitter.com/conkers3/status/1252147181399793670/

[9]Alphabet, Amazon, Apple, Meta, Microsoft, Nvidia, and Tesla

Fig. 1. Example tweet

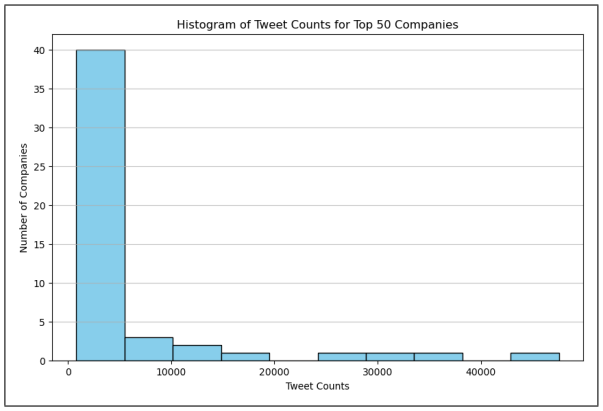

Fig. 2. Distribution of tweets across companies

In fact, the top five most tweeted about companies in our dataset were Apple, Amazon, Facebook, the S&P 500 index, and Microsoft accordingly. This was largely a function of the popularity of these stocks as consumer facing companies discussed on Twitter and not necessarily representative of the US economy as a whole.

Finally, we merged the cleaned Twitter and stock market datasets based on the date, aligning tweets with their corresponding stock market performance. All in all, we were able to recover around 400,000 rows of (date,tweet,ticker) triples. This merged dataset formed the basis of our analysis.

C. Sentiment Analysis

The first step in our analysis was to determine the sentiment of each tweet in our dataset. Utilizing natural language processing package NLTK[10], we used the TSA techniques discussed in the literature review to classify tweets into positive, negative, or neutral categories based on the text content. This process involved tokenization, stop word removal, lemmatization, and the application of pre-trained sentiment analysis models like textblob[11].

Textblob provided us with a sentiment score between -1 and 1 based on the polarity of the words within the tweet. Aggregating these scores across all tweets for a company on a given data resulted in a daily sentiment score for each company. We then construct a dataframe where each company on a given day has a row with their cumulative sentiment and equity metrics on that date. Note that some companies did not have any tweets on certain days. Following Weibe et al. [3] we were able to construct the sentiment time series for each company by taking a rolling average of the cumulative sentiment scores, correcting for the days without any sentiment.

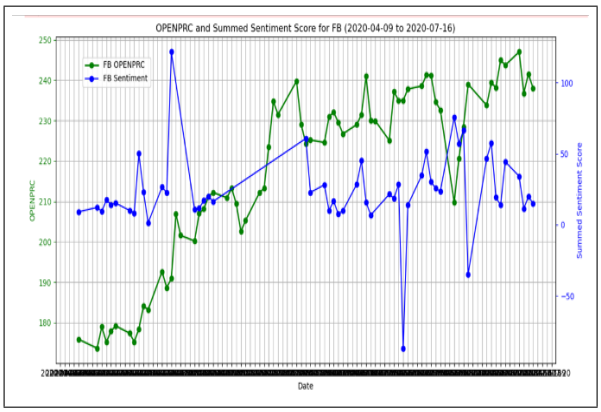

Fig. 3. Open price versus cumulative sentiment time series for Facebook from April 9 to July 16, 2020

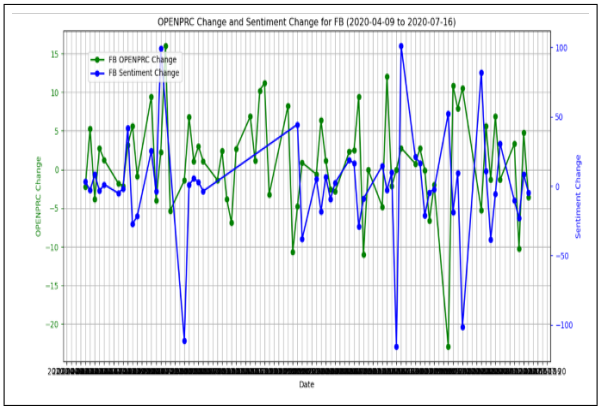

Fig. 4. Percent change in open price versus cumulative sentiment time series for Facebook from April 9 to July 16, 2020

Figures 3 and 4 show the sentiment time series overlayed with the open price for Facebook. We can see that the percent change of the time series over days better captures the relationship between the cumulative sentiment score and open price, which is the focus of our analysis.

To finalize the data preparation for our future analytical techniques, we needed to make all of our features

[10] https://www.nltk.org/

[11] https://textblob.readthedocs.io/en/dev/

numerical and standardized. We began by converting all the tickers into a company index from 0 to 5,222. We then converted non-numerical financial metrics into a vector with the use of word to vec package TF-ID [12]. We then use a standard scaler to mean center and normalize all of our numerical features to remove any biases from their relative magnitudes.

We construct our data matrix $X \subset \mathbf{R}^m$ with feature vectors v_i for $1 \leq i \leq 64$ that represent the 63 processed equity metrics we found from WRDS and the cumulative sentiment on a given date for each company.

IV. Models and Methods

Drawing on the foundational literature, we divide our analysis into two goals. First, in section *A* we seek to understand the structure of the relationship between our Twitter sentiment time series and stock market movements using network analysis. With these insights, we then seek to predict future stock price movements in section *B* with various machine learning techniques.

A. Structure of Data

In this section, we begin by understanding the statistical significance of correlations between our different feature vectors in subsection 1. We then construct a network and cluster it in subsection 2 to understand the similarity of companies' sentiment time series. Finally, we follow Luxberg [15] in subsection 3 in applying spectral clustering to our data matrix X to understand the similarity of the impact of sentiment movements on different company's equity metrics.

1) Correlations: We consider whether correlations between Twitter sentiment scores and various stock price metrics are statistically significant using the χ^2 test (2). We repeat the following process for each of the main stocks considered.

First, we construct a contingency table based on the number of days of positive sentiment change and positive price change as seen in figure 5 for Apple.

	Price Up	Price Down
Sentiment Up	17	8
Sentiment Down	7	21

Fig. 5. Contingency Table for Apple

The table entries are the observed frequencies (counts) O_i for each combination. We then calculate the expected frequencies E_{ij} for each cell in the table using (1). Now we can find the χ^2 statistic with (2).

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}} \tag{1}$$

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

[12]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

To understand if this statistic is significant or not compared to our null hypothesis (that there is no correlation between the two time series), we find a p-value. The p-value is calculated based on the χ^2 distribution with degrees of freedom (dof) determined by (5).

$$dof = (\text{Number of Rows} - 1) \times (\text{Number of Columns} - 1) \tag{3}$$

If the p-value is below a certain threshold (commonly 0.05 or 5%), it suggests that there is a significant association between the two variables, indicating that sentiment direction might affect the price direction of the stock, or vice versa. If the p-value is above the threshold, we fail to reject the null hypothesis, indicating no significant association between sentiment direction and price direction.

Following the results of the test, we have evidence for statistically significant correlations between the sentiment and equity metric time series of the few companies considered. To better understand this correlation across our dataset, we construct a correlation matrix between each feature vector in our data matrix X using Pearson correlation.

2) Network Analysis: In order to understand what sorts of companies are impacted by their sentiment time series in similar ways, we can construct a network and apply spectral clustering to it. First, instead of considering our original data matrix X, we construct a new matrix Y that contains all of the companies as the columns and times on the rows. Each Y_{ij} entry contains the cumulative sentiment of i_{th} company at time j.

Let A and B represent two companies and S_{Ai} and S_{Bi} represent their sentiment scores at time i. We then consider the correlation matrix of Y constructed by taking the pairwise Pearson correlation between all such A and B (4).

$$\rho_{S_A,S_B} = \frac{\sum_{i=1}^{n}(S_{A,i} - \bar{S}_A)(S_{B,i} - \bar{S}_B)}{\sqrt{\sum_{i=1}^{n}(S_{A,i} - \bar{S}_A)^2}\sqrt{\sum_{i=1}^{n}(S_{B,i} - \bar{S}_B)^2}} \tag{4}$$

We then construct an adjacency matrix Z to represent the distance between two companies with respect to their correlation (5).

$$Z = \frac{1 + \text{correlation\_matrix}}{2} \tag{5}$$

To transform this adjacency matrix into a network we let the companies represent nodes and the weights on the edges between them represent their relative distance. Instead of taking the raw weights from the adjacency matrix, we apply a Gaussian kernel transformation (6) to normally distribute them to amplify structures in the data for future analysis. The sigma parameter impacts how correlations get mapped and is determined experimentally to produce an even output distribution.

$$Z_i' j = \exp\left(-\frac{Z_{ij}^2}{2\sigma^2}\right) \tag{6}$$

From here we have successfully constructed a network that encodes the data in a meaningful way. Now we follow Luxberg [15] in applying spectral clustering using the symmetrized Laplacian to cluster this network. We use 12 clusters based the eigenjump observed in the eigenvalues of the Laplacian. Finally we compare these cluster labels with the SIC codes for each company using the Adjusted Rand Index (ARI)[13] to understand if the similarities found due to sentiment match company sectors.

We also consider how consistent these clusterings are across time. To this end we consider a window of time in our data and apply this process to it. We then shift the windows across time and calculate the ARI between the two clusterings, iterating the process across the entire dataset.

3) Spectral Clustering: For this section, we seek to understand what stock time series are similar to each other. First, our initial data matrix X is too large and noisy to consider as is. We apply Principal Components Analysis (PCA) on the features in order to reduce the dimensionality of the data. From there we can preform spectral clustering using the symmetrized Laplacian. Again we consider 12 clusters in order to facilitate comparison with the SIC codes using ARI.

B. Machine Learning Analysis

In this section, we begin our exploration with a standard linear regression in subsection 1. However, we find little linear relationships in our data and move onto more powerful models. Inspired by the success of LSTM models in the literature, we explore their application to our data in subsection 2. We also employ a random forest classifier as an alternative to LSTM to benchmark it's performance in subsection 3. Finally, we analyze the importance of different features to capture the impact sentiment has on the predictive power of our models in subsection 4. Here we also find the companies most impacted by sentiment in our dataset.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{8}$$

For the models, we use the framework of keeping 80% of the data for training and 20% for testing. We calculate their accuracy with metrics like (MAE) Mean Squared Error (7) and (RMSE) Root Mean Squared Error (8). Here, n is the number of observations, Y_i is the actual value of each observation, and $\hat{Y}_i$ is the predicted value. Due to computational constraints, we restrict our analysis to the most tweeted companies to ensure the quality of their sentiment time series.

1) Linear Regression: Our initial predictive model was a simple linear regression to search for linear relationships between sentiment scores, past equity metrics, and the future open price for each company.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \tag{9}$$

Y is a vector representing the open price for all companies at some time t. Each X_i corresponds to an independent variable. In our case, we use a column of our data matrix X that represents the past values of a given equity metric or cumulative sentiment for all companies. β_i are the coefficients the model learns and ϵ is the error term.

2) LSTM Model: The LSTM (Long Short-Term Memory) Neural Network, is designed to capture temporal dependencies, making it a great candidate for time series analysis.

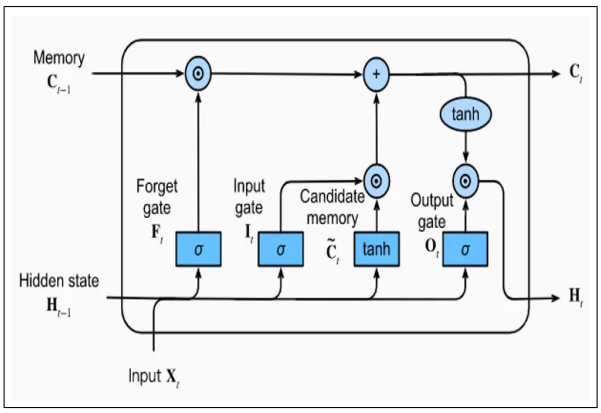

Fig. 6. Architecture of LSTM model

The model is composed of a series of gates that represent different operations on the training data as seen in figure 6[14] . Each gate operation is represented as follows:

- Forget gate: $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
- Input gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
- Cell state update: $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
- Final cell state: $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$
- Output gate: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$
- Final output: $h_t = o_t * \tanh(C_t)$

where x_t is the input at time t, h_{t-1} is the previous output, and W and b represent weights and biases, respectively. We partition our data matrix into time slices of the state of each company on every date.

This model is renowned for its robustness and ability to handle non-linear data, making it a suitable choice for financial market predictions alongside the LSTM neural network.

3) Random Forest Classifier: The Random Forest Classifier constructs multiple decision trees and outputs the average prediction:

[13] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html`

[14] https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c/

$$Y = \frac{1}{n}\sum_{i=1}^{n} Y_i(X) \tag{10}$$

where $Y_i(X)$ is the prediction of the i^{th} tree. Y is then the overall prediction by the model for the vector of open prices for all companies. X is our data matrix and n is the number of observations, which is the number of rows in X

4) Feature Importance: Feature importance was assessed in both the LSTM and Random Forest models to understand each variable's impact, including sentiment scores, on stock price predictions. For Random Forest, a feature's importance is measured by the decrease in impurity. For LSTM, we employed permutation feature importance.

In a Random Forest model, feature importance is derived from how much a feature decreases impurity in the decision trees. Impurity is typically quantified using metrics such as Mean Squared Error (MSE), entropy, or variance. The feature importance for a feature f is calculated as follows:

1) For each tree in the Random Forest:
   - For each split on feature f, calculate the impurity decrease $\Delta i(s,t)$, where s is the split and t is the tree.
   - Aggregate the impurity decreases $\Delta i(s,t)$ for all splits on feature f across all trees to determine the total impurity decrease $\Delta I(f)$.
2) Normalize the total impurity decrease for feature f by the sum of total impurity decreases for all features to obtain the feature importance:

$$\text{Importance}(f) = \frac{\Delta I(f)}{\sum_{f' \in F} \Delta I(f')}$$

where F is columns of our data matrix X.

Permutation feature importance is a model-agnostic technique that quantifies the change in a model's prediction error after permuting the feature's values. The steps for calculating permutation feature importance are:

1) Train the model and measure its performance on a dataset to establish a baseline, using MSE. Denote this as $\text{MSE}_{\text{baseline}}$.
2) For each feature f:
   - Permute the values of f in the dataset, disrupting its correlation with the target.
   - Re-evaluate the model on the perturbed dataset and record the new performance metric MSE_f.
   - Compute the feature's importance as the discrepancy between the permuted metric and the baseline metric:

$$\text{Importance}(f) = \text{MSE}_f - \text{MSE}_{\text{baseline}}$$

3. A more substantial discrepancy suggests a high dependency of the model on the feature, signifying greater importance.

Permutation feature importance considers both direct and indirect effects of features and is less biased towards features with high cardinality. However, it requires multiple model evaluations, making it computationally intensive.

V. Results

A. Structure of Data

1) Correlations: We employed the Chi-square test as the first step to assess the statistical significance of the relationship between Twitter sentiment and stock price movements for select companies such as Apple, Amazon, Microsoft, and Facebook that have most number of tweets.

The results of our Chi-square tests are summarized in the table 1 below:

Company	Chi-Square Statistic	p-value
Apple	8.197	0.0042
Amazon	2.321	0.128
Microsoft	0.188	0.664
Facebook	1.723	0.189

TABLE I
Chi-Square Test Results for Twitter Sentiment and Stock Price Movements

For Apple, the test demonstrated a statistically significant association, with a Chi-Square Statistic of 8.197 and a p-value of 0.0042, indicating that Twitter sentiment could have a notable impact on its stock price movements.

Conversely, the analyses for Amazon, Microsoft, and Facebook revealed no significant correlation between Twitter sentiment and stock price movements, as evidenced by their respective p-values of 0.128, 0.664, and 0.189. These findings suggest that for these companies, other factors might be more influential in driving stock prices than the sentiment expressed on Twitter. However, this is a surface level analysis that fails to incorporate lower level effect of sentiment score on stock prices.

We also developed a heatmap in figure 7 to visualize the correlations between various stock price metrics, including the total sentiment score, to explore how public sentiment intertwines with traditional financial indicators like opening and closing prices, volume, and more.

2) Network Analysis: After constructing our adjacency matrix as described in section 3*a*)2, we plot a histogram of the correlations to understand the relationships. Based on figure 8 we see that they are centered around .5. Intuitively, .5 would be a good guess for the σ parameter of the gaussian kernel. To test this hypothesis, we plot the median mapped value of the correlations for each value of σ.

We can see from figure 9 that the curve corresponding to .5 seems to do a good job of evenly spreading the correlations among the network weights. This allows our network to not be distorted by the raw correlations, and focus on the relative difference in correlations.

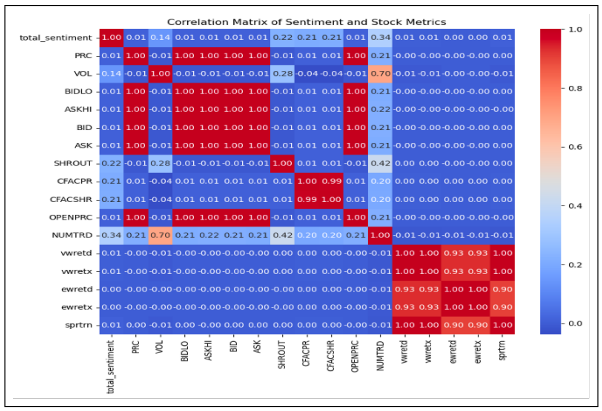

Fig. 7. Correlation Matrix of Sentiment and Stock Metrics

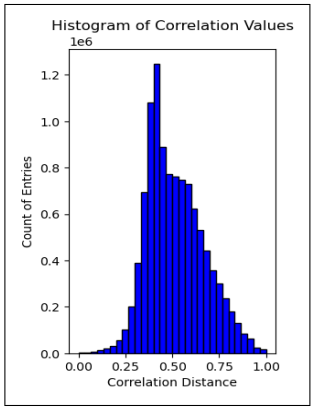

Fig. 8. Histogram of Pearson correlations of Y

Now, we are prepared to run spectral clustering with the symmetrized Laplacian on the network using 12 clusters. We plot the high dimensional space with a t-SNE visualization that maintains relative distances between points even in a 2 dimensional space. We observe cluster 3 is tightly clustered and distinct from the rest of the data. Looking closer, some of the prominent tech companies like Amazon, Tesla, Apple, and Facebook are present inside this cluster, indicating it corresponds to highly popular tech companies. This is an exciting finding that our network was able to identify these very similar companies

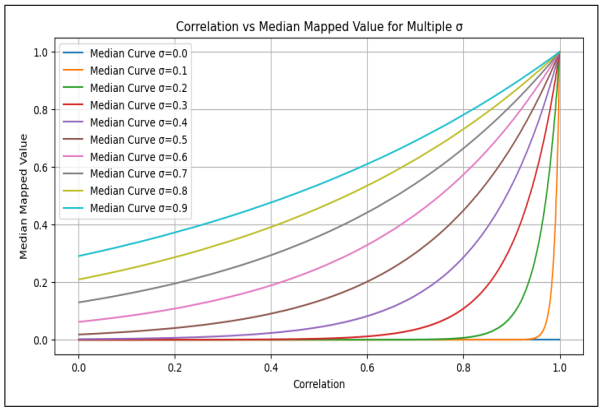

Fig. 9. Median mapped value of correlations for different σ

in terms of sentiment and market movement.

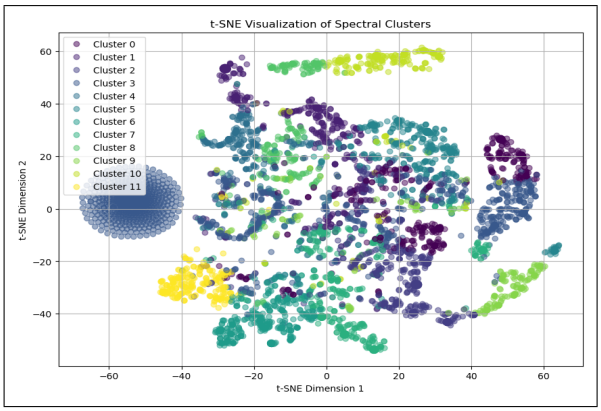

Fig. 10. t-SNE visualization of spectral clusters on Y

To gain more intution about what sort of companies got clustered together, we map out the distribution of major industries (as determined by the SIC codes) across clusters. As seen in figure 11, the distribution appears to be quite evenly spread, which matches up with the very low ARI of 0.008 observed between the two clusterings.

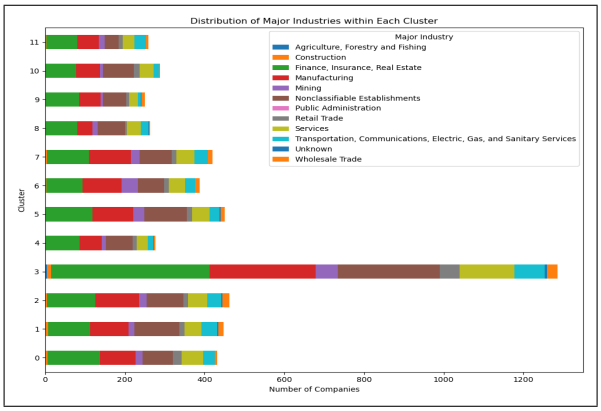

Fig. 11. Major industries within spectral clusterings on Y

Finally, for our rolling window analysis of how consistent the clustering remained across time, the results are as follows:

- ARI between months 1 and 2: 0.1117
- ARI between months 1 and 3: 0.0489
- ARI between months 2 and 3: 0.0.0412

The low ARI numbers and their sporadic nature indicate that the clusterings do not remain consistent over time. This makes intuitive sense as companies sentiments change rapidly across time in unpredictable ways (from our dataset's perspective). As such, what companies are clustered into which cluster largely depends on the news surrounding it in the media at that time.

3) Spectral Clustering: For our PCA, we first construct an elbow plot to determine the number of principal components to include in our dimension reduction. We wanted to preserve as much explained variance in the data as possible for future analysis and therefore chose to

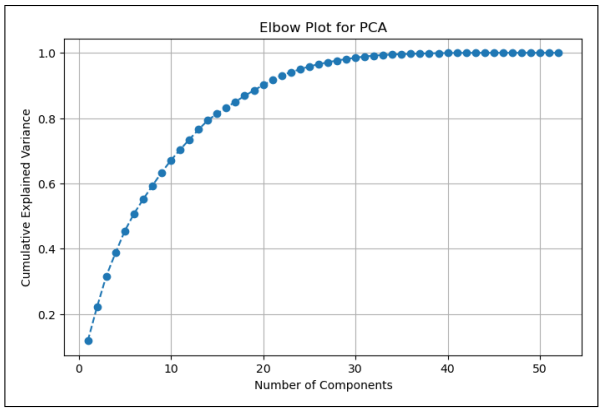

Fig. 12. t-SNE visualization of sentiment spectral clustering

use 30 principal components. As seen in figure 1423, this corresponds to around 95% of the explained variance in the data matrix X.

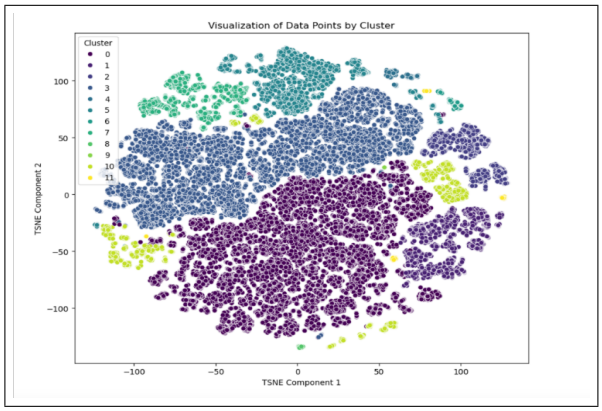

Fig. 13. t-SNE visualization of spectral clustering on $X_{reduced}$

After applying spectral clustering with 12 clusters to this dimension reduced data, we find the following cluster labels. We use t-SNE to visualize the data points in the high dimensional space ($\mathbf{R}^{30}$). After performing the ARI between these labels and the SIC codes of each company, we find the following results:

- ARI: 0.0958

The very low ARI indicates that the impact of sentiment time series on companies equity metrics are not correlated with sector trends. This ARI is slightly higher than the ARI found in section 2, indicating that the equity metrics are somewhat more correlated to company sector than sentiment. Intuitively, sentiment should be more company specific than the financial fundamentals.

B. Machine Learning Analysis

1) Linear Regression: We initially applied a simple linear regression model to the entire dataset as a baseline approach. This step allowed us to establish a preliminary understanding of the relationship between our selected features and stock price movements.

- Variance of Prediction Errors: 7953.696

The relatively high variance indicated considerable discrepancies between the model's predictions and the actual stock prices, suggesting that the linear model might be too simplistic to capture the complex dynamics of stock market movements effectively. Motivated by these findings, we decided to explore advanced machine learning models, focusing on a subset of the dataset to leverage more sophisticated analytical techniques capable of handling the intricacies and nonlinear relationships inherent in financial data.

2) LSTM Neural Network: We utilized utilized Long Short-Term Memory (LSTM) neural networks to effectively assess whether the collective sentiment expressed on Twitter could serve as a meaningful predictor of stock prices for heavily discussed companies. LSTMs are a type of recurrent neural network (RNN) capable of learning long-term dependencies, making them well-suited for time series forecasting. By inputting features derived from both Twitter sentiment scores and traditional stock market metrics, we trained our LSTM models to forecast future stock prices, assessing the predictive power of social media sentiment in comparison to other financial indicators.

For Apple, we utilized the LSTM model for stock price prediction with the stock metrices, targeting the PRC (stock price). We evaluated the model's performance using the Test RMSE and MAE. The results were as follows:

- Test RMSE: 6.280987095540245
- Test MAE: 5.521982934126413

To assess the specific impact of sentiment on the model's performance, we retrained the LSTM model without the `total_sentiment` feature. The resulting performance metrics were:

- Test RMSE: 5.088473338201218
- Test MAE: 4.415160535333794

Interestingly, the removal of the `total_sentiment` feature from the model inputs resulted in improved prediction accuracy. This suggests that, for Apple, the sentiment data might have introduced more noise into the predictive model than provided clarity.

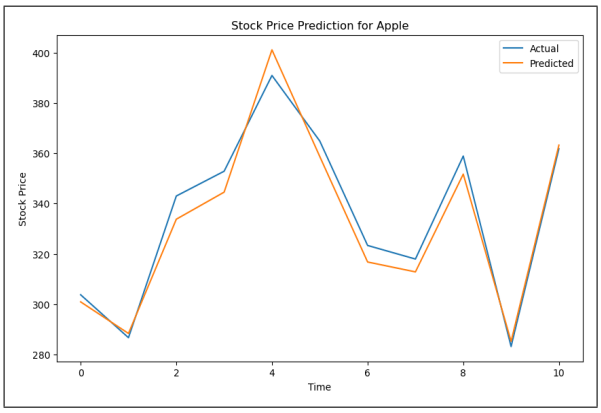

Fig. 14. Apple Stock Price Prediction with Sentiment

We extended our analysis to Amazon, where the inclusion of sentiment presented a different outcome. With

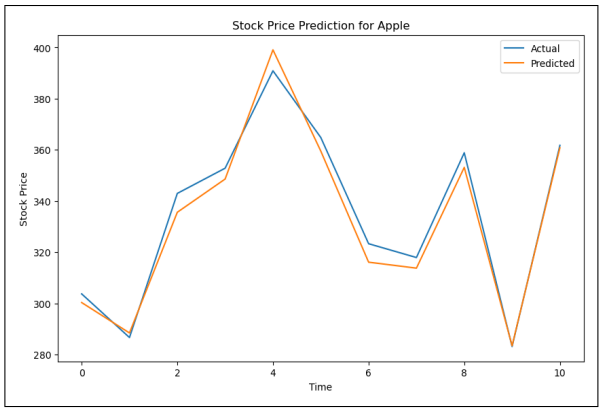

Fig. 15. Apple Stock Price Prediction without Sentiment

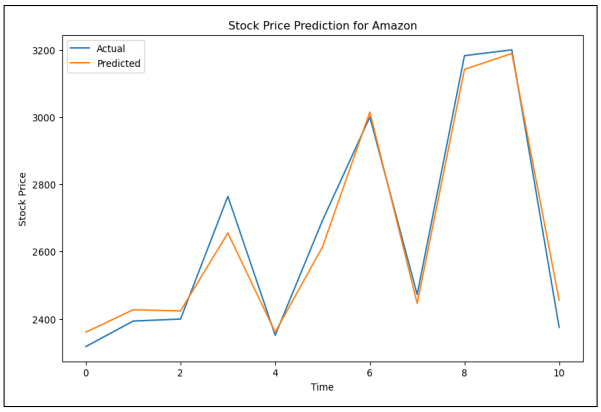

Fig. 17. Amazon Stock Price Prediction without Sentiment

the `total_sentiment` feature included, the LSTM model produced the following results:

- Test RMSE for Amazon: 50.00578527434134
- Test MAE for Amazon: 42.1879873863636

Upon removing the sentiment score from the model, we observed:

- Test RMSE for Amazon: 53.06623252047314
- Test MAE for Amazon: 42.89288795454542

In contrast to the findings for Apple, the inclusion of the sentiment score in Amazon's stock price prediction model led to a slightly more accurate prediction, as evidenced by the decrease in both RMSE and MAE values. This suggests that for Amazon, the overall sentiment on Twitter seems to have a more pronounced impact on its stock price predictions.

The divergent results of our LSTM model analysis between Apple and Amazon highlight the complex and variable nature of the relationship between social media sentiment and stock price movements. While sentiment data did not enhance the predictive accuracy for Apple, it positively contributed to Amazon's model performance.

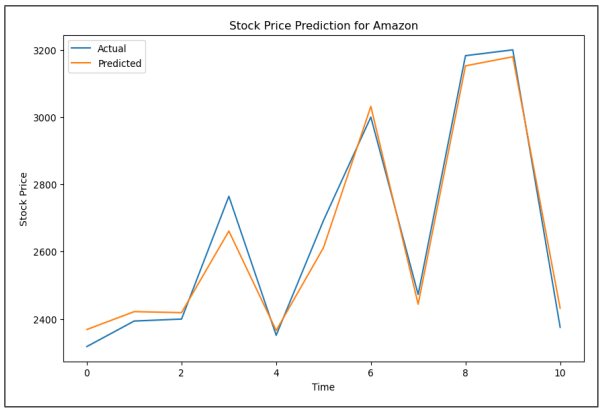

Fig. 16. Amazon Stock Price Prediction with Sentiment

3) Feature Importance Analysis: Our analysis of feature importance offers intriguing insights into the factors that significantly impact the stock prices of Apple and Amazon. By employing machine learning models, we were able to quantify the contribution of each feature, including traditional financial metrics and the total sentiment score derived from Twitter data.

Apple's Feature Importance: The analysis for Apple revealed that the `OPENPRC` (opening price) is the most influential feature, with an importance value of -136.932. This suggests that the opening price is a critical predictor of stock movements, consistent with financial theories that posit the opening price reflects pre-market sentiment and can significantly influence the day's trading dynamics.

- `BIDLO` (lowest bid price during the trading day): Importance Value = -114.499
- `BID` (bid price): Importance Value = -84.869
- `ASKHI` (highest ask price during the trading day): Importance Value = -52.398

These features underscore the market's sensitivity to liquidity and trading pressures. Conversely, features such as `SHROUT` (shares outstanding) and `ewretx` (equal-weighted returns excluding dividends) showed positive values, suggesting a direct but less significant impact on stock price predictions.

The `total_sentiment` score, reflecting collective sentiment towards Apple on Twitter, held a modest negative value, which reinforces its supplementary role in the prediction of stock prices. The sentiment's influence is present, but not as pronounced as direct market indicators like `OPENPRC`, `BIDLO`, and `ASKHI`.

For Amazon, the analysis revealed that the most influential feature is `SHROUT` (shares outstanding), with an importance value of -17779.656. This indicates that changes in the number of shares available in the market profoundly impact Amazon's stock price, potentially reflecting investor reactions to events such as equity dilution or stock buybacks. Notably, this feature's significance was pronounced around the period when Amazon announced a stock split on June 6, 2020, highlighting the direct impact of `SHROUT` on stock price movements during significant corporate actions.

Following `SHROUT`, the `OPENPRC` (opening price) emerges as another critical predictor, with an importance value of -9680.642. This reinforces the significance of

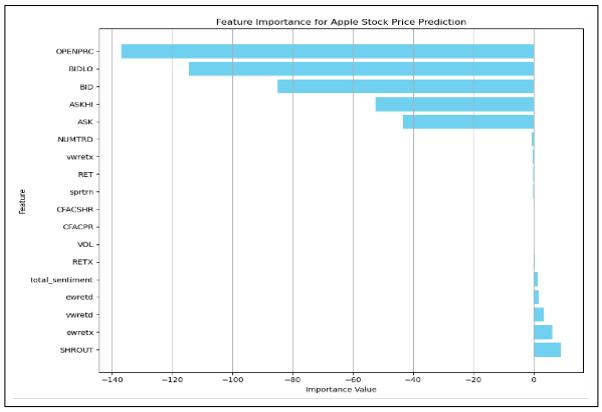

Fig. 18. Apple Feature Importance

opening market conditions and pre-market sentiment in shaping the trading behavior and outcomes for the day, suggesting that the opening price offers valuable predictive insights into stock movement.

Other features, including `BID` (bid price), `ASK` (ask price), and `BIDLO` (lowest bid price during the trading day), also displayed considerably negative values. These reflect the fundamental role of bid and ask prices in determining stock valuation, capturing the immediate dynamics of market demand and supply. Their significance underscores the market's responsiveness to liquidity and trading pressures.

Interestingly, the `total_sentiment` score, quantifying the overall sentiment towards Amazon on Twitter, exhibited notable importance with a value of -2985.023. This finding suggests that public sentiment, as manifested on social media, plays a significant yet intricate role in stock price predictions. The sentiment score's relevance indicates that broader market perceptions and investor reactions to company-related news, as captured through social media sentiment, serve as valuable predictors of Amazon's stock performance.

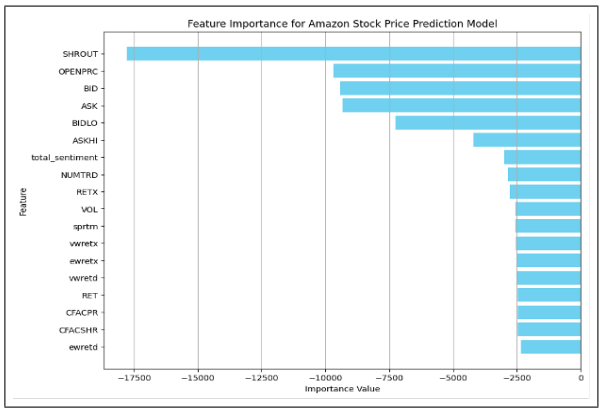

Fig. 19. Amazon Feature Importance

4) Random Forest Regressor: We also utilized the Random Forest Regression model alongside the LSTM neural network in analyzing Apple's stock data to validate the robustness and reliability of our LSTM predictions by checking for consistency in the predictive power and trend detection between the two methodologies.

Our analysis using the Random Forest Regression model on Apple's stock data showcased impressive performance metrics, with a high R^2 score of 0.99 and a low Mean Squared Error of 11.14, indicating a strong fit to the historical price movements. However, there's a concern that the model might be over-fitting the data, capturing noise along with the underlying trend, which could reduce its effectiveness in predicting future stock prices under different market conditions.

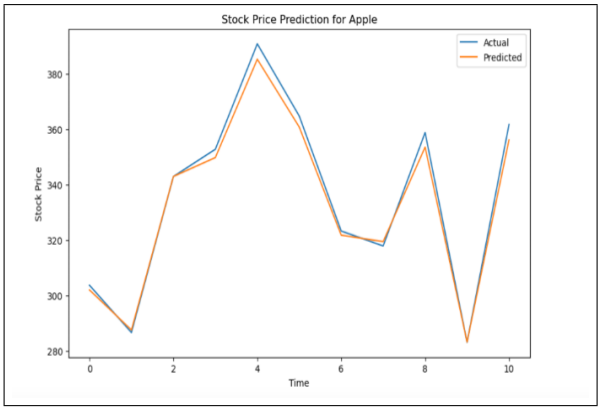

Fig. 20. Prediction Using Random Forest Model

Furthermore, compared to the LSTM approach, the Random Forest model does not incorporate the effect of sentiment scores as effectively. While LSTM models can capture and leverage the nuances of sentiment trends over time, the Random Forest approach may not fully exploit the predictive power of social media sentiment, potentially missing out on valuable insights derived from the emotional and psychological states of market participants.

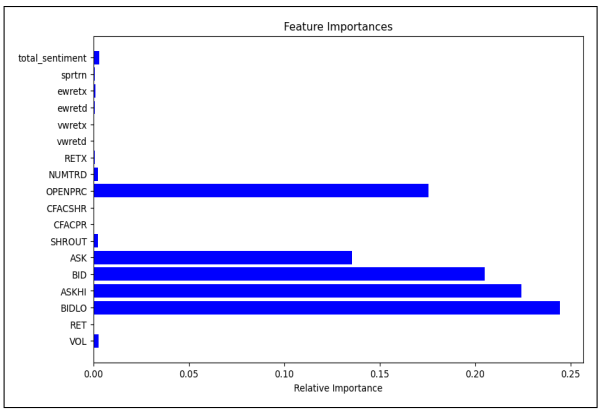

Fig. 21. Feature Importance Using Random Forest Model

5) Sentimental Stocks: The sentiment analysis revealed significant variations in public sentiment towards different companies over the study period. By aggregating daily sentiment scores, we were able to identify trends in public opinion, correlating these with major events and stock price movements.

We investigated the "sentimental" stocks, a term we coined to describe stocks for which sentiment scores serve as significant predictors for price movements. We ran our LSTM model on the top 50 stocks by tweet count to check the degree to which sentiment scores derived from Twitter data influenced stock prices. Our analysis aimed to identify which companies were most responsive to public sentiment, especially during the COVID-19 pandemic when social media discussions presumably had a considerable impact on the financial markets.

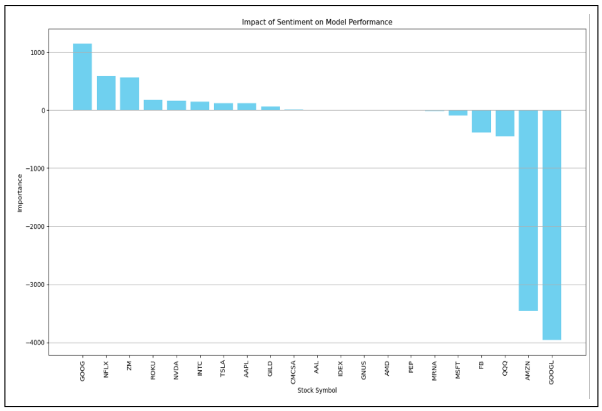

Fig. 22. Sentimental Stocks

The above histogram visualizes the results of our feature importance analysis, indicating the stocks where sentiment scores play a crucial role. Notably, for stocks like GOOGL (Alphabet Inc.), AMZN (Amazon.com Inc.), QQQ (a NASDAQ-100 index ETF), and MRNA (Moderna Inc.), sentiment scores emerged as influential features.

The significance of sentiment scores for MRNA is particularly telling and aligns well with real-world events. During the period of our tweet dataset, which coincides with the height of the COVID-19 pandemic, Moderna was at the forefront of vaccine development. The extensive public discourse and sentiment on Twitter regarding MRNA could be directly linked to the company's visibility in the media and its critical role in addressing the pandemic. This public attention likely translated into substantial sentiment-driven market movements for MRNA's stock, as individuals discussed the company's progress, vaccine efficacy, and potential impact on public health and the economy.

Similarly, for tech giants like GOOGL and AMZN, sentiment scores were found to be significant. This could be due to the high profile of these companies, their integral role in providing services during lockdowns, and the general public's heightened reliance on and interest in technology and e-commerce during the pandemic.

QQQ, a NASDAQ-100 index ETF that tracks the performance of some of the largest non-financial companies listed on the Nasdaq stock exchange, stood out as one of the securities most impacted by sentiment scores. The histogram we've obtained clearly demonstrates the heightened role that public sentiment played in the price movement of QQQ during the COVID-19 pandemic. This pronounced effect is particularly coherent with the context of the pandemic, which saw a surge in reliance on technology as remote work, digital services, and e-commerce became not just conveniences but necessities. The companies within the QQQ ETF, predominantly from the technology sector, were in the spotlight, as their services were crucial for the continuity of both work and social life in lockdown scenarios.

VI. Conclusion

The diverse outcomes observed across the companies we analyzed—Apple, Amazon, Microsoft, and Facebook—underscore the complex and variable nature of the relationship between Twitter sentiment and stock market predictions. While sentiment data improved prediction accuracy for Amazon, Microsoft, and Facebook, it had the opposite effect for Apple, suggesting that the impact of sentiment on stock prices is contingent on the specific context and characteristics of each company.

These findings enrich our understanding of the potential utility of social media sentiment analysis in financial forecasting. They suggest that sentiment analysis can be a valuable component of stock market prediction models, but its effectiveness is not uniform across all companies. For practitioners and researchers alike, these results emphasize the need for a nuanced and company-specific approach when incorporating sentiment data into predictive models.

As we continue to explore the intersection of social media sentiment and financial market movements, it becomes increasingly clear that the incorporation of sentiment analysis into forecasting models requires careful consideration of the unique attributes of each company and the broader market dynamics. This tailored approach is essential for harnessing the full potential of sentiment analysis in enhancing the accuracy of stock price predictions.

VII. Future Directions

Our research has unveiled the significant role of Twitter sentiment in influencing stock price movements, setting a foundation for future exploration aimed at enriching our understanding through broader and more sophisticated data collection and analysis methods. We plan to extend our investigation by acquiring a larger historical dataset of tweets to discern more complex sentiment-stock market relationships over extended periods and through varying market conditions. Additionally, we aim to diversify our sentiment data sources beyond Twitter, incorporating insights from platforms like Reddit[15], TikTok[16], Instagram[17], and Mastodon[18] where vibrant communities engage in financial discourse. This multi-platform approach promises a more holistic view of public sentiment's impact on

[15] https://reddit.com/
[16] https://tiktok.com/
[17] https://instagram.com/
[18] https://joinmastodon.org/

stock prices, allowing for a nuanced analysis of investor behavior and market trends.

To capitalize on this expanded dataset, we intend to apply advanced machine learning models that can more accurately capture the nuanced dynamics between sentiment and stock prices. Exploring state-of-the-art algorithms and innovative AI technologies, such as deep learning and natural language processing, will likely enhance our predictive model's accuracy. In particular models like ARIMA and BERT could be applicable. Additionally, more powerful network analysis techniques can be applied such as the Partition Decoupling Method [16]. Moreover, by conducting a cross-platform sentiment analysis and considering an interdisciplinary approach that integrates financial analysis with behavioral economics, we aim to develop predictive models that reflect the complex interplay of market forces and human psychology.

ACKNOWLEDGMENT

We would like to acknowledge Professor Daniel Rockmore, Professor of Mathematics at the Department of Mathematics, Dartmouth College, for this support throughout this research project. We also thank the Feldberg Library librarians for their support in collecting large amounts of financial data from Wharton Research Data Services.

REFERENCES

[1] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," in *The Journal of Finance*, vol. 25, no. 2, pp. 383-417, May 1970. DOI: 10.2307/2325486. [Online]. Available: https://www.jstor.org/stable/2325486?origin=crossref. [Accessed: 01- Mar- 2024].

[2] A. Giachanou and F. Crestani, "Like It or Not: A Survey of Twitter Sentiment Analysis Methods," in *Proceedings of the ACM on Conference*. [Online]. Available: https://doi.org/10.1145/2938640. [Accessed: 01- Mar- 2024].

[3] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165-210, 2006. DOI: 10.1007/s10579-005-7880-9. [Accessed: 01- Mar- 2024].

[4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1-8, 2011. [Accessed: 01- Mar- 2024].

[5] B. Dickinson and W. Hu, "Sentiment analysis of investor opinions on twitter," *Social Networking*, vol. 4, no. 03, p. 62, 2015. [Accessed: 01- Mar- 2024].

[6] A. Mittal and A. Goel, "Stock Prediction Using Twitter Sentiment Analysis," Stanford University, 2011. [Online]. Available: https://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf. [Accessed: 01- Mar- 2024].

[7] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements," in 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7955659. [Accessed: 01- Mar- 2024].

[8] Li Bing, Keith C.C. Chan, and Carol Ou, "Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements," in IEEE, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6982085. [Accessed: 01- Mar- 2024].

[9] J.-X. Liu, J.-S. Leu, and S. Holst, "Stock price movement prediction based on Stocktwits investor sentiment using FinBERT and ensemble SVM," *PeerJ Computer Science*, vol. 1403, DOI: 10.7717/peerj-cs.1403. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37346695/. [Accessed: 01- Mar- 2024].

[10] N. Das, B. Sadhukhan, T. Chatterjee, and S. Chakrabarti, "Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak," DOI: 10.1007/s13278-022-00919-3. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9325657/. [Accessed: 01- Mar- 2024].

[11] R. A. Mendoza-Urdiales, J. A. Núñez-Mora, R. J. Santillán-Salgado, and H. Valencia-Herrera, "Twitter Sentiment Analysis and Influence on Stock Performance Using Transfer Entropy and EGARCH Methods," National Library of Medicine, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9324505/. [Accessed: 01- Mar- 2024].

[12] Y. Cui and L. Liu, "Investor sentiment-aware prediction model for P2P lending indicators based on LSTM," *PLoS One*, vol. 17, no. 1, Art. no. e0262539, Jan. 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35085306/. DOI: 10.1371/journal.pone.0262539. [Accessed: 01- Mar- 2024].

[13] B. Taborda, A. de Almeida, J. C. Dias, F. Batista, and R. Ribeiro, "Stock Market Tweets Data," IEEE Dataport, Apr. 15, 2021. [Online]. Available: https://dx.doi.org/10.21227/g8vy-5w61. [Accessed: 01- Mar- 2024].

[14] Wharton Research Data Services, "WRDS," Wharton School of the University of Pennsylvania. [Online]. Available: https://wrds.wharton.upenn.edu.[Accessed: 01- Mar- 2024].

[15] U. von Luxburg, "A Tutorial on Spectral Clustering," Max Planck Institute for Biological Cybernetics, Technical Report No. TR-149, Aug. 2006. [Accessed: 01- Mar- 2024].

[16] R. Braun, G. Leibon, S. Pauls, and D. Rockmore, "Partition Decoupling for Multi-gene Analysis of Gene Expression Profiling Data," 2011. [Online]. Available: arXiv:1002.3946 [q-bio.QM]. [Accessed: 01-Mar-2024].

APPENDIX

Continuing our analysis with additional companies, we extended our investigation to include Microsoft and Facebook, further examining how the inclusion of sentiment data impacts the predictive accuracy of our LSTM models.

A. Microsoft's Stock Predictions

For Microsoft, the incorporation of the `total_sentiment` feature yielded the following results:

- Test RMSE for Microsoft: 1.7931
- Test MAE for Microsoft: 1.335

Interestingly, when the sentiment data was excluded from the model, a slight increase in both RMSE and MAE was observed:

- Without sentiment, Test RMSE: 1.79
- Without sentiment, Test MAE: 1.390

These results suggest that for Microsoft, sentiment data has a marginal but positive effect on the model's ability to predict stock prices accurately. The minor improvement with sentiment data indicates that while sentiment may not be a major determinant of Microsoft's stock movements, it does contribute beneficially to the model's predictive capabilities.

B. Facebook's Stock Predictions

Turning our attention to Facebook, the impact of sentiment data was more pronounced:

- With sentiment, Test RMSE for Facebook: 4.024
- With sentiment, Test MAE for Facebook: 3.316

Comparatively, the model's performance dipped when the sentiment score was excluded:

- Without sentiment, Test RMSE: 4.432

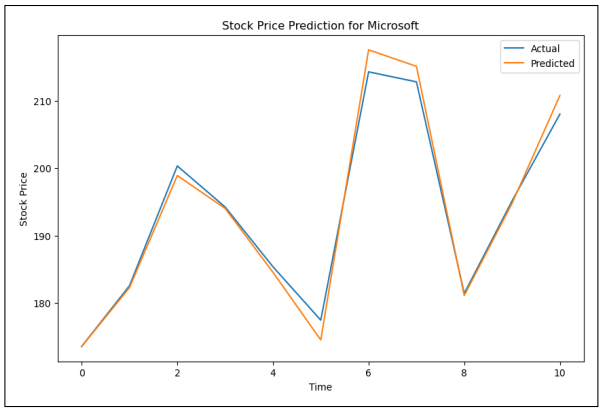

Fig. 23. Microsoft Stock Price Prediction with Sentiment

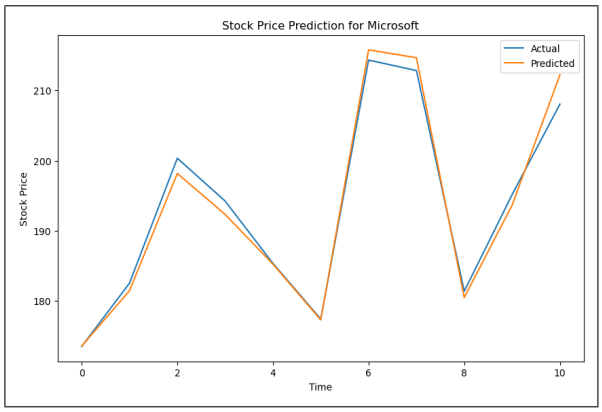

Fig. 24. Microsoft Stock Price Prediction without Sentiment

- Without sentiment, Test MAE: 3.86

For Facebook, the inclusion of sentiment scores notably enhanced the model's forecasting accuracy, as indicated by the lower RMSE and MAE values. This improvement underscores the relevance of social media sentiment in predicting Facebook's stock price movements, highlighting the importance of public perception and sentiment trends in shaping the company's market valuation.

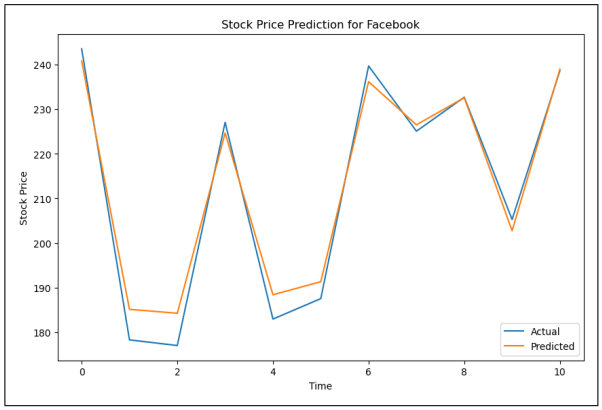

Fig. 25. Facebook Stock Price Prediction with Sentiment

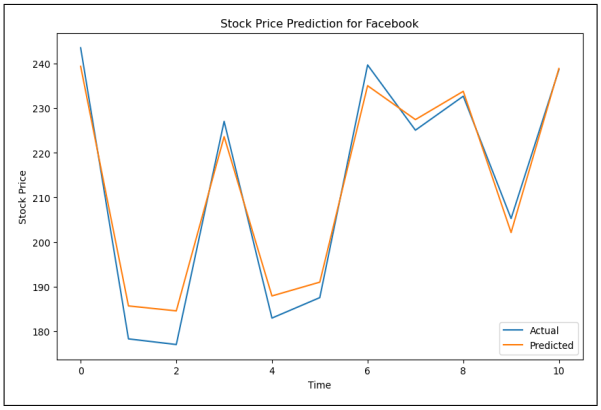

Fig. 26. Facebook Stock Price Prediction without Sentiment