**National Research University Higher School of Economics**

**Faculty of Computer Science**
**HSE and University of London Double**
**Degree Programme in Data Science and Business Analytics**

# BACHELOR'S THESIS
# (Research Project)
# Analysis of the Influence of Textual Information on the Stock Prices

**Prepared by the student of Group 192, Year 4 (year of study),**
**Danish Madhavan**

**Thesis Supervisor:**
**MSc CS, Lecturer, Faculty of Computer Science, Kirill Rudakov**

**Moscow**
**2023**

## Table of content

## Abstract

Stock price prediction has always been a very interesting task for researchers and investors since the accurate prediction of the stock can lead to significant gain in the profit from investing in the stock market. It is quite common for investors to base their decision on the public perception of the particular company, recent news or political events.

In my thesis, we will try to investigate whether peoples' opinion regarding a particular company together with news headlines can be sufficient factors which should be taken into account when predicting the stock price. We will test if adding these features will enhance the performance of the machine learning algorithms.

The results show that incorporating sentiment of Tweets is a significant factor which can be used to predict the stock prices while news headlines regarding the companies have moderate success in enhancing performance of models. Political news did show any reasonable improvement to machine learning models.

## Introduction

The stock market plays a pivotal role in the economy since it gives people investing opportunities. Accurately predicting stock price movements remains a formidable challenge because of volatility that emerges due to many factors like political events and social media. In recent years, social media platforms, particularly Twitter with its vast user base and daily tweet volume, have gained a lot of popularity and, thus, a lot of people make their opinions on the direction of the stock prices based on what they read there.

Analyzing a company's performance and its stock is crucial before making investment decisions. Sentiment analysis of social media, particularly Twitter, has emerged as a valuable source of information that can provide insights into positive or negative sentiments associated with stocks and trends and help the investor understand what the general attitude towards a particular company is. Prior studies have leveraged historical or textual data, employing techniques such as technical analysis, deep learning, and regression analysis. However, these studies often overlooked the inclusion of external factors, such as social media sentiment, despite its potential to significantly influence stock prices and trends. This is based on the understanding that stock prices are impacted by human behavior, which can be reflected through social media platforms.

Various studies have explored sentiment analysis of Twitter feeds, introducing machine learning methods encompassing sentiment classification techniques, Convolutional Neural Networks (CNNs), LSTM and so on. Nonetheless, research integrating both social media and historical data still offers significant opportunities for improvement. Previous studies have utilized Twitter data and stock market information to predict trends and prices using machine learning algorithms. Those models had promising results. However, a lot of studies considered only a portion of possible textual data that can influence stock price.

There were also studies that considered news headlines as a different source of textual information that may influence that stock price. Some of them showed that incorporating this source of information can also boost performance of models which means that it is also an important factor for predicting the stock.

In my study we will use 3 sources of textual information: tweets about companies, news headlines about companies, political news headlines. In this work we will employ neural network models, specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, to achieve more precise predictions of stock market trends and prices. These models will be applied to forecast the movement and value of stock prices on the subsequent day using tweets from the current day pertaining to specific companies. We will also use a classical machine learning algorithm AdaBoost and compare results of neural networks with it. This study aims to test the hypothesis that all three sources of textual information will improve the performance of models.

## Literature review

Social media has gained a lot of popularity in a recent decade. This is the reason why more and more studies came out in recent years which consider the effects of social media and other textual information on the stock prices. Most of these studies add features related to text to their models and see if it can boost the performance.

One study by Singh and Kaur used sentiment analysis on Twitter data to predict stock prices. The authors collected tweets mentioning the companies and classified them as positive, negative, or neutral. They employed machine learning algorithms, including Random Forest and Naïve Bayes, to classify the sentiments of tweets. The results showed that the sentiment of tweets can be used to predict stock prices, with a high accuracy.

Another study by Mittal and Goel from Stanford University used a similar approach to predict stock prices using Twitter data. The authors used a sentiment analysis tool called SentiWordNet to determine the sentiment of tweets and then used machine learning algorithms to predict stock prices. The study found that Twitter sentiment can be used to predict stock prices with an improved accuracy.

Sengupta from the University of Reading employed deep learning methods to predict stock prices in the Indian market using Twitter sentiment analysis. The author used a convolutional neural network (CNN) to classify the sentiments of tweets and then predicted the stock prices using a long short-term memory (LSTM) model. The results showed that the CNN-LSTM model outperformed other models.

Cristescu et al. used sentiment analysis on market news articles to predict stock prices. The authors used machine learning algorithms, including Support Vector Machines (SVM) and Decision Trees, to classify the sentiment of news articles.

Kumar et al. conducted research on incorporating financial news sentiment into stock price prediction models. They employed a deep learning framework that combined convolutional neural networks (CNN) and LSTM networks to analyze news sentiment and predict stock prices. The results demonstrated significant improvements in metrics like RMSE and MAE, suggesting that the inclusion of sentiment information enhances the accuracy of stock price prediction models.

A study by Zhang et al. focused on incorporating social media data, particularly Twitter, into stock price prediction. The authors utilized a recurrent neural network (RNN) model with an attention mechanism to capture the sentiment from tweets and predict stock prices. The incorporation of Twitter sentiment led to substantial reductions in RMSE and MAE, indicating improved predictive accuracy.

Darapaneni et al. used sentiment analysis and deep learning to predict stock prices in the Indian market. To forecast the stock values of certain Indian companies, the authors combined sentiment analysis, time series analysis, and deep learning methods. The outcomes demonstrated that the suggested model is capable of correctly predicting stock values for Indian companies.

From the papers above we may conclude that sentiment analysis has emerged as a useful technique for predicting stock prices using social media data. While there are variations in the accuracy of the models developed in different studies, the results suggest that sentiment analysis can be a valuable tool for stock market prediction.

## Methodology

*Collection of tweets:*

The first step of this study was to collect and analyze tweets related to five major technology companies (Apple, Microsoft, Tesla, Amazon, and Google) from the years 2015 to 2020. Originally, snscrape was used to parse the Twitter website. For Apple, I had to use a query ‘Apple stock’ instead of ‘Apple’ due to the fact that there are innumerous tweets regarding the query ‘Apple’ and I had to reduce the time spent parsing the website. For the other companies a similar logic was used. However, the limited number of tweets was not really representative of the sentiment regarding the company and, moreover, at some point snscrape was no longer successful in Twitter parsing because of the new policy regarding Twitter API. That is why a different approach was used. There is a huge dataset in Kaggle regarding the tweets of top 5 companies from 2015 to 2020. The LSTM model was tested on the tweets that were already scraped and tweets from the Kaggle dataset. Tweets from Kaggle dataset were a better choice since they gave better RMSE and MAE results. It was likely the case due to the fact that the parsed dataset had a smaller number of observations. The number of tweets was large enough to get the idea of the overall sentiment towards a certain company. I decided that it would be a better idea to use it since it would have a better impact on the model. The tweets were preprocessed using a series of steps to remove URLs, special characters, and stop words. Sentiment analysis was performed on the preprocessed tweets using the VADER (Valence Aware Dictionary and sentiment Reasoner) sentiment intensity analyzer, which provides a score between -1 and 1 to indicate the positivity, negativity, or neutrality of each tweet. A total of more than 4 million tweets were collected and stored for further analysis.

Once the tweets were collected, they were preprocessed to remove noise and irrelevant information that could affect the sentiment analysis results. This preprocessing involved converting all text to lowercase, removing URLs and other web-related information, removing special characters, and removing stop words. The sentiment intensity analyzer was then applied to each tweet, generating a sentiment score for each one.

The sentiment scores were used to analyze the overall sentiment trends for each company over time. The sentiment scores were aggregated by day and visualized using line charts to show the sentiment trend over time. In addition, the sentiment scores were used to build a predictive model for each company's stock price using machine learning techniques.

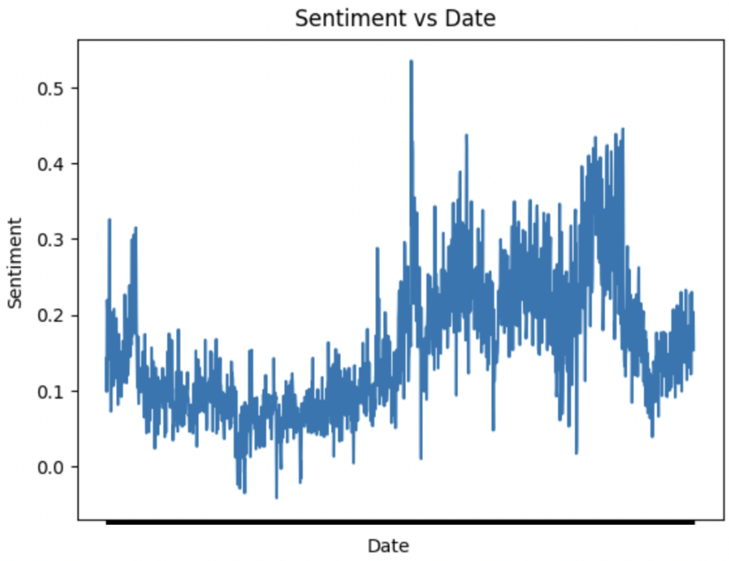


Figure 1. Sentiment of Apple by date

Here is a picture of a sentiment score of Apple plotted together with dates. You may see that there are several trends in the scores depending on the dates.

A variant of RoBERTa called Twitter-roBERTa-base was also considered to obtain the sentiment of tweets. It is based on the RoBERTa architecture, which is a transformer-based language model that was pre-trained on large amounts of textual data using a self-supervised approach. The Twitter-roBERTa-base model was fine-tuned on a large dataset of annotated tweets to improve its performance on sentiment analysis tasks. However, using Twitter-roBERTa-base was quite a time-consuming task since it had to process more that 4 million tweets. That is why sentiment intensity analyzer was chosen, it requires much less time to produce sentiment for each sentence at the expense of accuracy of sentiment provided as it provides the same sentiment score for each word in every context.

To compute the overall sentiment score for a particular day, I utilized the average of the sentiment scores for each tweet collected on that day. This approach allowed for the aggregation of sentiment information across a large volume of tweets, providing a more comprehensive understanding of the sentiment trends related to the target companies over time. By taking the mean of the sentiment scores, the influence of individual outliers and noise in the dataset was effectively minimized.

The resulting daily sentiment scores can serve as a valuable metric for understanding the public's sentiment towards these companies on a daily basis.

*Collection of news articles:*

To complement the sentiment analysis of tweets, we also collected news headlines for each day from the start of 2015 to the end of 2019, resulting in 88334 headlines for all 5 companies. Different sources of data were used for this task like Google News, New York Times, etc. As tweets only capture a small portion of the overall public sentiment, incorporating news headlines allows us to capture a broader range of sentiment related to the companies of interest. To enhance the accuracy of sentiment analysis on news headlines, we utilized FinBERT, a pre-trained BERT model that is fine-tuned on financial domain-specific language. By fine-tuning financial data, FinBERT has shown to improve the accuracy of sentiment analysis on financial news articles.

We took the average of all the sentiment scores for the headlines that were published on that day to get the sentiment scores for each day. Using this technique, we were able to daily collect the general sentiment trend for each organization, which we then used in our study.

The combination of sentiment analysis on tweets and news headlines provided us with a comprehensive view of the public sentiment towards the companies of interest, allowing us to identify any emerging patterns or trends in sentiment.

*Collection of political news:*

Political news were collected in the same way as news headlines regarding the company. The reason why political news were incorporated is that stock prices are usually influenced by different large events that happen around the world and not only by news that are directly related to the company itself. For example, changes in government policies and regulations, election outcomes, and international relations can affect the stock market and individual stock prices. 19466 political news articles were collected. RoBERTa was used for sentiment analysis.

*Models:*

In our study, we utilized several machine learning models to predict the sentiment score of the collected data. We employed the Long Short-Term Memory (LSTM) model and the Gated Recurrent Unit (GRU) model. Besides, Deep Learning models AdaBoost was used as well.

The LSTM model was designed with three layers, each containing 128, 64, and 32 units respectively. We added a dropout and recurrent dropout of 0.2 to prevent overfitting. Additionally, the model includes two dense layers with 16 and 1 units respectively. The LSTM model was compiled with the Adam optimizer and the MSE loss function. Early stopping was also implemented with a patience of 25 epochs and restore the best weights.

For the GRU model, we implemented a similar structure to the LSTM model.
The second GRU model was designed with three layers, each containing 128, 64, and 32 units respectively. Dropout and recurrent dropout of 0.2 were also implemented to reduce overfitting. Moreover, this model includes two dense layers with 16 and 1 units respectively. Similar to the LSTM models, the second GRU model was compiled with the Adam optimizer and the MSE loss function. Early stopping was also implemented with a patience of 25 epochs and restore the best weights.

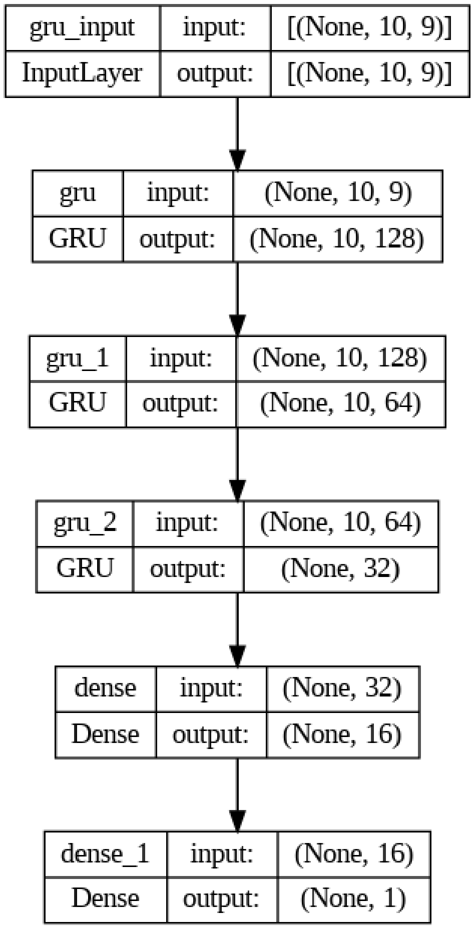


Figure 2. Architecture of GRU

In addition to the LSTM and GRU models, a AdaBoost model was also used for measuring the additional performance when incorporating features related to text. AdaBoost iteratively trains a series of weak classifiers on different subsets of the training data. Each weak classifier focuses on a particular subset and assigns weights to the data points, emphasizing the misclassified instances.

*Statistical test and metrics:*

However, it is important to notice that it would not be sufficient enough to just run every model one time, get the metric results and make conclusions from it. There is a model variation when training the model which means we may get different metric scores when training the model even on the same dataset. So, a different approach was used. We ran every model (every possible combination of features) for each company 10 times and used statistical tests to determine if a gain in model performance is significant or not. Basically, 10 metric values for each model were tested against 10 metric values of another model. By using such a method, our results are more robust to model variation and are more statistically correct. When assessing performance of different models three metrics were used: RMSE, MAE and Accuracy.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Figure 3. RMSE formula

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Figure 4. MAE formula

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Figure 5. Accuracy formula

Wilcoxon Signed Rank Test and Mann-Whitney U test were used to test if the difference in metrics were significant. When data does not adhere to the normality assumptions or when dealing with ordinal or non-normally distributed data, these two non-parametric statistical tests are used to compare two related or unrelated groups, respectively. Since we have the same test set, we would pay more attention to the results of Wilcoxon Signed Rank Test since we can not say that our groups are not related.

The null hypothesis: there is no significant difference in the RMSE (or other metric) scores between the model using only prices and the model incorporating sentiment of tweets.

Alternative hypothesis: There is a significant difference in the RMSE (or other metric) scores between the model using only prices and the model incorporating sentiment of tweets.

## Results

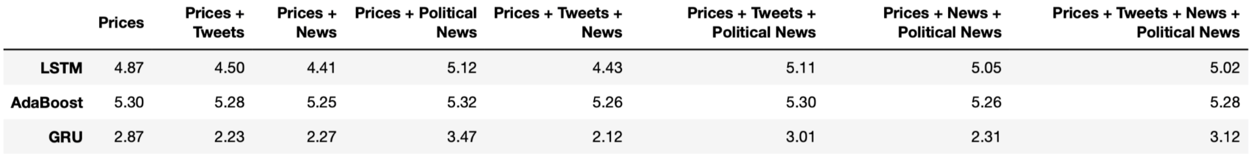
Several models were tested to see if the textual information brings any benefit to stock price prediction. Eight models were incorporated:

- Model with only historical prices
- Model with historical prices and tweets
- Model with historical prices and news (about a company)
- Model with historical prices and political news
- Model with historical prices, tweets and news (about a company)
- Model with historical prices, tweets and political news
- Model with historical prices, news (about a company) and political news
- Model with historical prices, tweets, news (about a company) and political news

It is important to test all possible combinations of models in order to determine the most effective approach to predicting the stock price using textual information. Each combination of models represents a different approach to incorporating sentiment from news and/or tweets in the prediction of stock prices. Testing each of these approaches allows for comparing the effectiveness of each model and aids in understanding which data source or combination of data sources is most useful in predicting stock prices. Testing all combinations of models allows for comparing the effectiveness of different sources of data against a base model that only uses historical prices. By comparing the results of these models against the base model, it becomes possible to measure the effectiveness of each source of textual information. This measurement is critical in determining whether the added information from news and tweets is helping in the prediction of stock prices or whether the historical price data alone is enough to predict stock prices accurately.

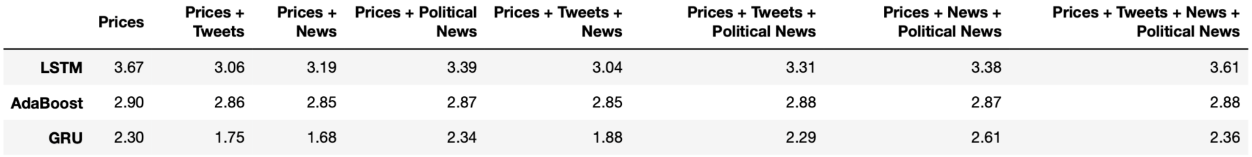
If the models that utilize textual data perform better than the base model, it indicates that adding sentiment analysis from news and/or tweets can provide additional insight into the stock market that is not captured by historical prices alone. This finding is significant because it supports the idea that incorporating additional data sources can provide better insights into the stock market's future performance. Additionally, identifying the most effective data type and combination of sources, allows for the development of more sophisticated and accurate models for predicting stock prices that can assist investors in making informed decisions.

Accuracy in our case is calculated based on how good the model predicts the direction of stock movement.



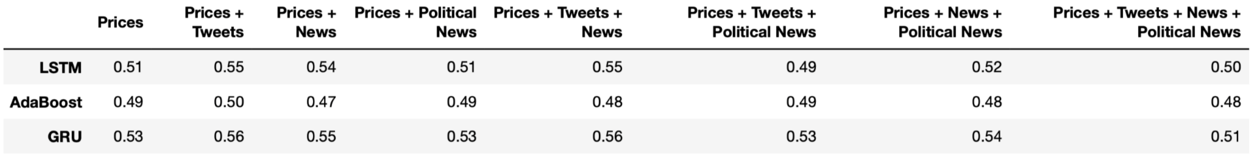
| | Prices | Prices + Tweets | Prices + News | Prices + Political News | Prices + Tweets + News | Prices + Tweets + Political News | Prices + News + Political News | Prices + Tweets + News + Political News |
|---|---|---|---|---|---|---|---|---|
| LSTM | 4.87 | 4.50 | 4.41 | 5.12 | 4.43 | 5.11 | 5.05 | 5.02 |
| AdaBoost | 5.30 | 5.28 | 5.25 | 5.32 | 5.26 | 5.30 | 5.26 | 5.28 |
| GRU | 2.87 | 2.23 | 2.27 | 3.47 | 2.12 | 3.01 | 2.31 | 3.12 |

Table 1. RMSE average scores



| | Prices | Prices + Tweets | Prices + News | Prices + Political News | Prices + Tweets + News | Prices + Tweets + Political News | Prices + News + Political News | Prices + Tweets + News + Political News |
|---|---|---|---|---|---|---|---|---|
| LSTM | 3.67 | 3.06 | 3.19 | 3.39 | 3.04 | 3.31 | 3.38 | 3.61 |
| AdaBoost | 2.90 | 2.86 | 2.85 | 2.87 | 2.85 | 2.88 | 2.87 | 2.88 |
| GRU | 2.30 | 1.75 | 1.68 | 2.34 | 1.88 | 2.29 | 2.61 | 2.36 |

Table 2. MAE average scores

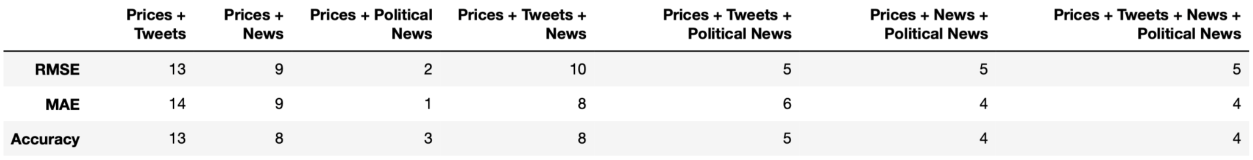


| | Prices | Prices + Tweets | Prices + News | Prices + Political News | Prices + Tweets + News | Prices + Tweets + Political News | Prices + News + Political News | Prices + Tweets + News + Political News |
|---|---|---|---|---|---|---|---|---|
| LSTM | 0.51 | 0.55 | 0.54 | 0.51 | 0.55 | 0.49 | 0.52 | 0.50 |
| AdaBoost | 0.49 | 0.50 | 0.47 | 0.49 | 0.48 | 0.49 | 0.48 | 0.48 |
| GRU | 0.53 | 0.56 | 0.55 | 0.53 | 0.56 | 0.53 | 0.54 | 0.51 |

Table 3. Accuracy average scores

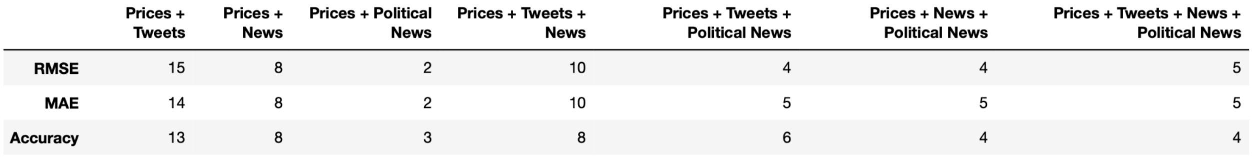
Every model was ran 10 times and the average values for Apple are present in the table above. The relationship of metrics between different features is quite similar for other companies as well. From tables for Apple we can see that there is a gain in performance when we use Tweets, News (about the company) or both features.

However, it is not just enough to look at average metrics, we must find out if these differences are statistically significant. Here are the results for statistical tests at 5% significance level.



| | Prices + Tweets | Prices + News | Prices + Political News | Prices + Tweets + News | Prices + Tweets + Political News | Prices + News + Political News | Prices + Tweets + News + Political News |
|---|---|---|---|---|---|---|---|
| RMSE | 13 | 9 | 2 | 10 | 5 | 5 | 5 |
| MAE | 14 | 9 | 1 | 8 | 6 | 4 | 4 |
| Accuracy | 13 | 8 | 3 | 8 | 5 | 4 | 4 |

Table 4. Wilcoxon Signed Rank Test results



| | Prices + Tweets | Prices + News | Prices + Political News | Prices + Tweets + News | Prices + Tweets + Political News | Prices + News + Political News | Prices + Tweets + News + Political News |
|---|---|---|---|---|---|---|---|
| RMSE | 15 | 8 | 2 | 10 | 4 | 4 | 5 |
| MAE | 14 | 8 | 2 | 10 | 5 | 5 | 5 |
| Accuracy | 13 | 8 | 3 | 8 | 6 | 4 | 4 |

Table 5. Mann-Whitney U test results

The number which is shown in every cell indicates how many times a particular feature gave a significant improvement in the corresponding metric out of the 15 tests (we have 15 tests because we have 5 companies and 3 metrics, 3 * 5 = 15). Significance results are shown considering results from all models. So, if the number is close to 15, it means that adding that particular feature always (or very often) gives improvement in model results. On the contrary, if the number is close to 0, it is almost never significant.

So, we can see that tweets show significant improvement almost all the time. While News and Tweets + News have moderate success in significance of improvement of performance. It has to be mentioned that political news do not seem to have a large effect on improvement of stock price prediction. Models that incorporate political news usually have worse values of metrics which could be explained by the assumption that features with political news are simply perceived as noise.

So, it appears that adding tweet sentiment can improve the performance of the model which indicates that it is an important feature to consider when building a stock price prediction model, while adding news sentiment may or may not lead to improved accuracy depending on the model and company. It appears that it makes more sense to incorporate only tweets without news since such a model gives the highest performance scores. However, if one tries to replicate the same models with the same features but for a different period of time and finds it unavailable to get data from Twitter, he/she may use news headlines as an additional feature which may also be useful.

## Conclusion

Our study offers valuable insights into the incorporation of sentiment analysis in predicting stock prices using textual data from tweets and news headlines. By analyzing a vast dataset comprising millions of tweets and news headlines related to major technology companies, we were able to uncover the impact of sentiment on stock price prediction performance.

In this study, we conducted sentiment analysis on over 4 million tweets, 88,334 news headlines related to five major technology companies (Apple, Microsoft, Tesla, Amazon, and Google) from 2015 to 2020 and 19466 political news. Our analysis revealed the overall sentiment of analysis of tweets can bring the most benefit when trying to predict the stock price and enhance performance of the models.

We also found that incorporating news headlines alone or along with tweets significantly outperformed a baseline model with only prices several times, allowing us to capture a broader range of sentiment related to the companies of interest. FinBERT, proved to be an effective tool for sentiment analysis on financial news articles. However, it is worth noting that our analysis indicated that incorporating sentiment from political news did not yield a reasonable improvement in stock price prediction performance. This suggests that the sentiment expressed in political news might not be as influential or directly relevant to the stock market dynamics compared to other factors. As a result, the allocation of resources to collect and analyze political news sentiment data may not be justified in the context of predicting stock prices.

To sum up, our study highlights the importance of sentiment analysis in understanding the public's perception of companies. By analyzing sentiment trends over time and incorporating them in machine learning, investors can better predict the stock price and possibly gain more revenue from trading on the financial market.

## Bibliography

- Singh, A., & Kaur, G. (2016). Stock market prediction using sentiment analysis. International Journal of Computer Applications
- Mittal, A., & Goel, A. (2012). Stock prediction using Twitter sentiment analysis. Stanford University CS229 Project Report
- Sengupta, S. (2018). Stock price prediction using sentiment analysis of Twitter data. Journal of Big Data
- Cristescu, A., Peca, A., & Robu, V. (2016). News sentiment analysis for stock price prediction using support vector machines. Procedia Computer Science
- Ko, H. J., & Chang, W. C. (2019). Sentiment analysis of news articles for stock prediction using an LSTM-based model. Information Processing & Management
- Jang, Y. J., Lee, J. W., & Lee, J. H. (2020). Sentiment analysis-based stock price prediction using BERT. Journal of Information Science
- Darapaneni, S., Ch, V. K. S. K., & P, S. S. S. (2021). Predicting Indian stock market prices using sentiment analysis and deep learning techniques. Journal of Financial Data Science
- Devi, K. S., & Nagabhushanam, P. (2018). Stock market prediction using sentiment analysis: A review. International Journal of Engineering & Technology
- Ding, X., Zhang, Y., Liu, T., & Duan, S. (2014). Using structured events to predict stock price movement: An empirical investigation. Journal of Systems Science and Complexity
- Xu, W., Huang, Z., & Zhang, H. (2018). Stock price prediction based on LSTM neural network and attention mechanism. Expert Systems with Applications
- Wu, C. Y., Tsai, C. F., & Wang, J. Y. (2020). Predicting the Taiwan stock market using news articles and sentiment analysis. Journal of Systems Science and Information
- Kim, M., Choi, S., & Lee, S. (2019). Stock price prediction using sentiment analysis and deep neural network. Applied Sciences

- Li, Y., Zhang, Z., & Li, Y. (2018). A hybrid approach for stock price prediction via combining ensemble learning with fuzzy rough set. Expert Systems with Applications
- Chen, Y., Chen, J., & Huang, C. J. (2015). An integrated framework for stock price forecasting based on PSO optimized fuzzy neural network with technical analysis. Neurocomputing
- Wang, J., Jiang, H., & Wei, Y. (2020). An event-driven approach to stock price prediction using deep learning. Journal of Intelligent & Fuzzy Systems