Turk J Elec Eng & Comp Sci
() : –
© TÜBİTAK
doi:10.3906/elk-

Social Media based Stock Market Analysis using Big-Data Infrastructure

Sujaudeen NANNAI JOHN*, Vishakan SUBRAMANIAN, Venkataraman NAGARAJAN, Shashanka VENKATESH
Dept. of Computer Science and Engineering, SSN College of Engineering, Chennai, India

Received: • **Accepted/Published Online:** • **Final Version:**

Abstract: Several factors influence the value of a stock apart from the typical quantitative and qualitative parameters seen in the fundamental analysis of stocks like balance sheets, income statements, cash flow statements etc. In recent years, one such factor that has gained prominence is social media trends. In this paper, we aim to study the correlation between social media trends and stock market movement with the usage of a big-data architecture. This work considers the sentiments expressed by Twitter users on relevant topics, and measures their correlation to stocks of companies within the relevant sector the Tweet appeals to. For efficient processing of such large-scale data, we use Apache Kafka for data ingestion and Apache Spark for Tweet data processing, i.e. - sentiment extraction and aggregation. Based on observed correlations, our proposed model helps to predict stock market movement based on the extracted Tweet sentiment data using state of the art gradient boosted tree models - namely XGBoost and CatBoost.

Key words: Stock market analysis, Big-data, Social media, Sentiment analysis, Machine learning, Twitter

1. Introduction

Stock market analysis has been a topic of great interest ever since public stock exchanges came into existence, since it is a realm of great profit. The task is considerably demanding because of the highly dynamic nature of the stock market, which is a non-linear, noisy and chaotic system. [1]. Social media has transformed from a niche utility to a ubiquitous means of information exchange throughout the world. People have started utilizing social media as a tool not only for socialising, but also as a means to consume news, share their opinions, create awareness, follow the ideas of popular personalities and luminaries, among many others. The architecture of social media provides the avenue for interesting ideas to disseminate among society in near-instant times. In this age, such sudden outbursts of information about a company or its product can make or break the valuation of that product or company - since information travels fast and people tend to react towards sensational articles. To highlight the severity of social media trends towards the stock market, we could consider the example of the Tweet posted by Elon Musk, the CEO of Tesla Motors, on 27 Jan 2021 towards a declining video-game retail company, GameStop [2]. His Tweet about the company immediately raised the company's stock value by more than 60% within hours. Such is the influence of social media on an already volatile stock market. Hence, it is worthwhile to understand the relationship between social media trends and the stock market. In this paper, we analyse the statistical correlation between social media trends and its effect on the relevant industry's stock price valuations using a big-data architecture, justified by the voluminous nature of Twitter and stock data.

*Correspondence: nsujaudeen@gmail.com

2. Related Work

The research undertaken by Lee et. al. [3] delves into analyzing Twitter data, by classifying it into relevant industrial sector categories and performing sentiment analysis to predict the trend of stock prices, and comparing it to reality, to consequently find the correlation between the two. Their publication states that they used a sample of 1000 news articles to create a classifier. They were able to obtain a 77% accuracy on the classification of Tweets into different sectors. They have used 100 Tweets overall, with 20 Tweets per category in order to test out their hypothesis.

Bollen et al.'s well-known paper [4] indicate that the accuracy of DJIA predictions (Dow Jones Industrial Average) can be significantly improved by including specific mood dimensions. Their research work analyzes mood in terms of 6 dimensions: Calm, Alert, Sure, Vital, Kind & Happy.

Following the results of the above paper, another research study was performed by Mittal et al. [5] that uses Twitter to predict public mood, and use the predicted mood and previous days' DJIA values to predict stock market movements. They have proposed an algorithm to approximate stock market data on days in which the market was closed due to public holidays and weekends. For prediction of stock market movements, they highlight several algorithms like Regression, SVM and SOFNNs (Self-Organizing Fuzzy Neural Networks).

Nayak et al. [6] have done significant work on sentiment analysis of Twitter data for stock price prediction using trend analysis of stock data obtained from Yahoo Finance for three sectors: Banking, Mining & Oil and sentiment data extracted from relevantly collected Tweets. They also predict the above markets based on available historical stock data and extracted social media sentiment data using 3 different supervised machine learning models. Kalyanaraman et al. [7] have put forth a paper that performs sentiment analysis on news articles to analyse the causative relation between news and stock valuations. They used a custom sentiment dictionary to build a classifier of news articles using Linear Regression and Gradient Descent. Their model predicted the sentiment of articles with around 60% accuracy and the direction of stock price movement (positive or negative) with an accuracy of 81.92%. Peng uses a big-data architecture using Apache Spark [8] and Apache Hadoop to predict US Oil stock prices.

The Stanford Core NLP library has been used by Kanavos et al. [1] to perform text pre-processing along with Naive Bayes classifier to perform sentiment analysis (using a traditional n-grams approach) on Twitter data for stock price predictions. Pagolu et al. [9] performed correlation analysis between Tweets related to Microsoft and its stock price movements during the time period of August 2015 - August 2016. They transformed correlation analysis into a classification problem with the input features being the total negative, neutral and positive emotions in Tweets observed in a 3 day period. With the utilization of methodologies like Word2Vec for sentiment analysis and Logistic Regression & SVM to perform the classification task, they were able to come up with a machine learning model which yielded an accuracy of around 70% to predict the stock movement as an upward or downward trend.

Mehtab et al. [10] further built upon the work done by Mittal et al. and developed eight regression and eight classification models for predicting the stock price movement of NIFTY 50, which is augmented using public mood data which was analysed and aggregated from relevant Tweets.

Selvamuthu et al. [11] present several variations of ANNs (Artificial Neural Networks) with different learning algorithms to predict the Indian stock market using tick data. They suggest using a more extensive dataset to capture seasonal trends, and consider the avenue of sentiment analysis, since statements expressed by renowned personalities are known to affect the stock market, to gain an extra edge in stock market prediction.

3. Design of the Proposed Architecture

An overview of the architecture is shown in Figure 1.

Initially, Tweet and stock data are collected. They are then pre-processed before being streamed through Apache Kafka [12]. Following this, Apache Spark [13] is used to perform sentiment analysis on Tweets and aggregate the results date-wise. Finally, correlation analysis is performed on the aggregated Tweet data and stock data for various market sectors. Based on the observed correlation, machine learning models are then built to predict stock market movement based on observed Tweet sentiments.

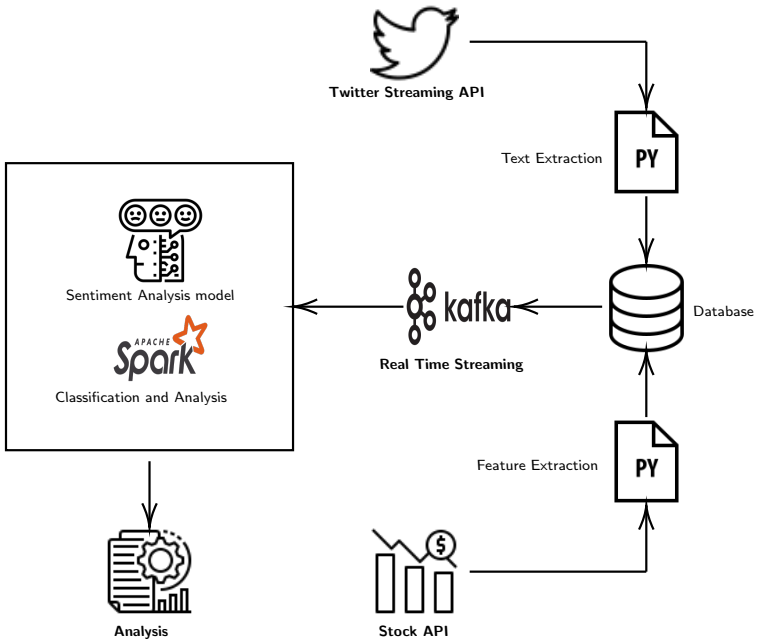


Figure 1: Architecture Design.

This architecture was built based upon the analysis of different tools available for various tasks involved in a big-data processing lifecycle, as researched by Gürcan et al. [14].

4. Data Collection and Preprocessing

4.1. Data Collection

The official Twitter API with the Academic Research access [15] is used to collect archival Tweet data. Polygon.io [16] is used to obtain 2 year historical stock market data. Tweets along with its retweet count are collected for certain market sectors based on appropriately chosen keyword(s) like #MSFT, #TSLA etc. and stock market data is collected for a representative company present in the corresponding industrial sector. The representative company of each sector was chosen based on it's market share in the sector and presence in the USA, since the majority of Twitter users are American (nearly 77M as of Jan 2022). Table 1 summarizes the volume of Tweet data collected and used in each sector.

4.2. Data Preprocessing

The collected JSON data from the Twitter API is cleaned to remove emoticons and extraneous characters. Links and other irrelevant data are removed from the Tweet, if any. Duplicate Tweets are also removed. Upon

Sector	Data Points
Electric Vehicles (EVs)	116,355
Oil	53,253
Technology (Tech)	632,181
Pharmaceuticals (Pharma)	313,407

Table 1: Data points count.

analysing the distribution of the collected Tweets' retweet counts, a high leftward skew is observed since most Tweets do not have high retweet counts. To filter out noise, Tweets with a very low retweet count are removed and Tweets with a high retweet count (outliers) are manually capped to a fixed retweet count to avoid bias and skew in the data.

Using the approximation algorithm devised by Mittal et al. [5], stock market data for missing days are calculated and imputed. Finally, all the preprocessed data is stored in a database for streaming through Apache Kafka.

5. Sentiment Analysis

A crucial component in the pursuit of our study to understand the relation between social-media (Tweets) and the stock market is analysing the sentiment (or mood) reflected by social media at any given date. To perform sentiment analysis on Twitter data, various NLP methods have been explored, as discussed in Section 2. In the field of Natural Language, one of the recent developments has been the introduction of the BERT (Bidirectional Encoder Representations from Transformers) Model [17]. BERT differs from conventional NLP language models in the fact that it uses bidirectional training and attention mechanism in order to have a deeper sense of language context compared to other unidirectional models. Since its introduction, the BERT model has been extensively used for popular NLP tasks like Text Summarization, Question-Answering, Text Classification, Sentiment Analysis etc.

In our methodology, we chose to use Cardiff NLP research group's publicly available TimeLMs based sentiment analysis model [18]. This model builds upon the existing RoBERTa-base model, which in turn is a BERT-base model that has been pre-trained on a large corpus of English data using the training approach as discussed in its research paper [19]. This sentiment analysis model has been pre-trained on a dataset of approximately 124 million Tweets collected between Jan 2018 - Dec 2021. To add further detail, their research aims at building language models (LMs) that specializes in understanding diachronic Twitter data [20]. Their model has been bench marked against the TweetEval task [21], and the results can be checked in their paper. We expect this model to perform optimally for our use-case, since it has been pre-trained on Tweet data that is temporally close to our collected Tweets, which translates to better understanding of the context of Twitter English language used during that period.

The model classifies a Tweet into three sentiments - Negative, Neutral and Positive, and returns an array of three confidence values which correspond to the probabilities of the Tweet being classified into the corresponding sentiments.

Thus, utilizing this transformer model, all Tweets available for a given day and a given sector are scored and aggregated by the Spark Processing Engine. With the use of Apache Spark and its fundamental RDD (Resilient Distributed Datasets) data structure, we are able to take advantage of parallelization to process Tweets efficiently. The overall mood observed on Twitter on a given day for each market sector is thus captured

and quantified.

6. Correlation Analysis

6.1. Methodology

After processing the entire stream of Tweet data and obtaining intermediate results based on the sentiment scores, correlation analysis is performed.

The aggregated sentiment scores obtained from the Spark Engine is grouped sector-wise and merged with its relevant stock market ticker data, according to a date-wise order. For example, the aggregated Tweet results obtained under the EVs (Electric Vehicles) category are merged with the stock market data of Tesla Motors (NASDAQ: TSLA). Table 2 describes the merged dataset's schema. In essence, we now have a combined dataset that quantifies the performance of a given stock and the social media mood captured using Twitter on that day for the stock's relevant market sector.

Attribute	Description
Category	Market sector under consideration
Date	UTC date that the Tweet belongs to
Open	Opening price of the stock on the given date in USD ($)
Close	Closing price of the stock on the given day in USD ($)
Wted_Neg	Daily aggregated negative sentiment score weighted based on retweet count
Wted_Neu	Daily aggregated neutral sentiment score weighted based on retweet count
Wted_Pos	Daily aggregated positive sentiment score weighted based on retweet count

Table 2: description of the dataset schema.

We then perform feature scaling, i.e. normalization of the data using min-max scaling, whose formula is highlighted in (1) and (2). This is done since different features present in the dataset have different ranges, and it is necessary to bring them all to the same scale, to avoid bias and to perform further analysis.

$$y' = \frac{y - x_{min}}{x_{max} - x_{min}}(x'_{max} - x'_{min}) + x'_{min} \tag{1}$$

For Min-Max normalization, typically $\langle x'_{min}, x'_{max} \rangle = \langle 0, 1 \rangle$.

$$y' = \frac{y - x_{min}}{x_{max} - x_{min}} \tag{2}$$

where:

$\langle x_{min}, x_{max} \rangle$ is the old range.
$\langle x'_{min}, x'_{max} \rangle$ is the new range.
$y \in \langle x_{min}, x_{max} \rangle$ is the value to be normalized.
$y' \in \langle x'_{min}, x'_{max} \rangle$ is the min-max normalized value.

Following this, correlation analysis is performed with the Spearman's Correlation Coefficient, which is a non-parametric measure of rank correlation between two variables. It is equal to the Pearson's Correlation Coefficient between the rank values of those two variables. While the Pearson's correlation assesses linear

relationships, Spearman's correlation assesses monotonic relationships (irrespective of linearity). It is a better correlation metric than the Pearson's correlation metric for this specific scenario since the stock market may not be linearly related to the sentiments observed on social media.

Hypothesis 1 (Null Hypothesis) *There is no monotonic association between the two variables (in the population).*

Hypothesis 2 (Alternate Hypothesis) *There is a monotonic association between the two variables (in the population).*

A statistical test of significance is also performed for the observed Spearman's correlation coefficient values, to verify its significance as per Hypotheses 1 and 2. The correlation value is considered statistically significant (i.e. the alternate hypothesis is accepted) if the two-tailed p-value is lower than 0.01 (for a 99% confidence interval). Otherwise, the correlation is statistically insignificant (i.e. the null hypothesis is accepted), and can be ignored for all practical purposes.

$$\rho = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{3}$$

Where:

ρ is Spearman's coefficient, $\rho \in \langle -1, 1 \rangle$
$cov(R(X), R(Y))$ is the covariance of the rank variables.
$\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviations of the rank variables.

If all n ranks are all distinct, it can be computed using the formula:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{4}$$

where:

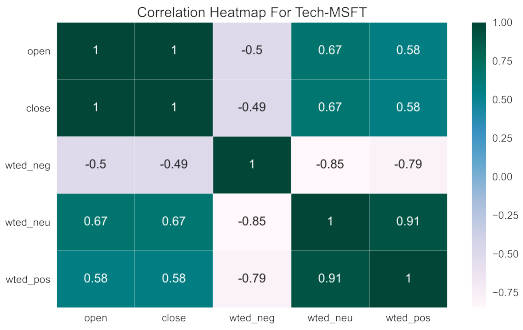
ρ is Spearman's coefficient, $\rho \in \langle -1, 1 \rangle$
n is sample size.
d_i is the difference between the ranks of the data being analyzed.

This procedure is carried out using the equations shown in (3) and (4) for each aggregated sector data separately to observe and tabulate the correlation between that sector's stock market trend and its' respective social media mood.

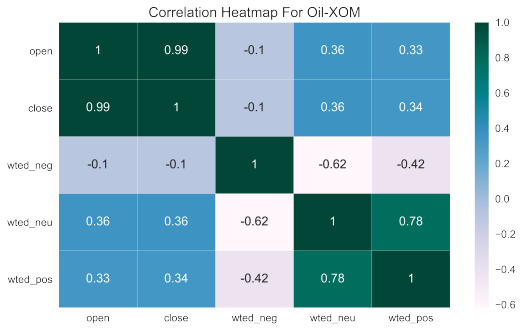
6.2. Results and Discussion

For our analysis, we consider only the daily close value to understand the correlation between a sector's market performance and the corresponding day's social media trend.

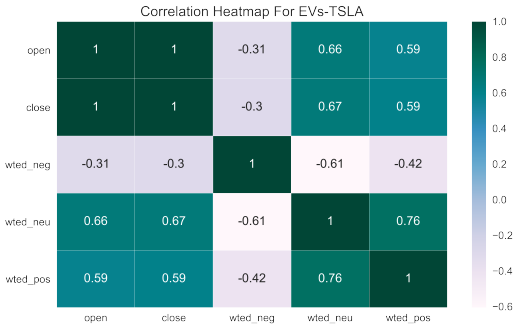
Upon category-wise examination of the obtained results shown in Figure 2 with respect to the Technology sector, and its reference stock market entity - Microsoft Corp., a strong positive correlation was obtained between



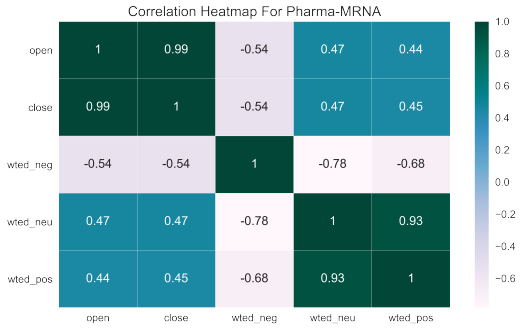
(a) Tech - (NASDAQ: MSFT)



(b) Oil - (NYSE: XOM)



(c) EVs - (NASDAQ: TSLA)



(d) Pharma - (NASDAQ: MRNA)

Figure 2: Correlation heatmap between Tweets and Stock Market for different industrial sectors.

Table 3: p-values for Spearman correlation between weighted sentiments and close.

(a) (NASDAQ: MSFT)

Feature	ρ	p
Wted_Neg	-0.493	$5.418e-46$
Wted_Neu	0.667	$3.368e-95$
Wted_Pos	0.576	$5.669e-66$

(b) (NYSE: XOM)

Feature	ρ	p
Wted_Neg	-0.100	$7.235e-3$
Wted_Neu	0.364	$5.945e-24$
Wted_Pos	0.339	$6.987e-21$

(c) (NASDAQ: TSLA)

Feature	ρ	p
Wted_Neg	-0.299	$2.657e-16$
Wted_Neu	0.665	$9.465e-93$
Wted_Pos	0.593	$2.030e-69$

(d) (NASDAQ: MRNA)

Feature	ρ	p
Wted_Neg	-0.542	$4.398e-57$
Wted_Neu	0.469	$2.514e-41$
Wted_Pos	0.446	$4.001e-37$

the closing price and Wted_Neu (0.67) and Wted_Pos (0.58), respectively (Figure 2a). A medium negative correlation was obtained between the closing price and Wted_Neg (−0.49) attribute. These results indicate that there is a good correlation between general social media discussion of the Tech sector and the stock market in both ways, indicating that the technology market is volatile to the mood of social media. The observed correlations are all statistically significant (from Table 3a).

Upon examining the Oil sector tweets with the Exxon Mobil. Corp stock performance (Figure 2b), there is a moderate positive correlation between the Wted_Neu (0.36) and Wted_Pos (0.34) to the closing price. The weak correlation between the Wted_Neg and the closing price (−0.1) indicates that the oil market does not fluctuate if there are negative sentiments being expressed about it on Twitter - making it a lowrisk stock in that aspect. But the market closing price shows considerable appreciation with positive/neutral discussion about it on the social media platform, although not being very susceptible to social media trends. These correlations are also statistically significant, as concluded from Table 3b.

For the Electric Vehicles (EVs), we see that there is a strong correlation between the weighted neutral sentiment score (Wted_Neu) and Close (0.66), as well as between the weighted positive sentiment score (Wted_Pos) and Close(0.59) for the Tesla (EVs) dataset (Figure 2c). There is a moderate negative correlation between the weighted negative sentiment score (Wted_Neg) and Close (−0.31). These results can be interpreted as follows: The neutral sentiment's result indicates that the Tesla stock performs well with a rise in the amount of discussion about electric vehicles in general. An overall positive mood about the EV market correlates to the appreciation of Tesla Inc.'s stock market value. An overall negative sentiment, however, does not translate as much to depreciation of the market value, due to its moderate correlation. Table 3c shows that these correlation values are all statistically significant.

In the Pharmaceutical sector, with its stock market representative - Moderna Inc., it was observed (Figure 2d) that there is a medium positive correlation between the three weighted attributes - Wted_Neg (−0.54), Wted_Neu (0.47) & Wted_Pos (0.45). This indicates that the Pharma sector's market behaviour is also volatile to the sentiments expressed on social media, especially on the negative side - a negative mood correlates more to the pharmaceutical sector's market price going down. From Table 3d, we can confidently conclude that the observed correlation values for this dataset are statistically significant.

7. Prediction

7.1. Methodology

Following our primary objective of performing analysis, we decided to additionally perform prediction of the next-day's stock market closing value using machine learning algorithms, since we observed a good correlation between the weighted Tweet sentiment scores and the close price of the market. From Nayak et al.'s [6] research on stock market prediction augmented by sentiment analysis of Twitter and news, it can be seen that Boosted Decision Trees perform the best in this scenario when compared to other algorithms like Logistic Regression and Support Vector Machines. Thus, we decided to compare and contrast two widely-used decision tree algorithms that employ the concept of boosting to predict the close value of the above used stock listings, augmented by the weighted Tweet sentiment scores.

Separate regressor models were constructed using XGBoost and CatBoost for each market sector. The datasets were split in an $85:15$ ratio for training and testing purposes.

7.1.1. XGBoost

The first boosting algorithm under consideration is DMLC's XGBoost (eXtreme Gradient Boosting) [22] which was presented in 2014 as an optimized gradient-boosting library that is designed to be highly efficient, flexible and portable. From its inception, it has gained universal acclaim and usage in the data science community.

Optimal hyper parameters for the model was found with the usage of a grid search algorithm and its results are listed in Table 4.

Hyper Parameter	Value
Maximum Tree Depth	3
Number of Estimators	500
Objective Function	Squared Error
Booster	GB Tree
Learning Rate	0.1
Column Sampling by Tree	1
Evaluation Metric	Mean Absolute Error (MAE)

Table 4: XGBoost - Optimal hyper parameters.

7.1.2. CatBoost

The second algorithm under consideration is Yandex's CatBoost [23], also a popular gradient-boosting algorithm introduced in 2017 with native support of categorical features and GPU usage for faster training times.

Optimal hyper parameters for the model was found with the usage of a grid search algorithm and its results are listed in Table 5.

Hyper Parameter	Value
Tree Depth	4
Iterations	200
Learning Rate	0.05
L2 Leaf Regularization	10
Loss Function	Mean Absolute Error (MAE)

Table 5: CatBoost - Optimal hyper parameters.

Separate regressor models were constructed for each market sector for analysis. The datasets were split in an $85:15$ ratio for training and testing purposes.

7.2. Results and Discussion

The two models are compared and contrasted based on the metrics obtained for them on the same test data. The scores are tabulated in Tables 6 and 7.

It was observed that both models were able to capture the trend of the closing price movement (increase/decrease), and were able to predict the next day's closing price with high accuracy. The XGBoost model performs best (83.81% accurate) on the Tech-MSFT dataset (from Table 6a) , which follows from the observed high correlation between its weighted sentiment scores and its closing price. The observed metrics are better in all cases for the XGBoost model, except for the Pharma-MRNA dataset , for which the CatBoost model showed the best performance of 84.61% accuracy (Table 7d).

Table 6: XGBoost regressor metrics.

(a) (NASDAQ: MSFT)

RMSE	0.0471
MSE	0.0022
MAE	0.0377
R2 Score	0.8381
Explained Variance Score	0.8427
Max Error	0.1403

(b) (NYSE: XOM)

RMSE	0.1914
MSE	0.0366
MAE	0.1420
R2 Score	-0.6285
Explained Variance Score	0.2131
Max Error	0.4236

(c) (NASDAQ: TSLA)

RMSE	0.0562
MSE	0.0031
MAE	0.0463
R2 Score	0.7086
Explained Variance Score	0.7555
Max Error	0.1630

(d) (NASDAQ: MRNA)

RMSE	0.0366
MSE	0.0013
MAE	0.0255
R2 Score	0.8185
Explained Variance Score	0.8370
Max Error	0.1885

Table 7: CatBoost regressor metrics.

(a) (NASDAQ: MSFT)

RMSE	0.0480
MSE	0.0023
MAE	0.0383
R2 Score	0.8319
Explained Variance Score	0.8427
Max Error	0.1288

(b) (NYSE: XOM)

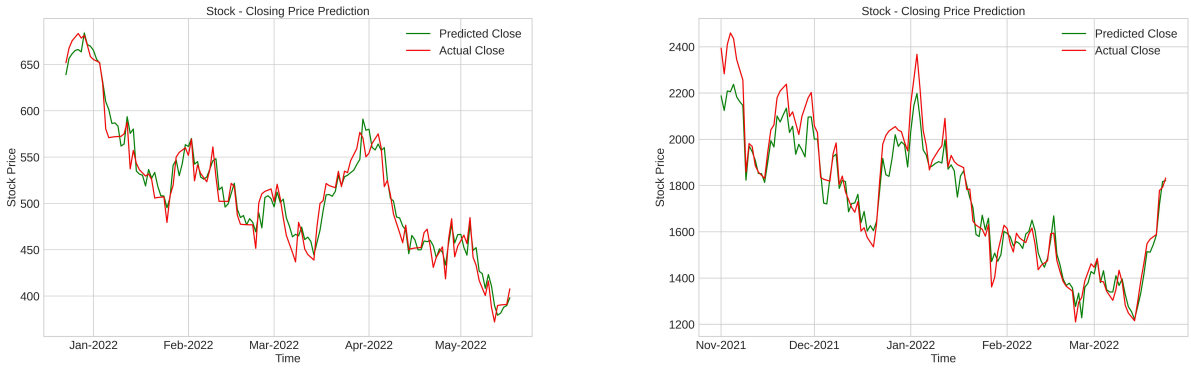
RMSE	0.1962
MSE	0.0385
MAE	0.1457
R2 Score	-0.7114
Explained Variance Score	0.1810
Max Error	0.4340

(c) (NASDAQ: TSLA)

RMSE	0.0711
MSE	0.0050
MAE	0.0562
R2 Score	0.5328
Explained Variance Score	0.6047
Max Error	0.2063

(d) (NASDAQ: MRNA)

RMSE	0.0337
MSE	0.0011
MAE	0.0251
R2 Score	0.8461
Explained Variance Score	0.8642
Max Error	0.1491



(a) XGBoost model performance on Tech-MSFT data. (b) CatBoost model performance on Pharma-MRNA data.

Figure 3: Time Series Plots on the test data.

For emphasis, the best performance of both models are also depicted using a time-series plot on the test data results, as shown in Figures 3a and 3b. It can be observed that the trend is captured accurately, and the difference between the predicted values and actual values is minimal, even for the other datasets as evidenced by the RMSE values obtained by using XGBoost models - 0.1914 for Oil-XOM (Table 6b), 0.0562 for EVs-TSLA (Table 6c) and 0.0366 for Pharma-MRNA (Table 6d). The CatBoost models also show a similar result in terms of RMSE score for the other datasets (apart from the best performing one), with it obtaining 0.0480 on the Tech-MSFT dataset (Table 7a), 0.1962 on Oil-XOM (Table 7b) and 0.0711 on the EVs-TSLA dataset (Table 7c).

Thus, we believe that using either model for this task would be prudent, although there is a slight advantage to using XGBoost in terms of reliability, as evidenced by our results.

8. Conclusion

In our study, we have considered four stock market segments - Electric Vehicles, Oil & Gas, Technology and Pharmaceuticals - and four representative stock market entities - Tesla Inc., Exxon Mobil Corp., Microsoft Corp., Moderna Inc. to analyse the correlation between the sentiments observed in social media to the respective entity's stock market valuation.

In this paper, we have shown conclusive proof that there is a strong correlation between the mood expressed on social media to stock market behaviour, although the extent of the correlation varies for each market sector. We were also able to identify two candidate machine learning algorithms - CatBoost and XGBoost, that can be utilized for efficient stock market prediction using the sentiment values. Our main contribution in this area of research has been the incorporation of state-of-the-art NLP models for sentiment analysis, and a more fine-grained approach to correlate stock market listings to related social media data, instead of a generalized approach that performs mood analysis to stock market indices like the DJIA, as done by other authors such as Lee et al. [3] and Mittal et al. [5]. We also used a significantly larger dataset to broaden the sample space under consideration, and the results we obtained show great promise in terms of practical use and further research.

Finally, we have also shown that stock market prediction based on sentiment-analysis of social media data is a worthwhile methodology that can be used to bolster existing stock prediction algorithms that make

use of other features.

9. Future Work

With regard to future work, the inclusion of data from other social media websites like Facebook, Reddit etc. can present an even broader picture of the sentiment expressed by users online. The inclusion of the sentiments reflected in news articles could also prove to be more effective in trend analysis of the stock market. Thus, data from multiple sources can be pooled together and aggregated with big-data architecture. Another consideration is the fact that the social media data that we work with are primarily produced by the English-speaking audience. Similar NLP techniques that work with other languages would be required to perform similar analyses in non-English speaking countries. With the inclusion of non-English Tweets (and other social media posts), it is possible to obtain a higher correlation score. The stock market data prediction task can further be improved and made reliable with the addition of other features that affect the stock market, apart from social media. To broaden the scope of this research, many other market sectors can also be considered to perform such sentiment-based analysis and prediction.

References

[1] A. Kanavos, G. Vonitsanos, A. Mohasseb, and P. Mylonas, "An entropy-based evaluation for sentiment analysis of stock market prices using twitter data," in *15th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2020, Zakynthos, Greece, October 29-30, 2020*, pp. 1–7, IEEE, 2020.

[2] "Elon Musk - GameStop Tweet [online]." Website: https://cnb.cx/3qVd5LM. [accessed 03 06 2022].

[3] C. Lee and I. Paik, "Stock market analysis from twitter and news based on streaming big data infrastructure," in *IEEE 8th International Conference on Awareness Science and Technology, iCAST 2017, Taichung, Taiwan, November 8-10, 2017*, pp. 312–317, IEEE, 2017.

[4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science, Vol. 2, No. 1*, pp. 1–8, Mar 2011.

[5] A. Mittal and A. Goel, "Stock prediction using twitter sentiment analysis," *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, vol. 15, p. 2352, 2012.

[6] A. Nayak, M. Pai, and R. Pai, "Prediction models for indian stock market," *Procedia Computer Science, Vol. 89*, pp. 441–449, 2016.

[7] V. Kalyanaraman, S. Kazi, R. Tondulkar, and S. Oswal, "Sentiment analysis on news articles for stocks," in *Proceedings of the 2014 8th Asia Modelling Symposium*, AMS '14, (USA), p. 10–15, IEEE Computer Society, 2014.

[8] Z. Peng, "Stocks analysis and prediction using big data analytics," *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pp. 309–312, 2019.

[9] V. S. Pagolu, K. Challa, G. Panda, and B. Majhi, "Sentiment analysis of twitter data for predicting stock market movements," *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pp. 1345–1350, 2016.

[10] S. Mehtab and J. Sen, "A robust predictive model for stock price prediction using deep learning and natural language processing," *SSRN Electronic Journal*, Jan 2019.

[11] D. Selvamuthu, V. Kumar, and A. Mishra, "Indian stock market prediction using artificial neural networks on tick data," *Financial Innovation, Vol. 5*, pp. 1–12, 2019.

[12] "Apache Kafka [online]." Website: https://kafka.apache.org/downloads. [accessed 09 03 2022].

[13] "Apache Spark [online]." Website: https://www.apache.org/dyn/closer.lua/spark/spark-3.2.1. [accessed 02 03 2022].

[14] F. Gürcan and M. Berigel, "Real-time processing of big data streams: Lifecycle, tools, tasks, and challenges," *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–6, 2018.

[15] "Twitter-API [online]." Website: https://developer.twitter.com/en/products/twitter-api. [accessed 10 02 2022].

[16] "Polygon-API [online]." Website: https://api.polygon.io/v1/open-close/. [accessed 10 02 2022].

[17] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Vol. 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, 2019.

[18] "Twitter-Roberta [online]." Website: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest. [accessed 07 01 2022].

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR, abs/1907.11692*, 2019.

[20] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados, "Timelms: Diachronic language models from twitter," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pp. 251–260, Association for Computational Linguistics, 2022.

[21] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 1644–1650, Association for Computational Linguistics, Nov. 2020.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 785–794, ACM, 2016.

[23] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6639–6649, 2018.