# Stock Price Forecasting using Opinion Leader Sentiment Analysis

Alan Kassymkanov
*Computer Science Dept.*
*University of Richmond*
Richmond, VA
alan.kassymkanov@richmond.edu

Naron Chen
*Computer Science Dept.*
*University of Richmond*
Richmond, VA
naron.chen@richmond.edu

Luoli Wang
*Computer Science Dept.*
*University of Richmond*
Richmond, VA
luoli.wang@richmond.edu

Wonyoun Kim
*Computer Science Dept.*
*University of Richmond*
Richmond, VA
wonyoun.kim@richmond.edu

## I. Abstract

The primary objective of this project is to enhance the accuracy of stock price forecasting models for leading tech companies by incorporating the sentiments of influential opinion leaders. To achieve this, the project will employ sentiment analysis techniques on social media data, specifically from Twitter, in conjunction with historical stock data to create the model. The team will apply various regression and classification models to predict stock prices and assess their performance using diverse metrics. The anticipated outcomes of this project include establishing a correlation between social media sentiment and stock prices, and investigating the efficacy of integrating opinion leader sentiment into stock price forecasting models.

### A. Introduction

Stock price forecasting plays a pivotal role in the financial decision-making process for investors and traders, particularly in the rapidly evolving tech industry. Social media has emerged as a vital source of information for investors, offering a platform for users to express their opinions and feelings about companies and their products. This project aspires to devise a more accurate stock price forecasting model for prominent tech companies by incorporating the sentiments of opinion leaders on social media platforms. We will conduct sentiment analysis on Twitter data, combined with historical stock data, to construct the model. Various regression and classification models will then be utilized to predict stock prices, and their performance will be evaluated using a range of metrics. The expected results of this project will demonstrate a correlation between social media sentiment and stock prices, while also exploring the effectiveness of integrating opinion leader sentiment in stock price forecasting. The insights generated from this project have the potential to offer valuable information to investors and traders, enabling them to make well-informed decisions based on market sentiment.

### B. Related Work

@article mittal2012stock, title=Stock prediction using twitter sentiment analysis, author=Mittal, Anshul and Goel, Arpit, journal=Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), volume=15, pages=2352, year=2012
The authors of this paper used a dataset of Dow Jones Industrial Average values from June to December 2009 and publicly available Twitter data containing more than 476 million tweets. This is different from our datasets date range. They preprocess the data by filling gaps in the DJIA data using a concave function, adjusting for sudden jumps/falls in the stock values, pruning periods of volatile activity, and computing the z-score of each point in the data series. The authors use Self Organizing Fuzzy Neural Networks (SOFNN) to predict DJIA values using previous values and predicted mood values. We will mention later in the section how the result of this paper compares to ours.

### C. Data Collection

The data collection step involves exploring and extracting the valid information from the Twitter post. Based on the dataset *"Tweets about the Top Companies from 2015 to 2020"* from Kaggle, we extracted tweets each with a company tag (stock ticker symbol) associated. Each tweet has its post time, tweet content, tweet_id, writer_id, comment_num, retweet_num, and like_num. We then connected to the twitter API and extracted each tweet writer's follower count and at the mean time verified the writers with twitter's database. We also extracted all the finance daily price data from Yahoo Finance. We will expand on this part in the Data Visualization & Feature Extraction section.

### D. Data Cleaning

Firstly, the dataset was checked for duplicate tweets. Duplicate tweets can lead to biased results and can inflate the importance of certain tweets in the analysis. Therefore, any repeated tweets were removed from the dataset.

Secondly, tweets with invalid usernames were removed from the dataset. These tweets are typically uninformative

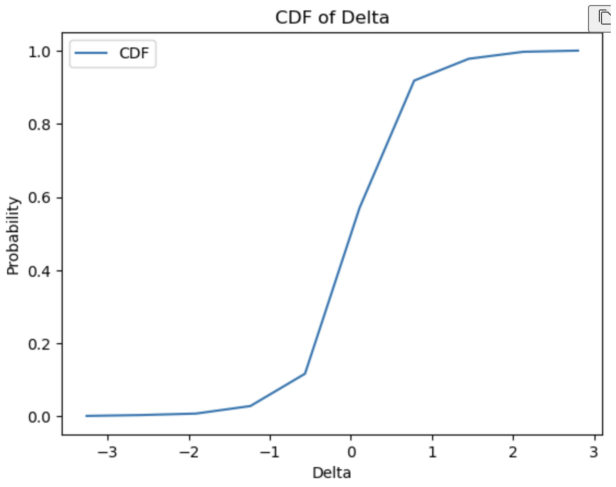
and can be detrimental to any analysis or modeling efforts. Tweets with invalid usernames may indicate that the tweet was generated by a bot or a spam account, which can also introduce bias into the datasets.

Thirdly, the dataset was checked for any missing values or inconsistencies. Missing values can be problematic, especially when dealing with large datasets, and can lead to inaccurate results. Inconsistencies in the dataset can also introduce bias and can make it difficult to perform accurate analysis or modeling. Therefore, any missing values or inconsistencies were identified and addressed.

### *E. Data Visualization & Feature Extraction*

For our training data, we have selected the top tech companies as we believe that the fast and dynamic nature of their stock prices makes them more responsive to market sentiment on a daily basis. To enrich our dataset with additional information, we have included a follower count column. This metric provides a direct indication of whether a tweet was written by an opinion leader, and we have chosen to perform sentiment analysis on opinion leaders as we believe that opinions are often led by influential individuals rather than independent users. On the other hand, by using the sentiment of opinion leaders, we can reduce the noise in our dataset and achieve more accurate results.

To obtain the follower count data, we connected to Twitter using a developer account. For each tweet-id and unique writer's name in our dataset, we fetched the follower count of each writer individually. We obtained the daily price of the 5 different stock markets from Yahoo Finance. And plotted a CDF graph where Delta here refers to price movement from yesterday to today.

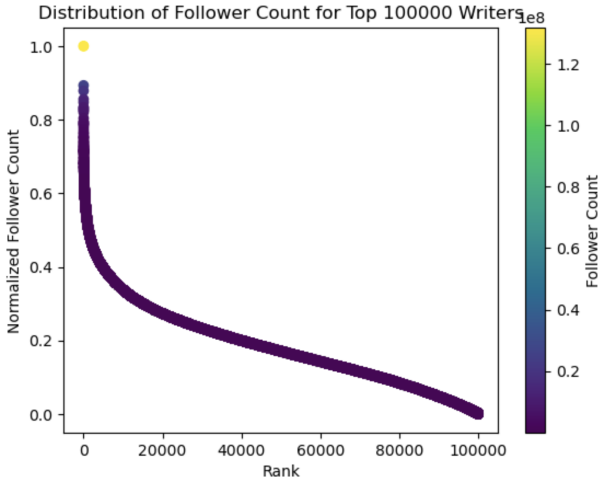



One key observation can be made from the visualizations above. The price movement of the selected stocks mostly falls within a small range of -1 to 1, which is significantly smaller in proportion to the overall stock price range of 15 to 40 dollars per stock. This indicates that a highly precise prediction model is required to achieve accurate results.The sorted follower count graph demonstrates that the majority of Twitter users possess a relatively low number of followers, as evidenced by the graph's steep slope. As per a report published by Sysomos, an estimated 93.6% of Twitter users have fewer than 100 followers. While this information may not be completely up-to-date, it serves as a valuable reference point for establishing a follower count threshold. Consequently, we have set our threshold at 100 followers to guarantee that our analysis concentrates on opinion leaders who wield a considerable degree of influence.

It is essential to acknowledge that our dataset factors in the stock market's closing time, which occurs at 4:30 pm EST. Tweets published after this time serve solely as a feature for predicting future dates. Moreover, since the stock market remains closed on weekends and holidays, these dates were omitted from the dataset.

Additionally, we conducted lexicon-based sentiment analysis on each tweet to derive a compound score ranging between -1 and 1, which signifies the tweet's overall positive or negative sentiment. To enhance the sentiment analysis's accuracy concerning financial tweets, we incorporated a list of finance-specific lexicon into the dictionary. This inclusion enabled us to assign positive scores to terms such as "oversold," "buy," and "dividend," while allocating negative scores to words like "sell" and "short."

By expanding our sentiment analysis methodology with finance-specific lexicon, we were able to better capture the nuances of financial discussions on social media platforms. This refined approach allows for more accurate representation of opinion leaders' sentiments, which in turn contributes to a more reliable stock price forecasting model. Furthermore, by focusing on influential opinion leaders with a minimum of 100 followers, our analysis emphasizes the impact of their sentiments on stock prices, providing valuable insights for investors and traders.

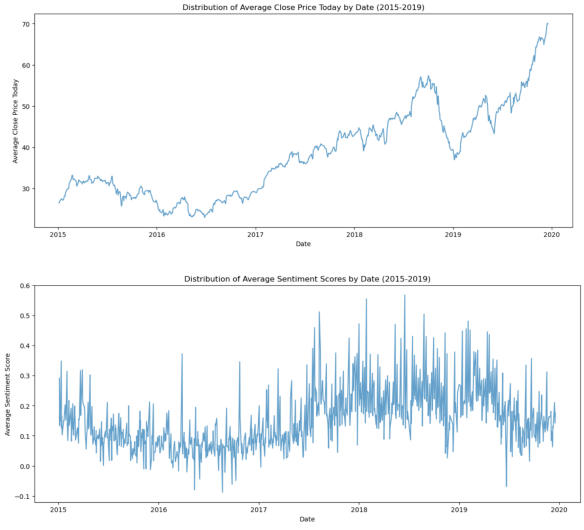


The sorted follower count graph demonstrates that the majority of Twitter users possess a relatively low number of followers, as evidenced by the graph's steep slope. As per a report published by Sysomos, an estimated 93.6% of Twitter users have fewer than 100 followers. While this information may not be completely up-to-date, it serves as a valuable reference point for establishing a follower count threshold. Consequently, we have set our threshold at 100 followers to guarantee that our analysis concentrates on opinion leaders who wield a considerable degree of influence.

It is essential to acknowledge that our dataset factors in the stock market's closing time, which occurs at 4:30 pm EST. Tweets published after this time serve solely as a feature for predicting future dates. Moreover, since the stock market remains closed on weekends and holidays, these dates were omitted from the dataset.

Additionally, we conducted lexicon-based sentiment analysis on each tweet to derive a compound score ranging between -1 and 1, which signifies the tweet's overall positive or negative sentiment. To enhance the sentiment analysis's accuracy concerning financial tweets, we incorporated a list of finance-specific lexicon into the dictionary. This inclusion enabled us to assign positive scores to terms such as "oversold," "buy," and "dividend," while allocating negative scores to words like "sell" and "short."

By expanding our sentiment analysis methodology with finance-specific lexicon, we were able to better capture the nuances of financial discussions on social media platforms. This refined approach allows for more accurate representation of opinion leaders' sentiments, which in turn contributes to a more reliable stock price forecasting model. Furthermore, by focusing on influential opinion leaders with a minimum of 100 followers, our analysis emphasizes the impact of their sentiments on stock prices, providing valuable insights for investors and traders.




Upon examining the data visualization on Apple Inc as an illustration, it is evident that the correlation between stock price trends and average sentiment scores is not immediately apparent. Nevertheless, a subtle positive relationship can be observed between mid-2017 and late 2018, during which Apple's stock price experienced consistent growth. In late 2018, both sentiment scores and stock prices declined, likely due to the escalating US-China trade war and concerns surrounding the company's future growth. The compound sentiment score from Twitter contains a significant amount of noise, which we will address in the subsequent sections to enhance the accuracy and reliability of our analysis.

### *F. Paper Used (optional, if using an existing paper)*

@article mittal2012stock, title=Stock prediction using twitter sentiment analysis, author=Mittal, Anshul and Goel, Arpit, journal=Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), volume=15, pages=2352, year=2012

The authors of this paper used Self Organizing Fuzzy Neural Networks (SOFNN) to predict DJIA values using previous values and predicted mood values. They obtain 75.56% accuracy using a new cross-validation method for financial data. They also devise a naive strategy to maintain a profitable portfolio based on their predicted values. The authors' work is based on Bollen et al.'s strategy, which predicted the behavior of the stock market by measuring the mood of people on Twitter, and achieved an accuracy of nearly 87% in predicting the up and down changes in the closing values of Dow Jones Industrial Index (DJIA).

Self Organizing Fuzzy Neural Network (SOFNN) algorithm performed better than our model, giving nearly 75.56% accuracy. In contrast, the accuracy results obtained from the models in our project were generally lower, with the highest accuracy of 68% obtained from linear regression on a more generalizable setting. But our LSTM model achieved a much higher accuracy based heavily on previous stock price.

Their study digged deep into finding the characteristics of Calmness and Happiness , which were more predictive of stock values, confirming the Granger causality analysis. In comparison, the our project focused on different features and did not explore the in-depth relationship between moods and stock prices. However, our project was greatly inspired by this paper in exploring the correlation between stock price and social sentiment, we explored the LSTM approach in the project which was not mentioned in the previous one.

Despite the lower accuracy rates obtained in our project, the inclusion of the LSTM approach highlights the importance of exploring different machine learning algorithms and models to achieve better results. The LSTM model has shown great success in handling time series data and has been applied in various fields, including finance. The use of LSTM in our project allowed us to capture the recent market trends and fluctuations in the data and use them to predict future stock prices, which could potentially improve the accuracy of stock

price prediction. Overall, while our project's approach and focus differed from the previous study, both works highlight the importance of experimentation and exploration in developing more accurate and reliable models for predicting stock prices.

### G. Approach

To create a comprehensive dataset, we combined both Twitter data and historical stock price data. Using this dataset, we attempted to solve both a regression problem (predicting the actual stock price) and a classification problem (predicting whether the stock price increased or decreased). For the regression problem, we used linear regression, gamma regression, and RNN + LSTM models to predict the stock price for a given day based on the price of the stock for the past few days and relevant Twitter data, such as sentiment scores, number of likes, number of followers, and number of retweets.

For the classification problem, we created a new column in the dataset that indicated whether the stock price increased or decreased, with 1 representing an increase and 0 representing a decrease. To predict this binary outcome, we used SVM, logistic regression, Deep Neural Network, and KNN models. Overall, our approach demonstrates a thorough exploration of various regression and classification models, indicating a strong understanding of the strengths and weaknesses of different machine learning techniques. By combining both Twitter and historical data, we attempted to show a willingness to experiment with alternative data sources beyond traditional financial data to improve the accuracy of stock price predictions.

### H. Methodology

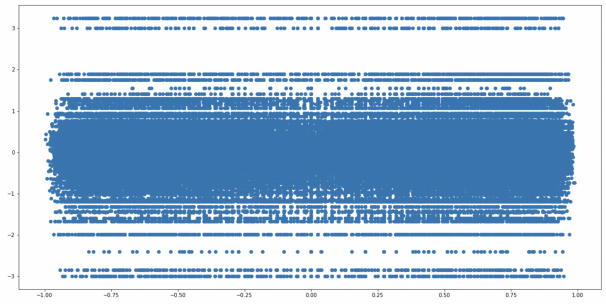
We explored correlation of the variables. In this project, as we've already mentioned, we aimed to predict stock prices using a combination of historical price data and sentiment scores from Twitter posts. To accomplish this, we performed sentiment analysis on Twitter data to calculate sentiment scores for each post, which were then used as features in a machine learning model.

One of the main objectives of our project was to determine if there was a correlation between the change in stock price and the sentiment scores from Twitter posts. We attempted to visualize this relationship using a scatter plot, but unfortunately, the plot showed a very non-linear relationship between the two variables, with no clear correlation. This indicates that sentiment scores alone may not be sufficient for accurately predicting stock prices.

While our results may not have supported our initial hypothesis, they still provide valuable insights into the limitations of using sentiment analysis as a predictor of stock prices.

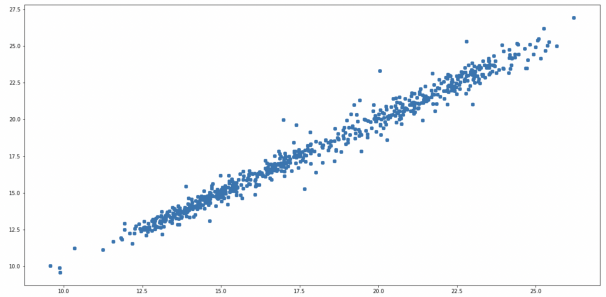
Moreover, we calculated the Person's correlation score for the Sentiment Score vs Change in Price (we created a new column by subtracting yesterday's price from today's price). As we expected after looking at the scatter plot, the correlation factor is very low. It is equal to 0.0023. We can assume that these 2 random variables are uncorrelated. However, since lack of correlation does not imply independence, we still continued our attempts with modeling using sentiment score as a feature.

Plot of the Compound Score vs Change in the price:



In our project, we observed a strong linear relationship between the price of the stock yesterday and the price of the stock today. This is evident from the scatter plot that we generated (below), which showed a very strong positive correlation between these two variables. The calculated correlation coefficient of 0.9845 further supports this observation, indicating a high degree of correlation between the two variables.

This finding highlights the importance of considering historical stock prices when attempting to predict future prices. Overall, our observation of a strong linear relationship between yesterday's and today's stock prices provides valuable insights into the underlying patterns and trends in the stock market. However, note that this is the result for a fixed company, obviously.



**Linear Regression**

We used this model twice, with two different approaches. First, we trained the model on data from the same company and tested it on a separate test set from the same company. This approach yielded impressive results, with an $R^2$ score of 99%. However, we recognized that this high score could be attributed to the strong correlation between one of the features (i.e. the stock price from the previous day), as we have argued above. Therefore, we decided to test the model on a more generalizable setting, where we trained it on data from four different companies and tested it on a fifth company (note that the ration is 4:1 which is equivalent to 80%:20%).

This second approach resulted in a decreased $R^2$ score of 68%. While this score may seem lower than the previous result, it is a more realistic representation of the model's ability to generalize to new data. This approach also highlights the importance of testing models on data from multiple sources to ensure that the model's predictive power is not limited to a single company or specific set of data. Overall, our analysis suggests that linear regression can be an effective tool for predicting stock prices, but careful consideration should be given to the choice of features and the generalizability of the model to new data.

**Gamma Regression**

After our initial success with linear regression, we decided to explore the use of another type of generalized linear model, specifically gamma regression. We initially thought this was a reasonable choice since gamma regression is well-suited for data where the target value is always non-negative, as gamma distribution is nonnegative. However, our results with gamma regression were not satisfying. We used a similar approach as with linear regression, splitting the data for the same company and testing it on a separate test set from the same company. Unfortunately, the $R^2$ score we obtained using gamma regression was only 0.18 for the same company.

We also tested the model on a more generalizable setting, using the same approach of training on data from four different companies and testing it on a fifth company. However, our results were even worse, with an $R^2$ score of -0.84. This result suggests that gamma regression was not an effective tool for predicting stock prices in our project, and it highlights the importance of experimenting with different models and approaches to find the most suitable one for the specific data and problem at hand.

**Logistic Regression** As we've described above, to run a logistic regression, we created a new column representing 1 when price increased and 0 when it decreased. The result obtained from logistic regression with an accuracy score of 0.503767 was not ideal for our stock price prediction task. While it is slightly better than randomly guessing, an accuracy score of 0.5 indicates that the model is not performing significantly better than chance.

However, it is important to note that logistic regression is a relatively simple model and can be useful as a baseline for comparison with more complex models. Further improvements can be made by fine-tuning the logistic regression model with more relevant features or by exploring other classification models (what we did in the next steps). Here are some of the coefficient estimations of Logistic Regression:

Accuracy: 0.503767267756353

| | comment_num | retweet_num | like_num | follower_count | compound_score | Close_price-today | Close_price-tmr | intercept | Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| coeff_estimates | 0.246896 | -0.905442 | 0.489606 | -0.809247 | -0.029504 | 2.070092 | -1.890859 | 0.006019 | 0.503767 |

**Support Vector Machine** We turned to SVM after the Logistic Regression model failed to produce satisfactory results. SVM is known for its ability to handle complex datasets and to create a clear separation between classes. SVM tries to find the optimal hyperplane that maximizes the margin between two classes.

In our project, we experimented with three types of kernels: linear, polynomial, and Gaussian, to determine which one would produce the best results. We also varied the regularization parameter C, which controls the trade-off between maximizing the margin and minimizing the classification error. After trying different combinations of kernels and C values, we found that the best result was obtained with a Gaussian kernel and C=100. The accuracy result is 0.5167. While the accuracy score improved compared to the Logistic Regression model, it is still far from being the best-performing model in our project. Nonetheless, the use of SVM shows the importance of trying different machine learning algorithms to find the best one for a particular task.

| | 1 | 10 | 100 | 1000 |
|---|---|---|---|---|
| **linear** | 0.514 | 0.514000 | 0.514000 | 0.515000 |
| **poly** | 0.514 | 0.514333 | 0.514667 | 0.514667 |
| **gauss** | 0.514 | 0.516000 | 0.516667 | 0.507333 |

**Gaussian Discriminant Analysis** Gaussian Discriminant Analysis (GDA) is another classification approach that was used in the project. The decision to use this model was driven by two reasons. Firstly, it was to assess if the GDA model performs better than logistic regression. Secondly, it was based on the assumption that if the feature space is normally distributed, GDA would perform better than logistic regression (since this model assumes it). This hypothesis was tested to see if it holds in this context.

The result of the GDA model was 51%, which is a slight improvement over the logistic regression model. However, the score is still far from a good one. This implies that there is not enough evidence to confirm that the feature space is approximately Gaussian distributed. Nonetheless, the use of GDA provides some insight into the nature of the distribution of the features. In conclusion, the use of different classification models in this project allowed for a comparative analysis of their performance, which provided valuable insights into the data and the underlying distribution of the features.

**KNN agorithm** We decided to utilize the method of K-Nearest Neighbors (KNN) algorithm approach in order to predict analyze the data because of it's distance metric which is great for analyzing numerical data such as the price and it works well with linear data. These characteristics of KNN drew our focus to try and analyze our data using KNN algorithm. We first started used the sklearn library MinMax Scalar in order to scale the features before training the model and applying it to the test dataset. We then ran the algorithm.

It was here where we saw the drawbacks of KNN overshadow the benefits severely. We had a score of -1.145 which is not the result we were looking for. Looking for ways to improve our accuracy, we decided to try and us GridSearchCV to tune the hyper parameters of the KNN model including the number of neighbors, the weight function, and the distance metric. We used the best estimator obtained from the GridSearchCV to build a BaggingRegressor model, which combines multiple KNN models to improve and reduce overfitting. We used BaggingRegressor model because it is particularly useful when the underlying model is unstable or has a high variance. Testing the data after GridSearchCV and BaggingRegressor, we were given the score of -0.26. A good improvement but still not where we wanted it to be. Overall, the scores were not what we wanted but we were able to learn about how we can continually improve the accuracy of KNN through trial and error. If time allowed, we would have looked into other methods of KNN such as different scalars, pipeline methods, etc.

**Neural Network** Neural Networks have emerged as a promising solution to classification problems and have been found to handle large datasets and complex relationships between input and output variables due to their brain-inspired architecture. In this case, a neural network model is trained using a combination of stock prices from five companies, namely Apple, Amazon, Google, Software, and Tesla, with 70% of the data used for training and 30% for testing and validation.
The model has two hidden layers with 32 neurons each, using the ReLu activation function that is commonly used for non-negative features. However, despite the model's efforts to improve performance, the evaluation results show an accuracy rate of only around 0.52, which is only slightly better than the Support Vector Machine approach. Both approaches have shown limited success in this problem, indicating the need for further research and experimentation to achieve better results.

**LSTM** This study proposes a time series analysis approach using LSTM models to predict stock prices. Time series analysis is a widely used technique in financial analysis to study patterns and trends in a set of data points over time. By utilizing the stock prices from the previous two days as time series features, this study aims to capture the recent market trends and fluctuations in the data. The LSTM model is a powerful artificial neural network used for analyzing time series data, from which it handle long-term memory and short-term memory separately. Moreover, it can handle the issue of exploding or vanishing gradients by using regular recurrent neural network. To train the LSTM model, the previous two days' stock prices of Apple are used as the input data. By training the LSTM model on this historical data, the model can recognize patterns and trends in the data and use them to predict future stock prices. In the next step of the study, the trained LSTM model is employed to make predictions for Google's stock price. The model takes the previous two days' stock prices of Google as input and uses the learned patterns and trends from the training phase to generate a prediction for the current stock price. To evaluate the LSTM model's performance, the study employs the squared root of the mean squared error between the predicted stock price and the actual stock price. The mean squared error measures the average squared difference between the predicted and actual values. By taking the square root of the mean squared error, the study obtains a measure of the average error in the units of the stock price. The evaluation of the LSTM model's performance yields an error of 2.002 for the entire dataset. This indicates that, on average, the model's predictions deviate from the actual stock prices by approximately $2.002. While this error may seem relatively small, it is important to note that even small errors can have significant implications for financial decision-making and investment strategies. Overall, the study's evaluation of the LSTM model's performance on the prediction task provides valuable insights into the model's accuracy and generalization ability. The low error rate suggests that the LSTM model may be a promising tool for predicting stock prices, with practical implications for the financial industry.

### *I. Results*

| Regression Results (R^2 Scores) | | |
|---|---|---|
| Algorithm | Score | |
| Linear Regression | 0.99 (Test on Same Company) | 0.63 (Test on fifth Company) |
| Gamma Regression | 0.18 (Test on Same Company) | -0.84 (Test on fifth Company) |
| KNN | -1.145 (MinMax Scalar) | -0.26 (GridSearchCV, BaggingRegressor) |
| Long Short Term Memory | 2.002 (Means Squared Error) | N/A |

| Classification Results (Accuracy) | | |
|---|---|---|
| Algorithm | Score | |
| Support Vector Machine | 0.515 (Linear, C = 1000), 0.514667 (Poly, C = 100), 0.516667 (Gaus, C = 100) | N/A |
| Neural Networks | 0.52 | N/A |
| Logistic Regression | 0.503767 | N/A |
| Gaussian Discriminant Analysis | 0.51 | N/A |

LSTM results:


As we can see, only LSTM did a good job here. This means that stock price movement is a very complex task, which can be solved with good accuracy only using one of the most complicated methods, such as LSTM.

## II. Individual Team Member's Responsibilities

Naron: Data Collection, Data Cleaning, Feature Extraction, Data Visualization, Linear Regression & KNN Model
Alan: Feature Extraction, Data Visualization, Correlations, Linear & Gamma Regression, Gaussian Discriminant Analysis, RNN, Random Stock Generation.
Tony: Logistic Regression, Support Vector Machine, Neural Network, Long Short-Term Memory, Data Visualization, Model Evaluation.

Wonyoun: K Nearest Neighbors algorithm, Models Evaluations and Testings.

## III. Optional-Above and Beyond

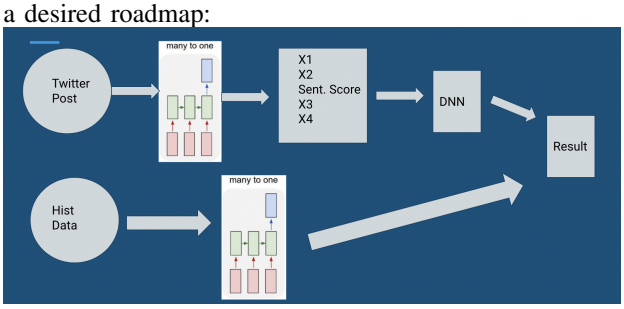
RNN+LSTM, Gamma Regression, Gaussian Discriminant Analysis.

## IV. Conclusion and Future Work

**Conclusion** In this project, we have investigated various machine learning models and techniques to predict stock prices using a combination of historical stock prices and Twitter data. Our results demonstrated that linear regression and LSTM models showed the most promising results, while other models like gamma regression, logistic regression, support vector machines, Gaussian Discriminant Analysis, and K-Nearest Neighbors algorithm yielded less satisfactory outcomes.

It is important to note that our study does not take into account many factors, such as the demographics of Twitter users, the languages spoken, or the direct correlation between those who invest in stocks and those who use Twitter. Nonetheless, our findings might provide valuable insights into the potential of using machine learning models and large-scale social media data to predict stock prices, offering a foundation for further research and development in this area.

**Future Work** There are two avenues we plan to pursue to continue developing our project and improving our results. Firstly, we aim to enhance the LSTM model by incorporating Twitter features. We plan to explore two different approaches. The first is to run the LSTM solely on historical stock price data. The second is to use a deep neural network to analyze Twitter data. Once we have these two sets of results, we will investigate different methods to combine them in a meaningful way, based on empirical research. Moreover, we'll do more feature engineering to get more important "finance" data (for example, volatility of the stock). Here is a desired roadmap:




Second approach:
We want to explore the idea that stock price movement follows a Brownian motion(stochastic motion sampled from Normal Distribution) with sudden jumps that follow a Poisson process. Based on this assumption, we plan to use density estimation techniques to estimate the parameters of the distributions and predict future stock prices.
To do this, we will use historical stock prices and apply density estimation methods such as maximum likelihood estimation or kernel density estimation. By estimating the underlying distribution of stock prices, we hope to gain insight into the behavior of the market and make more accurate predictions.
We also plan to compare the results of this approach with other methods we have tried. This will allow us to determine which approach is most effective in predicting stock prices and potentially improve our understanding of market behavior.

## V. References

Omer Metin. 2020. Tweets About the Top Companies From 2015 to 2020 [Data set]. Kaggle. https://www.kaggle.com/datasets/omermetinn/tweets-about-the-top-companies-from-2015-to-2020?select=Tweet.csv

sysomos, revised April 2014, published Jun 2009, Inside Twitter: An In-Depth Look Inside the Twitter World, https://www.sykey4biz.it/files/000270/00027033.pdf