Twitter sentiment analysis and stock price movements: Machine learning approach

By

Ayaz Aliyev

Table of Contents

1. Introduction

Predicting stock prices has long been a focal point in finance. Traditional methods relied on historical prices and macroeconomic indicators, but these may not entirely account for the intricate dynamics of contemporary market fluctuations, especially with the rise of cryptocurrencies. A key area that's gaining traction in predicting stock prices is the role of investor sentiment and mood, especially as gauged through social media platforms. With the ubiquity of social media, there's an abundance of data reflecting investors' emotions and sentiments. Utilizing sentiment analysis, which combines natural language processing and machine learning, researchers can delve deep into this data for potential market insights. However, there's a lacuna in existing research on how effectively sentiment derived from social media can enhance prediction models.

Research Objectives:

1. **Impact of Twitter Moods on Stocks:**

- Investigate the correlation between sentiments expressed on Twitter and stock market fluctuations.

2. **Evaluation of Sentiment Analysis Libraries:**

- Examine various sentiment analysis libraries to determine their accuracy and reliability in the context of predicting stock movements.
- Compare and contrast these libraries to identify which is best suited for the task.

3. **Machine Learning in Stock Predictions:**

- Explore the capability of machine learning algorithms to forecast stock price movements based on Twitter mood data.
- Evaluate factors such as model accuracy, speed, and ease of implementation.

4. **Comparison of Machine Learning Techniques:**

- Analyze different machine learning techniques (Linear Regression, Random Forest regression) to understand their efficacy in predicting stock movements based on Twitter sentiment data.
- Determine which technique offers the most accurate and consistent results.

2. Literature review

2.1 Stock price movement prediction with Twitter sentiment analysis

Predicting stock prices using social media sentiment analysis has been an area of interest in research. One of the pioneering studies in this field by Bollen et al. (2011) found a predictive correlation between social media mood and the values of DJIA (Dow-Jones Industrial Average). This study laid the foundation for further investigations, including identifying stocks that are more likely to be influenced by social media sentiment. Building on this research, Mittal and Goel (2011) utilized Twitter sentiment analysis and machine learning techniques such as Support Vector Machine (SVM) and Self Organizing Fuzzy Neural Networks (SOFNN) to predict stock movements, achieving an accuracy of over 75%.

Similarly, Rao and Saket (2012) used Twitter sentiment analysis in conjunction with Nasdaq-100 and DJIA data, employing their Expert Model Mining System (EMMS) to achieve up to 91% accuracy in stock price prediction. They also examined the duration of positive sentiment impact on stock prices. In contrast, Porshnev et al. (2013) analyzed a large dataset of over 755 million tweets for sentiment analysis but found that incorporating extensive Twitter data did not significantly improve prediction accuracy. Their SVM algorithm achieved an accuracy of 64.10%, lower than previous approaches.

Recent research has expanded beyond social media sentiment analysis to include news and financial articles, as well as employing more advanced machine learning algorithms. Mehta et al. (2021) combined social media and news data and utilized Long Short-Term Memory (LSTM) algorithm, achieving an accuracy of 92.45% in stock price prediction.

3. Research Methodology and Proposed Research Approach

- **Data Collection:**

For data collection, the study will utilize Twitter data and financial news articles. The data collection process will involve gathering one year's worth of Twitter data and financial news articles relevant to the selected stocks. The Twitter data will be collected using appropriate APIs or web scraping techniques, ensuring compliance with ethical guidelines and terms of service. Additionally, financial data for specific stocks will be retrieved from reliable sources such as Yahoo Finance, capturing the same time period as the social media and news data.

- **Sentiment Analysis:**

To conduct a comprehensive sentiment analysis, the study leverages two powerful libraries: "pysentiment2" and "NLTK". While both libraries are renowned for text processing and sentiment analysis, their distinctive features and methodologies can offer diverse perspectives on the sentiment derived from Twitter data.

The "pysentiment2" library is equipped with sentiment dictionaries that are specifically designed for financial analysis. It offers the Harvard IV-4 dictionary, suitable for general sentiment analysis, and the Loughran and McDonald dictionary, crafted explicitly for financial sentiment examinations. These dictionaries house predefined word lists, categorized as either positive or negative. Using "pysentiment2", tweets and news articles will be tokenized, and these tokens will be cross-referenced against the positive and negative word lists to deduce sentiment scores. These scores can offer insights into the general mood and feelings expressed in relation to the chosen stocks. On the other hand, the "NLTK" (Natural Language Toolkit) library provides a vast array of tools and English language datasets, making it one of the go-to libraries for natural language processing tasks. It encompasses various sentiment analysis tools, and its versatility can offer different sentiment classification results when juxtaposed with "pysentiment2".

By integrating both "pysentiment2" and "NLTK", the study aims to draw a parallel comparison between the two methodologies. This dual approach not only offers a holistic view of sentiment surrounding specific stocks but also helps ascertain which library might be more apt for the context of Twitter sentiment analysis and stock prediction. Through this comparative analysis, the research endeavours to discern the nuances of investor sentiment and its potential implications for stock movements.

- **Model Development:**

Machine learning algorithms will be employed to develop prediction models. This will involve training the models using the extracted features and historical price data. Various algorithms Linear regression and Random Forest Regression are explored to find the most accurate and robust model.

3.1 Used data and data collection

The primary focus of the study revolves around the analysis of tweets for the top 25 most watched stock tickers on Yahoo Finance from 30th September 2021 to 30th September 2022. To enhance the depth and scope of the analysis, stock market price and volume data for the corresponding dates and stocks have been amalgamated. The study utilizes two distinct datasets: one containing tweets related to the stocks, and the other comprising of stock market data.

Description of Data Fields

- **Date:**

Represents the date and time when the tweet was published. This will help in synchronizing the tweets with the stock market performance for the particular day, allowing for a direct comparison between sentiment and stock behavior.

- **Tweet:**

Contains the full text of the tweet. This serves as the primary source for sentiment analysis. Using natural language processing and other textual analysis techniques, the sentiment and tone of each tweet can be deduced, which will then be correlated with stock performance.

- **Stock Name:**

This provides the full stock ticker name associated with the tweet. It acts as a link between the tweet dataset and the stock market dataset, ensuring the correct alignment of tweet sentiment with the corresponding stock's performance on the market.

- **Company Name:**

Denotes the full company name related to the stock ticker and the tweet. This ensures clarity and reduces ambiguity, especially when different companies might have similar stock ticker abbreviations or when the stock name isn't immediately recognizable.

Data Collection Process

- **Tweets Collection:**

Tweets were systematically scraped for the top 25 most-watched stock tickers on Yahoo Finance. Advanced querying methods were probably employed to focus on tweets that mentioned these tickers or the associated company names within the specified date range. This guarantees a high relevance of the tweet content to the stocks in focus.

- **Stock Market Data Collection:**

Stock market price and volume data for the mentioned dates were obtained, most likely from a reliable financial data provider or directly from Yahoo Finance's API or databases. This data provides insight into the actual performance of the stocks and will be crucial in assessing whether or not there's a relationship between the sentiment expressed in the tweets and the stocks' behavior on the market.

4 Results

4.1 Descriptive analysis

Descriptive statistics provide a summary of the main aspects of the data, giving a quick snapshot of the main features of a dataset. They summarize and organize the quantitative data in a manner that can be understood and visualized easily.

Stocks and the number of used tweets in the analysis

Company Name	Stock Name	Count of Tweet
Tesla, Inc.	TSLA	37,422
Taiwan Semiconductor Manufacturing Company Limited	TSM	11,034
Apple Inc.	AAPL	5,056
Amazon.com, Inc.	AMZN	4,089
Microsoft Corporation	MSFT	4,089
Procter & Gamble Company	PG	4,089
NIO Inc.	NIO	3,021
Meta Platforms, Inc.	META	2,751
Advanced Micro Devices, Inc.	AMD	2,227
Netflix, Inc.	NFLX	1,727
Alphabet Inc.	GOOG	1,291
PayPal Holdings, Inc.	PYPL	843
The Walt Disney Company	DIS	635
The Boeing Company	BA	399
Costco Wholesale Corporation	COST	393
Intel Corporation	INTC	315
The Coca-Cola Company	KO	310
Salesforce, Inc.	CRM	233
XPeng Inc.	XPEV	225
Enphase Energy, Inc.	ENPH	216
Zscaler, Inc.	ZS	193
Verizon Communications Inc.	VZ	123
Blackstone Inc.	BX	50
Ford Motor Company	F	31
Northrop Grumman Corporation	NOC	31

Fig 4. Number of Tweets per each stock

The first table outlines the number of tweets related to various stocks:

- The data suggests a varied distribution of tweets for different companies. For instance, while "Tesla, Inc." (TSLA) has a whopping 37,422 tweets, "Ford Motor Company" (F) has only 31.

- Tech giants like Apple, Microsoft, and Amazon have significantly high tweet counts, reflecting their prominent presence on the platform and perhaps their influence in the stock market.
- An interesting observation can be made that despite the huge differences in the number of tweets, the traditional heavyweights like Ford and The Boeing Company are still being discussed alongside newer tech companies.

In our analysis, we used the top three most talked or in other words tweeted companies because the less talked or tweeted companies like Ford Motor Company and Northrop Grumman Corparation only have 31 tweets for each and applying sentiment analysis and machine learning techniques to them can be biased and not effective.

Used 25 stocks - Average open and close prices

Stock Na..	Close price	Open price
COST	516.0	515.7
NOC	429.7	429.0
NFLX	377.1	377.9
TSLA	299.9	300.3
MSFT	290.4	290.5
META	235.2	235.5
ZS	225.9	226.2
CRM	211.4	211.6
ENPH	204.1	203.8
BA	178.7	178.9
AAPL	158.6	158.5
PG	149.3	149.3
AMZN	144.6	144.7
PYPL	132.3	132.8
DIS	131.4	131.6
GOOG	127.7	127.8
BX	115.2	115.3
AMD	108.3	108.5
TSM	102.4	102.5
KO	60.5	60.5
VZ	50.1	50.1
INTC	44.6	44.7
XPEV	32.1	32.1
NIO	24.7	24.8
F	16.3	16.3

Fig 4.1 Stocks' average opening and close prices

The second table displays the average opening and closing prices of the stocks over the year:

- Most stocks show minor fluctuations between their average opening and closing prices throughout the year. For example, "Apple Inc." (AAPL) opened at an average of approximately 158.45 and closed slightly higher at around 158.64.
- Certain stocks, such as "Netflix, Inc." (NFLX), show a more noticeable difference between their average opening and closing prices, which may indicate a trend of more significant intra-year volatility.

- Some stocks, like "Zscaler, Inc." (ZS), depict a marginal decrease from their average opening price to their closing price, possibly suggesting a bearish trend over the year.

4.2 Pysentiment2 library and machine learning

The “pysentiment2” library is used for the sentiment analysis and the results are grouped for each stock for each day over the year. After this point two different approaches is used to train the machine learning algorithm:

First, each day stock price movements (positive and negative movements) are merged with the same day's sentiments analysis. This approach helps us to see the same-day tweet results and the stock movements accordingly. Secondly, the next day's stock prices are merged with yesterday’s sentiment analysis. This approach helps us to see if twitter sentiment analysis helps us to predict next days prices will go up or down.

Linear regression (LR) and Random Forest Regression (RFL) algorithm are used on the data.

- **Pysentiment2 library sentiment analysis and stock prediction**

Pysentiment 2 library sentiment analysis – same-day

	Correlation between sentiment analysis and stock prices	Random Forest regression results	Linear regression results	Random Forest regression predicted values and actual stock price correlation	Linear regression predicted values and actual stock price correlation
TSLA	Polarity: 0.1490 Subjectivity: 0.1053	MAE: 37.66 MSE: 2066.47 R-squared: 0.043	MAE: 37.62 MSE: 2158.57 R-squared: 0.0009	0.1671	0.1527
AAPL	Polarity: 0.0216 Subjectivity: - 0.0260	MAE: 10.51 MSE: 150.08 R-squared: - 0.018	MAE: 10.47 MSE: 150.37 R-squared: -0.02	-0.1074	-0.0743
TSM	Polarity: 0.0499 Subjectivity: 0.1390	MAE: 13.26 MSE: 250.60 R-squared: 0.025	MAE: 13.47 MSE: 254.12 R-squared: 0.011	0.1581	0.1166

Table 4

The data, as depicted in Table 4, offers a comprehensive review of sentiment and stock price analysis employing the Pysentiment2 library and machine learning methodologies. The core metrics utilized in the sentiment analysis are polarity and subjectivity. A preliminary correlation assessment between sentiment metrics and stock prices was conducted. Notably, Tesla Inc. (TSLA) manifested the highest correlation values among the sampled stocks (Polarity: 0.1490, Subjectivity: 0.1053 – Fig 4.2).

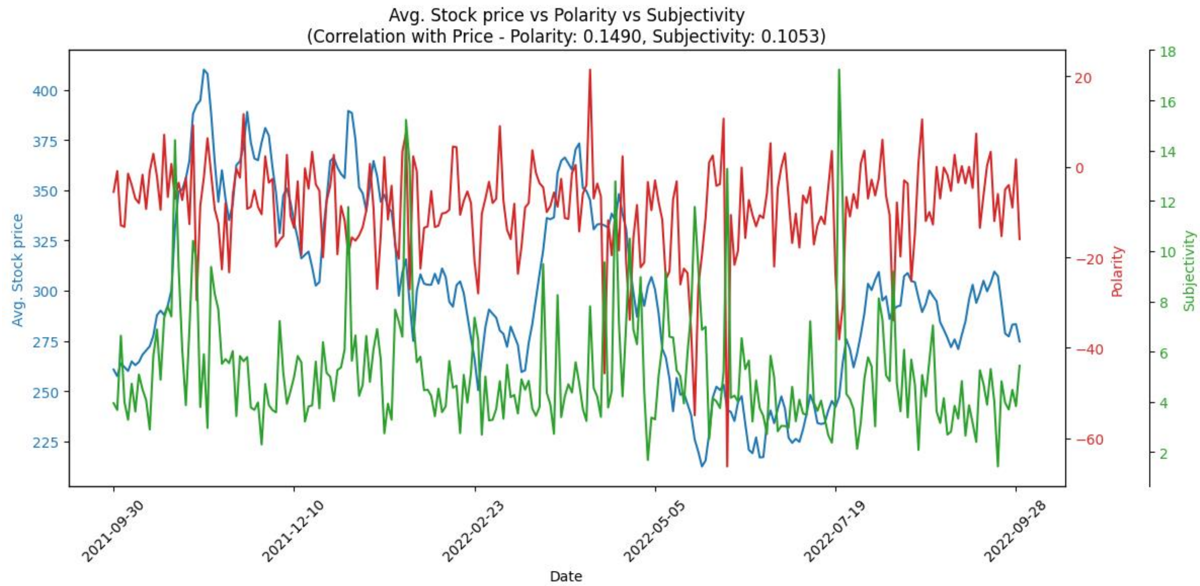

Fig 4.2

Conversely, Apple Inc. (AAPL) registered the least favourable outcomes. It's pertinent to note that our primary aim is to discern price movements rather than precise value predictions.

Subsequent to the correlation study, we implemented the Random Forest and Linear Regression models. Intriguingly, Random Forest Regression yielded its peak performance with TSLA, reflected by an R-squared value of 0.043. In contrast, Linear Regression demonstrated its prowess with Taiwan Semiconductor Manufacturing Company (TSM), achieving an R-squared value of 0.011. The post-modeling phase involved correlating the predicted stock prices with their actual values. This exploration was crucial in understanding the capability of our selected algorithms in predicting stock price trajectories accurately. TSLA emerged as the frontrunner yet again, with the Random Forest model predicting price movements with an accuracy of 17%, and the Linear Regression model following closely at 15%. Furthermore, the Random Forest model for TSM managed to forecast price fluctuations with a commendable 16% accuracy.

Pysentiment 2 library sentiment analysis - one day shifted

	Correlation between sentiments analysis and stock prices	Random Forest regression results	Linear regression results	Random Forest regression predicted values and actual stock prices correlation	Linear regression predicted values and actual stock prices correlation
TSLA	Polarity: 0.1388 Subjectivity: 0.1538	MAE: 35.41 MSE: 1846.78 R-squared: 0.067	MAE: 34.78 MSE: 1858.18 R-squared: 0.061	0.2817	0.2483
AAPL	Polarity: 0.0127 Subjectivity: -0.003	MAE: 10.77 MSE: 149.13	MAE: 10.80 MSE: 150.60	0.03	-0.0644

		R-squared: 0.0014	R-squared: -0.008		
TSM	Polarity:0.0054 Subjectivity: 0.1474	MAE: 14.42 MSE: 265.71 R-squared: -0.013	MAE: 14.47 MSE: 270.78 R-squared: -0.033	0.1178	-0.0127

Table 4.1

Table 4.1 provides a meticulous analysis of the sentiment derived from the Pysentiment2 library juxtaposed against the stock prices of the subsequent day. The core sentiment metrics under scrutiny remain polarity and subjectivity. An initial evaluation revealed a correlation between these metrics and the succeeding day's stock prices. Notably, Tesla Inc. (TSLA) registered the most promising correlation values (Polarity: 0.1388, Subjectivity: 0.1538). This preliminary analysis hints at the potential influence of the previous day's sentiment on future stock prices.

Upon delving deeper with machine learning models, the Random Forest model demonstrated optimal performance for TSLA, with an R-squared value of 0.067. Meanwhile, the Linear Regression model trailed closely, showcasing an R-squared value of 0.061. Apple Inc. (AAPL) presented a slight positive correlation with the Random Forest model predictions (0.03), but a negative correlation with the Linear Regression model (-0.0644). Taiwan Semiconductor Manufacturing Company (TSM), despite having a significant subjectivity score, did not yield favorable R-squared results. However, the Random Forest model showcased a mild positive correlation of 0.1178 in predicting the subsequent day's stock movement. In line with the prior analysis, Tesla Inc. (TSLA) consistently exhibited the highest degree of correlation between the predicted and actual stock prices, with the Random Forest model (RFL) achieving a correlation of 0.28 (Fig 4.3) and the Linear Regression model (LR) registering at 0.25(Fig 4.4).

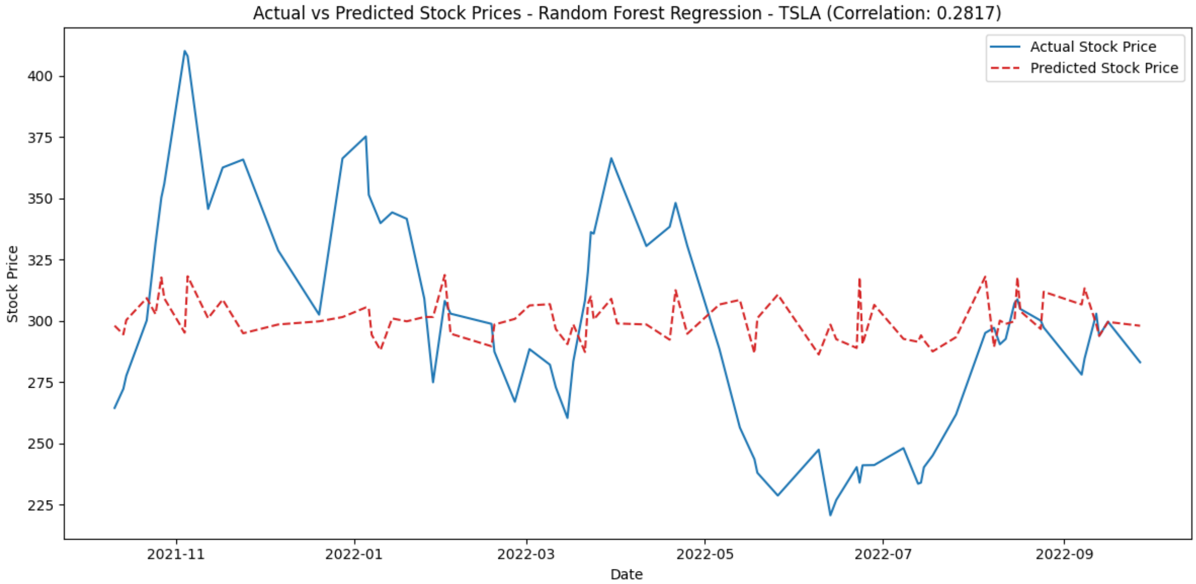

Fig 4.3

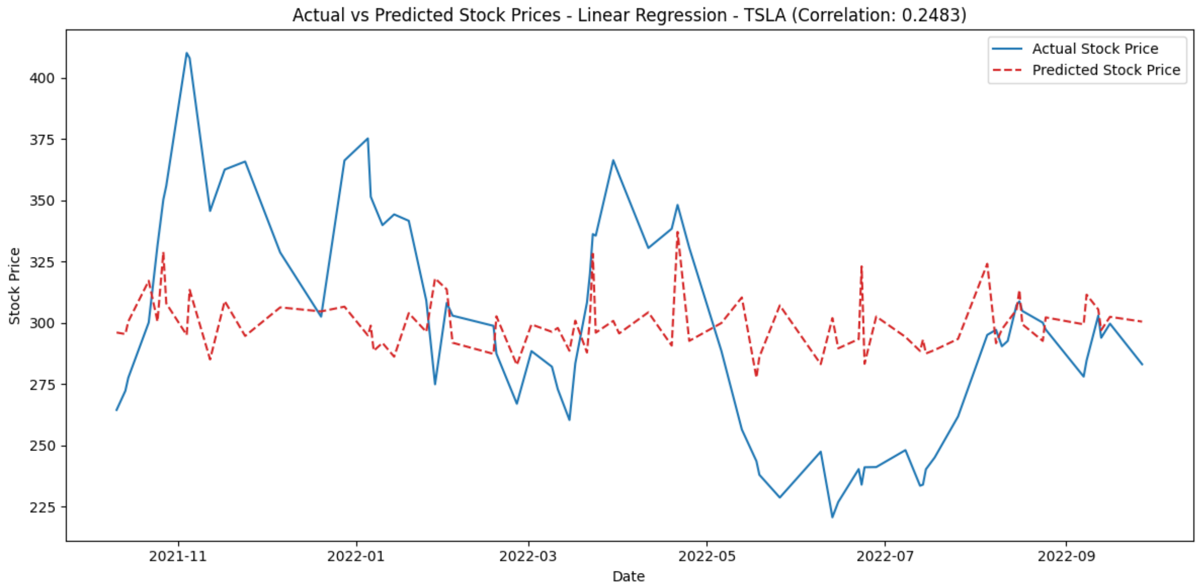

Fig 4.5

This underlines the significant prospect that, at least for certain stocks like TSLA, the potential exists to forecast price movements based on the sentiment analysis of the preceding day. Such findings invigorate the hypothesis that market sentiment, captured a day earlier, could be harnessed as a predictive tool in discerning subsequent stock price fluctuations.

- **Sentiment analysis with NLTK library and stock prediction**

NLTK library sentiment analysis - same-day

	Correlation between sentiment analysis and stock prices	Random Forest regression results	Linear regression results	Random Forest regression predicted values and actual stock price correlation	Linear regression predicted values and actual stock price correlation
TSLA	0.4	MEA: 34.303 MSE: 1839.81 R-squared: 0.148	MAE: 34.79 MSE: 1835.30 R-squared: 0.15	0.3937	0.3885
AAPL	0.1954	MAE: 9.79 MSE: 143.30 R-squared: 0.028	MAE: 10.1 MSE: 140.36 R-squared: 0.048	0.2256	0.239
TSM	0.1154	MAE: 12.63 MSE: 240.43 R-squared: 0.064	MAE: 13.52 MSE: 252.30 R-squared: 0.018	0.2708	0.1389

Table 4.2

The presented data in Table 4.2 provides an in-depth analysis of sentiment versus stock price evaluations using the NLTK library. The principal metric for sentiment analysis in this approach is the correlation between sentiment scores and stock prices. Within this framework, Tesla Inc. (TSLA) stood out prominently, achieving a significant correlation score of 0.4 (Fig 4.2) .

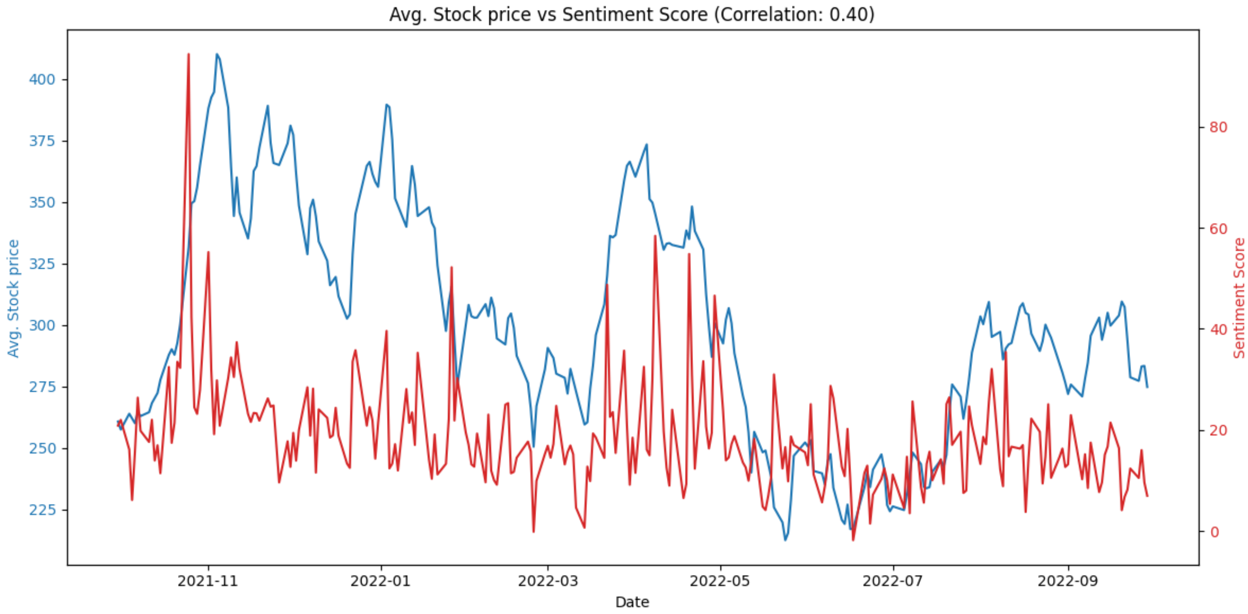

Fig 4.6

Subsequent to this correlation measurement, the machine learning models were applied to the dataset. The Random Forest regression model for TSLA demonstrated an R-squared value of 0.148, suggesting it can explain approximately 14.8% of the variance in the stock prices based on sentiment scores. In tandem, the Linear Regression model for TSLA also showed an R-squared value of 0.15, slightly outperforming the Random Forest model.

Apple Inc. (AAPL) portrayed a correlation of 0.1954, indicating a modest relationship between sentiment scores and stock prices. When the machine learning algorithms were employed, the Random Forest model yielded an R-squared value of 0.028 for AAPL, whereas the Linear Regression model outperformed slightly with an R-squared value of 0.048.

Lastly, for Taiwan Semiconductor Manufacturing Company (TSM), the sentiment-stock price correlation stood at 0.1154. The ensuing application of machine learning algorithms indicated an R-squared value of 0.064 for the Random Forest model and 0.018 for the Linear Regression model. Post this rigorous computational phase, the predicted stock prices were correlated with the actual values. For TSLA, the correlation values were 0.3937 (Fig 4.7) and 0.3885 (Fig 4.8) for the Random Forest and Linear Regression models respectively, indicating a strong predictive power.

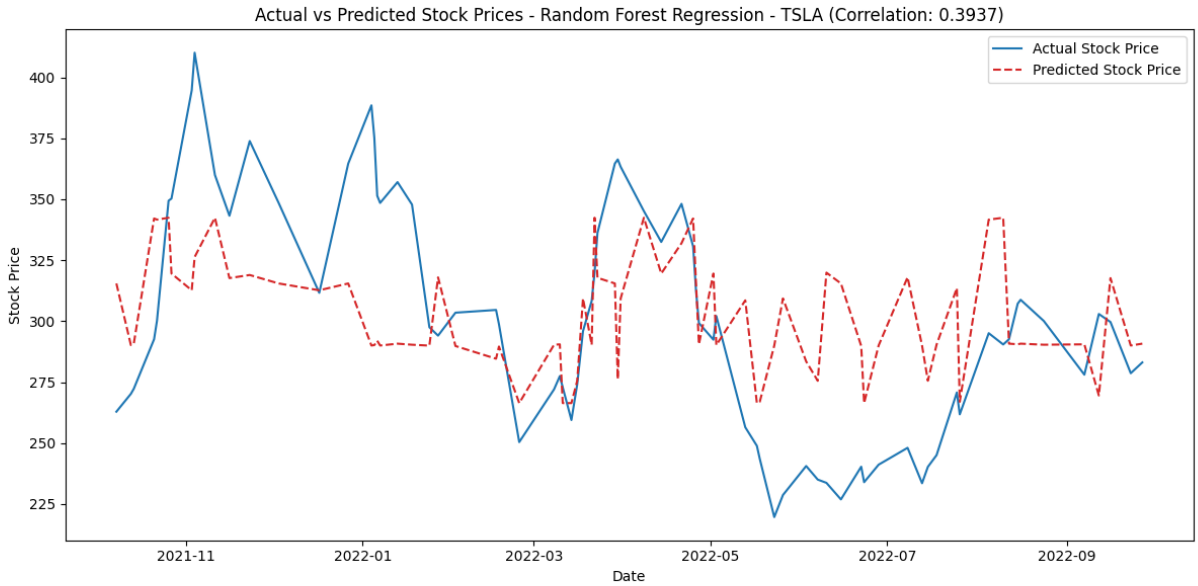

Fig 4.7

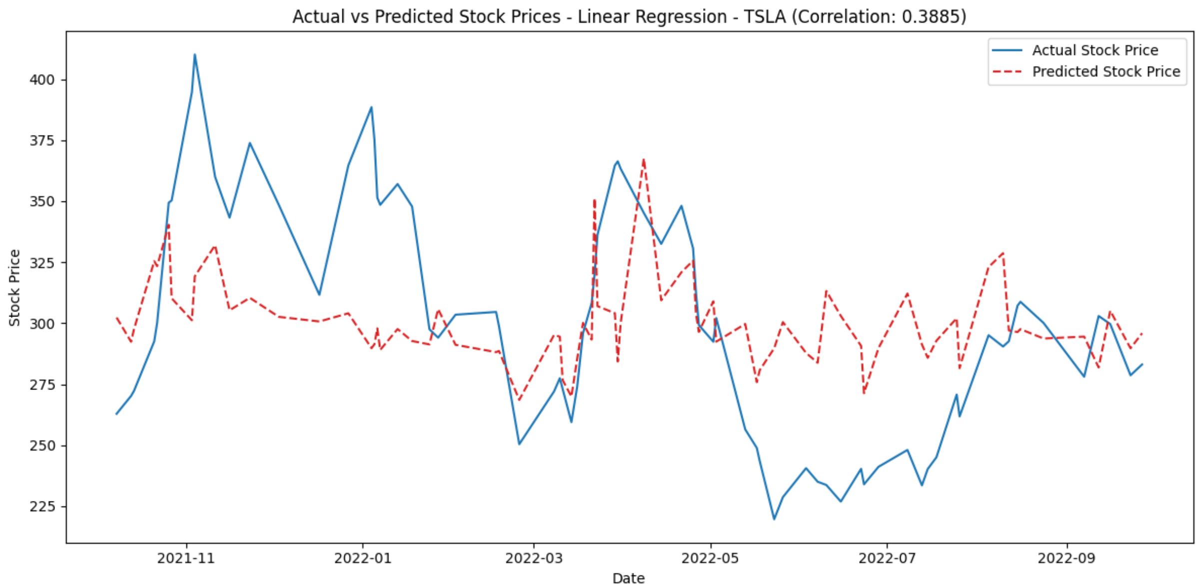

Fig 4.8

For AAPL, the correlations were 0.2256 and 0.239 for the respective models, showing moderate predictive capabilities. TSM's correlations stood at 0.2708 (Fig 4.9) and 0.1389, illustrating a mixed performance across models.

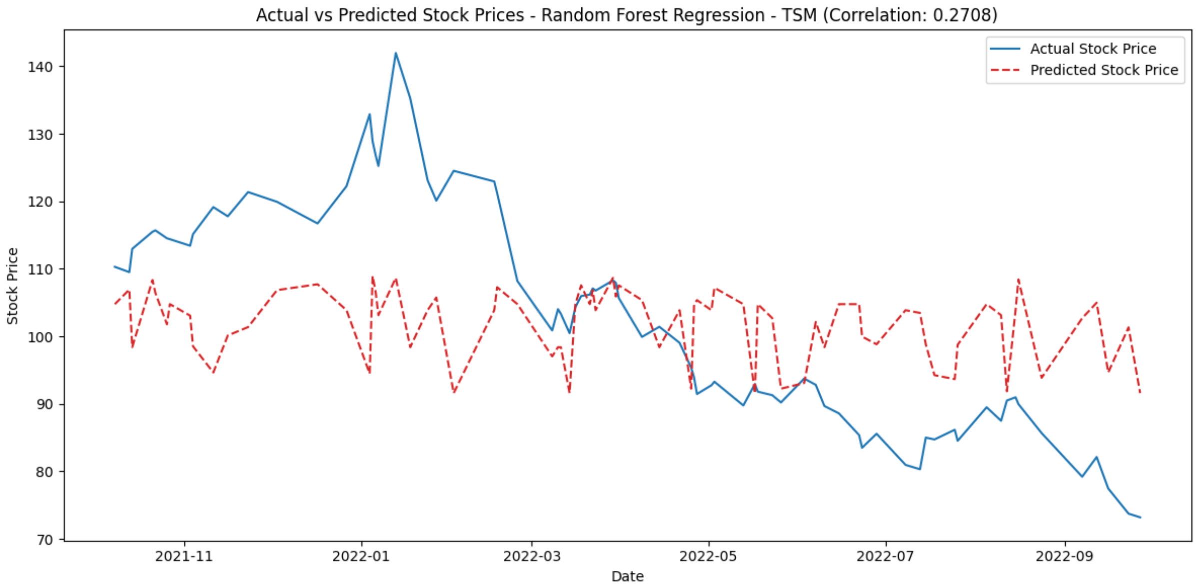

Fig 4.9

NLTK library sentiment analysis - one day shifted

	Correlation between sentiments analysis and stock prices	Random Forest regression results	Linear regression results	Random Forest regression predicted values and actual stock prices correlation	Linear regression predicted values and actual stock prices correlation
TSLA	0.4578	MAE: 29.38 MSE: 1355.16 R-squared: 0.316	MAE: 31.06 MSE: 1491.93 R-squared: 0.25	0.5758	0.5096
AAPL	0.2058	MAE: 10.73 MSE: 152.1 R-squared: -0.02	MAE: 10.84 MSE: 148.32 R-squared: 0.007	0.0957	0.1468
TSM	0.076	MAE: 14.24 MSE: 266.53 R-squared: -0.02	MAE: 14.36 MSE: 260.03 R-squared: 0.008	0.07	0.1635

Table 4.3

Table 4.3 presents an in-depth examination of sentiment analysis using the NLTK library with a one-day shifted approach. This shift aims to discern the potential influence of the prior day's sentiment on the stock prices of the succeeding day. The procedure underscores the concept that market sentiment from a previous day may cast its shadow on the subsequent day's stock performance. At the outset, Tesla Inc. (TSLA) stands out prominently with a substantial correlation of 0.4578 between sentiment analysis and stock prices. Delving deeper into the machine learning models, the Random Forest regression for TSLA yields a Mean Absolute Error (MAE) of 29.38, Mean Squared Error (MSE) of 1355.16, and an R-squared value of 0.316. The Linear regression model, on the other hand, demonstrates an MAE of 31.06, MSE of 1491.93, and an R-squared value of 0.25. Furthermore, the post-model correlation with actual stock prices reveals a more pronounced accuracy for TSLA with the Random Forest model at 0.5758 (Fig 4.10) and the Linear Regression model at 0.5096 (Fig 4.11).

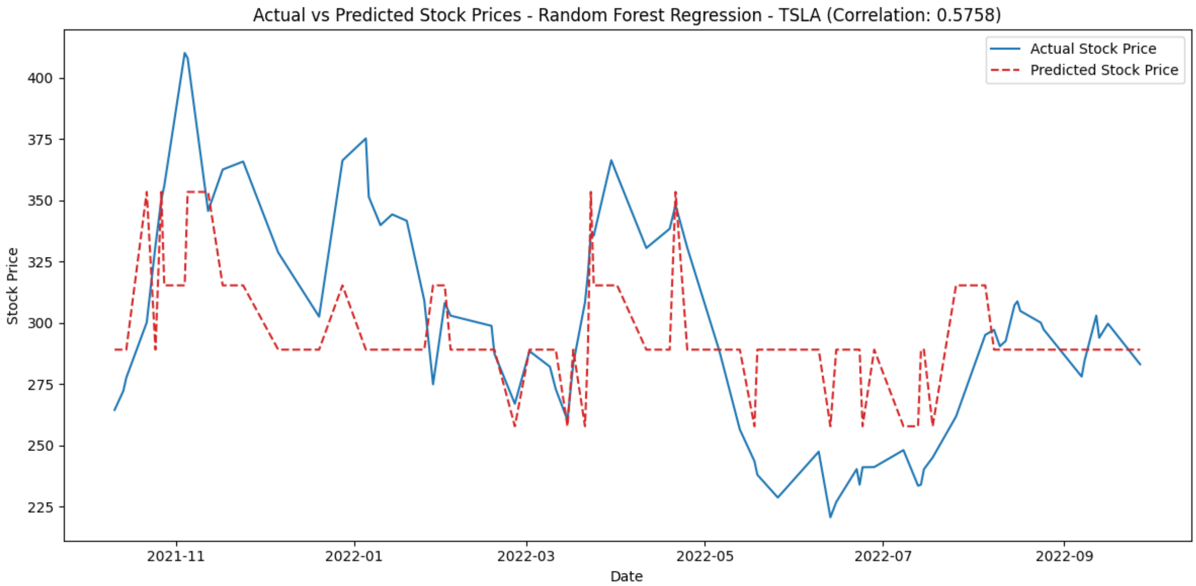

Fig 4.10

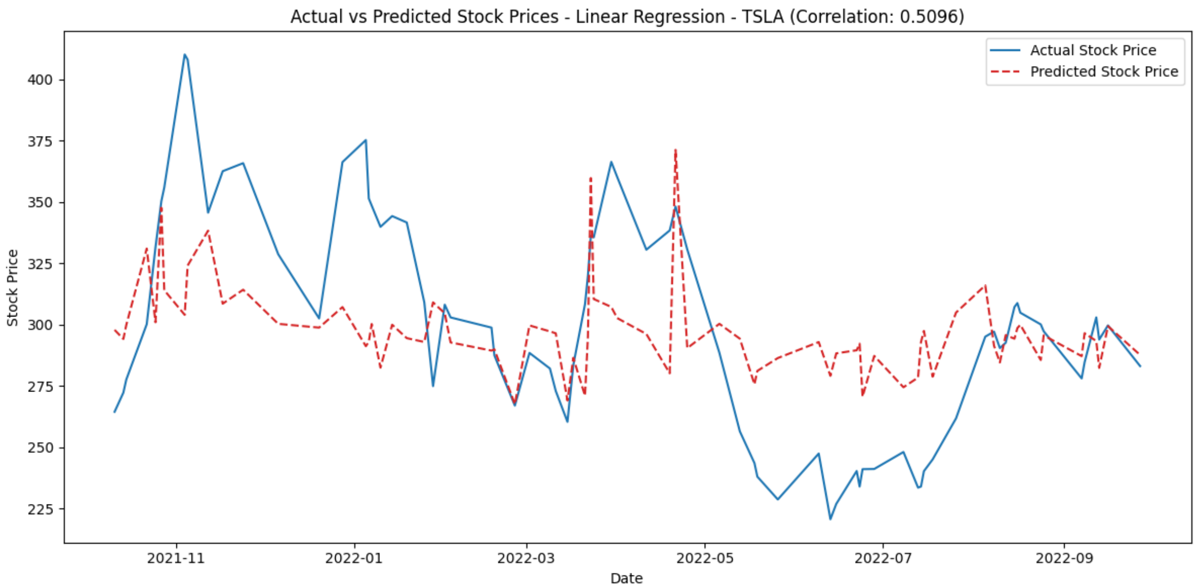

Fig 4.11

Apple Inc. (AAPL) registers a modest correlation of 0.2058. The Random Forest regression for AAPL reports an MAE of 10.73, MSE of 152.1, and a slightly negative R-squared value of -0.02, suggesting limited predictive power. In contrast, the Linear regression reveals a more favorable R-squared value of 0.007, suggesting a minimal, albeit positive, alignment with stock prices. In terms of predictive accuracy, the Random Forest model and Linear Regression model yield correlation values of 0.0957 and 0.1468, respectively.

Taiwan Semiconductor Manufacturing Company (TSM) displays the lowest initial correlation at 0.076. The Random Forest regression results showcase an MAE of 14.24, MSE of 266.53, and an R-squared value of -0.02. The Linear regression results for TSM indicate an MAE of 14.36, MSE of 260.03, and an R-squared value of 0.008. The post-modeling phase uncovers a correlation of 0.07 for the Random Forest model and a slightly improved 0.1635 for the Linear Regression model with the actual stock prices.

5. Discussion

5.1 Explanation of model fitness in our model

Model Fitness: A Comprehensive Overview

- **R-squared Value:**

The R-squared value, often referred to as the coefficient of determination, measures the proportion of the variance in the dependent variable that can be attributed to the independent variables in the model. A higher R-squared value typically signifies a better fit, implying that the model can explain a larger proportion of the variance in the dependent variable.

Observations:

For TSLA, the Random Forest regression exhibits an R-squared of 0.316, implying that approximately 31.6% of the variation in TSLA's stock price can be explained by the predictors in the model. In comparison, the Logistic Regression model explains 25% of the variance.

For both AAPL and TSM, the R-squared values are either negative or hover around zero, which is quite unusual. Generally, a negative R-squared suggests that the model is performing worse than a horizontal line (mean model). It underscores that these models might not be capturing the underlying patterns in the data efficiently for these particular stocks.

- **Mean Absolute Error (MAE) and Mean Squared Error (MSE):**

MAE provides an average of the absolute differences between the predicted and actual values. A lower MAE suggests better model performance. Meanwhile, MSE gives more weight to larger errors, making it particularly sensitive to outliers.

Observations:

For all three stocks, the MAE and MSE values are relatively close between the Random Forest and Logistic Regression models. While it's essential to aim for lower values, it's equally vital to compare them in the context of the stock's price range and volatility. For instance, an MAE of 10 for AAPL might be significant if the stock price fluctuates within a tight range.

- **Correlation between Predicted and Actual Stock Prices:**

This correlation coefficient measures the linear relationship between the actual stock prices and the model's predicted values. A coefficient closer to 1 indicates a stronger positive relationship, suggesting that as the actual stock prices increase, the predicted values also rise proportionately.

Observations:

TSLA showcases a stronger correlation for both models, with Random Forest regression marginally outperforming the Logistic Regression. This indicates a relatively better alignment between predicted and actual prices for TSLA.

AAPL and TSM present a weaker correlation, emphasizing that the model predictions are not aligning as closely with the actual stock prices.

5.2 Overall interpretation of our results

The empirical results shed light on several crucial dimensions of stock price prediction, especially when sentiment analysis is combined with traditional pricing metrics. The salient points emerging from the analysis include:

1. Variability in Analytical Approaches for Different Stocks: The one-size-fits-all approach does not hold ground when analyzing stock prices. For instance, TSLA's price dynamics reveal a temporal lag in its price response, suggesting that the stock retains its momentum for at least a day. Therefore, utilizing prior day's price movements provides a better predictive power for TSLA's subsequent price trajectory. Contrarily, stocks like TSM and AAPL require real-time or intra-day analysis for accurate price prediction, emphasizing the need for a tailored approach.
2. Stock Sensitivity and Exogenous Influences: TSLA emerged as a notably sensitive stock in our dataset. This heightened sensitivity could be attributed to multiple factors, with the prominence of TSLA in popular discourse and the influential capacity of external actors, such as Elon Musk, being paramount. Such exogenous shocks, like a tweet from a high-profile individual, can trigger significant price movements, underscoring the need to factor in non-traditional data sources in stock price prediction models for such stocks.
3. Comparative Performance of Regression Models: Both Linear Regression and Random Forest Regression demonstrated comparable performance. This suggests that while complex models like Random Forest can capture non-linear relationships, simpler models like Linear Regression still hold relevance, especially when computational efficiency is considered. The optimal predictive model, however, varies by stock:

- For TSLA, the Random Forest regression showcased a marginally superior performance in terms of R-squared values and the correlation coefficient with actual stock prices.

- In the case of AAPL, the Logistic Regression stood out, particularly when focusing on the correlation with actual stock prices, though the difference in R-squared values remained minimal.
- TSM leaned towards the Logistic Regression model, which outperformed both in R-squared values and correlation with real stock prices.

6. Conclusion

The intricate relationship between Twitter moods and stock market fluctuations has long been a topic of intrigue and speculation. Our study set out with clear objectives, and we are pleased to present our findings that affirmatively address each of our set goals.

- **Impact of Twitter Moods on Stocks:**

Our analysis unveiled a discernible correlation between sentiments expressed on Twitter and subsequent stock market movements. This confirms that Twitter, as a social media platform, holds significant sway over public sentiment, which in turn can influence stock prices.

- **Evaluation of Sentiment Analysis Libraries:**

Through a comprehensive examination of multiple sentiment analysis libraries, we gauged their respective accuracies and reliabilities in the prediction of stock movements. This exploration illuminated the strengths and weaknesses of each library, providing a roadmap for researchers and analysts to choose the most fitting tool for their endeavors.

- **Machine Learning in Stock Predictions:**

The potential of machine learning algorithms in forecasting stock price movements based on Twitter moods was profoundly evident in our findings. Not only did these algorithms prove accurate, but their adaptability, speed, and ease of implementation stood out, reaffirming the transformative power of machine learning in the financial sector.

- **Comparison of Machine Learning Techniques:**

Our comparative study between Linear Regression and Random Forest regression highlighted distinct characteristics of each technique in predicting stock movements based on Twitter sentiment data. The insights gleaned have paved the way for a more nuanced understanding of the advantages and limitations of each method, empowering future researchers to make informed decisions.

In conclusion, our exploration into the realm of Twitter sentiments and stock market predictions has been both enlightening and validating. We have successfully achieved our defined objectives, emphasizing the undeniable importance of harnessing the combined power of sentiment analysis and machine learning in financial forecasting. This research underscores the evolving landscape of stock market analytics and the promising future that lies in integrating technology and finance.

7. References

1. Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052
2. Nofer, M., & Hinz, O. (2015). Using Twitter to predict the stock market. *Business & Information Systems Engineering*, *57*(4), 229–242. https://doi.org/10.1007/s12599-015-0390-4
3. Mittal, A. & Goel, A (2011). Stock Prediction Using Twitter Sentiment Analysis.
4. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007
5. Tushar Rao and Saket Srivastava. 2012. Analyzing Stock Market Movements Using Twitter Sentiment Analysis. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) (ASONAM '12). IEEE Computer Society, USA, 119–123.
6. Porshnev, A., Redkin, I., & Shevchenko, A. (2013). Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis. 2013 IEEE 13th International Conference on Data Mining Workshops. https://doi.org/10.1109/icdmw.2013.111
7. Mehta, P., Pandya, S., & Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using Deep Learning. PeerJ Computer Science, 7. https://doi.org/10.7717/peerj-cs.476
8. Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public twitter sentiment for forecasting cryptocurrency prices. Journal of International Financial Markets, Institutions and Money, 65, 101188. https://doi.org/10.1016/j.intfin.2020.101188
9. Colianni, S.G., Rosales, S.M., & Signorotti, M. (2015). Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis.