Predicting stock prices using ensemble learning techniques

Salma Elsayed[1], Ahmad Salah[1,2], Ibrahim Elhenawy[1], Marwa Abdellah[1]
[1]Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig, Egypt
[2]Department of Information Technology, College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri, Sultanate of Oman

Article Info

Article history:

Received Dec 28, 2023
Revised Aug 13, 2024
Accepted Oct 23, 2024

Keywords:

Deep learning
Ensemble learning
Machine learning
Prediction regression
Stock price

ABSTRACT

Stock price prediction has grown in importance due to its role in determining the future worth of business shares. There are several approaches for stock price prediction that can be classified into machine learning, deep learning, and ensemble learning methods. To predict stock prices, we proposed collecting a dataset for different well-known stocks, *e.g.,* Microsoft. The utilized datasets consist of two parts; the first part contains a set of tweets for the stocks under investigation in this study which were collected from the X social media platform and the other part contains the stock prices. Sentimental features of the tweets were extracted and merged with the stock price changes. Then, we framed the problem as a regression task. we aim to analyze the performance gap between ensemble learning and other machine learning (ML) and deep learning (DL) models for predicting stock prices based on tweets. In this context, different ensemble learning models were proposed to predict the price change of each stock. Besides, several machine learning and deep learning models were used for comparison purposes. Several evaluation metrics were utilized to evaluate the performance of the proposed models. The experimental results proved that the stacking regressor model outperformed the other models.

Corresponding Author:

Ahmad Salah
Department of Computer Science, Faculty of Computers and Informatics, Zagazig University
Alzeraa Square, Zagazig, Sharkia, 44519, Egypt
Email: ahmad@zu.edu.eg

1. INTRODUCTION

The internet serves as an online learning platform for communication and exchanging ideas. Through common social media platforms, people can contribute feedback and suggestions for a wide range of services and offerings. Twitter, Facebook, and Google+ are well-known examples of social media platforms that are utilized for idea posting. Twitter is an online social network where millions of tweets are posted every day. The prediction method may be carried out using Twitter data. Live Twitter data may be collected via the Twitter API and analyzed using the classifier. The stock market is an important part of the economy and affects commerce changes and industrial growth. Several data mining approaches are employed to handle variations in the stock market, and financial news articles are assumed to influence stock prices [1], [2], [3].

In [4], a unique sentiment indicator based on weighted textual contents and financial aberrations to forecast stock changes is performed. First, the authors suggested a unique weighting approach for each stock movement. Then, they produced an actual adjusted sentiment measure that was more accurate by accounting for the day of the week and vacation. Using support vector machine (SVM) [5], [6], decision tree (DT), gradient boosting decision tree (GBDT), random forest (RF) [7] naïve Bayes (NB), K-nearest neighbor

(KNN) [8], and logistic regression (LR) algorithms. They observed that the modified sentiment measure can effectively improve stock market movement prediction.

In study [9], a unique teaching and learning based optimization (TLBO) model using long short-term memory (LSTM) [10], based on sentiment analysis (SA) for stock price prediction using data collected from Twitter is conducted. The authors predicted stock prices using four main steps: pre-processing, classification, learning rate schedule, and output unit optimization. Furthermore, an LSTM model was used to categorize tweets into positive and negative sentiments on stock values. Using several methods, experimental results of the TLBO-LSTM model outperform the latest techniques with a maximum accuracy of 95.33%, a recall of 85.28%, and an F1-score of 90%.

Forecasting the trend of stock prices is a critical mission that assists investors in making sound financial decisions in the stock market. Thus, the goal of the current work is to considerably minimize the risk of trend prediction using the ensemble learning technique. The efficiency of the suggested approach for predicting stock price is proved by testing on 22 corporations such as TSLA, Amazon, META, and Microsoft using Twitter data. Each dataset is divided into training and testing sets for the experiment. In the proposed work, a pre-processing of the Tweets data is conducted to convert unstructured data into meaningful text. Then, the polarity of the tweets was extracted using two different approaches. In the first approach, the VADER model, implemented in study [11], was used to extract the polarity of each tweet and obtain the sentimental score (in the range of -1 to 1). In the second approach, we chose the maximum value of positive, negative, and neutral percentages as the polarity score. The daily closing price's percentage change is computed based on the percentage between two consecutive days' closing prices and based on this percentage we could compute the polarity of the tweets. Then, we matched the daily closing price changes to the two extracted polarities to find out which approach was more effective. The highest number of matches with the financial closing price changes was the sentimental score to extract polarity, the first approach.

The training set contains 80% of the dataset, and the test set is made up of the remaining 20% of the dataset. The ensemble models are mainly classified into stacking, blending, bagging, and boosting. This study compares five ensemble learning models which are $CatBoostClassifier$, $CatBoostRegressor$, $BaggingRegressor$, $GradientBoostingRegressor$, and $StackingRegressor$. Each predictive model is evaluated by five metrics which are mean-absolute-error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE), root mean squared error (RMSE), and R squared (R2). Moreover, several machine learning models such as KNN, support vector machine (SVM), and linear regression (LR) are used for comparison purposes to find out the best accuracy. The MAPE metric of the Stacking Regressor model was the lowest of the tested models, indicating that the Stacking Regressor model outperformed the other models. The ensemble models effectively forecast stock closing prices, as they achieved low values in most metrics. The main contributions of this work are as follows:

- Extracting tweets' sentimental features in two different methods. Then, the extracted features are to be used for generating an efficient dataset from the original one.
- Using ensemble learning models to predict the stock price and comparing their results with different machine learning (ML) and deep learning (DL) models.
- 22 different stocks with several evaluation metrics were utilized to thoroughly evaluate the performance of the proposed predictive models against one of the existing research works.

The remaining parts of the paper are as follows. Section 2 discusses some papers that used ML and DL in predicting stock price based on sentiment analysis. The proposed methodology is explained in section 3. Section 4 summarizes the results of our study and constructs it in tables and bar charts. Finally, the paper is concluded in section 5.

2. RELATED WORK

2.1. ML-based stock price prediction models

Vijh *et al.* [12] proposed a model to predict the next closing price for five different firms from various fields using an artificial neural network (ANN) model [13] and RF model. They collected the dataset over 10 years from 4/5/2009 to 4/5/2019 for Nike, Goldman Sachs, Johnson, Johnson, Pfizer, and JP Morgan Chase and Co. Based on the RMSE, MAPE, and MAE metrics, the comparison study showed that the ANN model outperforms RF in stock price prediction. However, they should use more techniques in the comparison to ensure the accuracy of ANN in the prediction process.

Christanto *et al.* [14] suggested employing financial stock data as input parameters to several machine learning models such as SVM [15]. They obtained a dataset of 16 first such as NASDAQ, Nikkei 225, Hang Seng index, FTSE100, DAX, and ASX. The prices dataset is collected for the period from the 4^{th} of January 2000 to the 25th of October 2012. They used the multiple additive regression trees (MART) model (a decision tree-based boosting algorithm) and compared it against an SVM model. They found out that the volume of the training data is important for the SVM model because if its size is insufficient, the hyperplane might be unable to effectively divide the data. They used the RMSE metric [16] to evaluate the performance of their model. They applied linear regression, generalized linear model (GLM), and SVM to forecast the daily NASDAQ price movement. The results showed that the proposed SVM model outperforms the other models, as it achieved the lowest RMSE.

Sadorsky [17] utilized a random forests model to forecast the stock price direction of clean energy exchange-traded funds (EFTs). The authors used random forests and decision tree bagging models in the prediction task and compared their results with an ANN model [18] and SVMs. They used data on the stock value of five well-known firms, in the US-listed, and extensively traded clean energy ETFs. The daily data set begins on 1 January 2009 and finishes on 30 September 2020. The information was obtained from Yahoo Finance. The forecast accuracy of each ETF is evaluated across a period ranging from one day to twenty days. For predicting ranges of 10 days or longer, the forecasting accuracy of RF and tree bagging models exceeds 80%. They found that RF model [19] and decision tree bagging are simpler to predict than other ML models such as ANNs and SVMs. However, they should have used more metrics to ensure the performance of their techniques.

2.2. DL-based stock price prediction models

Mehta *et al.* [20] proposed to predict the Indian stock market during the period from the 1st of October 2014 to the 31st of December 2018 using sentiment analysis. They applied machine learning and deep learning techniques namely, SVM, linear regression, naïve Bayes, and LSTM. They analyzed the relationship between media data and market price rates over a constrained time and applied a range of factors from finalized data sets to enhance prediction accuracy. The findings proved that by employing online platforms and financial news data, LSTM [21] was able to achieve the highest accuracy of about 92.45%. Linear SVC classifier achieved the second-highest precision classifier. The naïve Bayes, linear regression, and maximum entropy methodology remained around 86.72%, 86.75%, and 88.93%, respectively.

In study [22], the proposed model included a reliable forecasting technique for the probability of stock market changes. Firstly, the most important financial market indicators are been chosen that might be utilized to forecast the stock market. Some statistical machine learning approaches have been applied as well as two new algorithms that produce better results in other scientific domains which are deep neural networks and extreme gradient boosting (XGBoost) [23]. The performance metrics suited for unbalanced datasets are been applied for testing the models. Their empirical findings showed that deep learning [24] produced better prediction accuracy.

Abdullah and Salah [25] implemented the CNN-LSTM model, a hybrid model that merges an LSTM model [26] and convolution neural network (CNN) architecture [27], [28]. The proposed hybrid model makes use of the convolution layer attributes for retrieving relevant features contained in time series data, in addition to the LSTM design's ability to learn long-term associations. The datasets used in the tests were gathered from Yahoo Finance for three years from the 1st of December 2016 to the 1st of December 2020 via a daily time interval. The analyses were performed on three unique dataset forms: stock market, foreign exchange tools, and cryptocurrency. Using two evaluation metrics namely, MSE and MAE. The proposed models outperformed the state-of-the-art approaches, machine learning techniques, and statistical techniques. The findings show that the suggested CNN-LSTM model outperforms the LSTM model and the other models on the majority of the datasets tested.

3. THE PROPOSED METHOD

In Figure 1, we utilized a dataset from the Kaggle website about stock tweets, as can be found in $https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction?select=stock\_tweets.csv$. Next, we conducted a preprocessing step to eliminate non-essential content. To ensure polarity, we extracted features from the tweets using two distinct methods. Subsequently, we utilized various ensemble, machine learning, and deep learning models to build the predictive models. Finally, we applied evaluation metrics to assess the performance of these prediction models.

In the overview subsection, we clearly discussed the steps of our model. In the dataset subsection, we detailed the process of generating the final dataset. The feature extraction subsection explained the two different methods used for feature extraction. In the predictive model's subsection, we described all the trained models. The implementation details subsection showcased the implementation of these models. Finally, in the evaluation metrics subsection, we illustrated the various metrics used to evaluate the trained models.

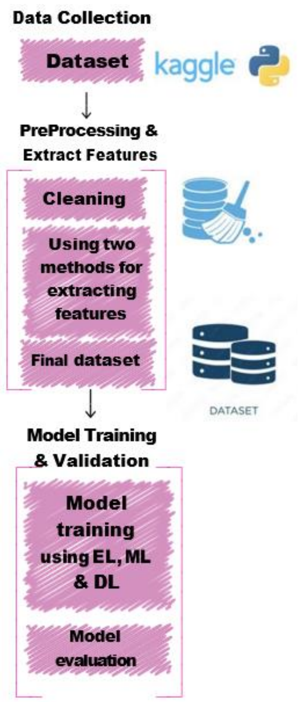


Figure 1. The steps of the proposed method

3.1. Overview

To perform modeling for the stock market prediction, we collected stock-related tweets from the X social media platform for different stocks from the Kaggle website. The extracted tweets contain unnecessary data like special characters, URLs, emojis, hashtags (#), and @. Thus, we have preprocessed these tweets to obtain only the plain sentences. Then, we extracted the features from the tweets in two different ways to ensure the polarity. Next, we compared the polarity of the tweets with the trend of the financial data to choose only the matched tweets and ignore the other tweets. The dataset size is reduced from 80,793 to 15,430. We divided the reduced dataset into 22 datasets based on the stock name. Thus, we obtained 22 datasets for 22 stocks.

To predict the price of each stock, we framed the problem of price prediction as a regression problem where the outcome variable is the predicted future stock price daily. Different ensemble learning models were utilized. Five ensemble learning models, namely $CatBoostClassifier$, $CatBoostRegressor$, $BaggingRegressor$, $GradientBoostingRegressor$, and $StackingRegressor$ were used. Each predictive model is evaluated on five metrics which are MAE, MAPE, MSE, RMSE, and R2. Each dataset is divided into training and testing sets for evaluation. The training set contains 80% of the dataset, and the test set is made up of the remaining 20% of the dataset. We utilized several machine learning algorithms which are KNN, SVM, and linear regression. Besides, we utilized a deep learning model, *i.e.*, an LSTM model [25], to ensure the accuracy of ensemble learning models.

3.2. Dataset

We have collected the dataset of different stocks from the Kaggle website on the following link ($https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction?select=stock\_tweets.csv$, last accessed on 23 March 2024). The dataset consists of two parts. The first part contains the tweets with a size of 80,793 and four features which are date, tweet, stock name, and company name. The other part of the dataset contains the stock price data with a size of 6,300 and eight features which are date, open price, highest price, lowest price, closing price, adjusted closing price, volume, and stock name. Of note, the extracted tweets contained unnecessary data like special characters, URLs, emojis, hashtags (#), and @. Thus, we have preprocessed these tweets to obtain only the cleaned sentences. Then, we extracted the features in two ways to ensure the polarity. After That, we chose the tweets whose polarity was matched with the trend of the financial data and ignored the others. Thus, we reduced the dataset size from 80,793 to 15,430. We divided the whole dataset into small datasets based on the stock name. We obtained 22 datasets for 22 stock names.

3.3. Features extraction

For the sake of feature extraction, we utilized two different methods to ensure the correctness of the polarity score. We used the valence aware dictionary for sentiment reasoning (VADER) model [29]. The VADER model is a sentiment analysis tool that can distinguish between polarity (positive or negative) and intensity of feelings. It is included in the natural language toolkit (NLTK) package [30] and can be used effectively on plain text data. The sentimental score is in the range of -1 to 1 following the condition denoted in (1).

$$Polarity = \begin{cases} Positive, & if\ x \geq 0.5 \\ Neutral, & if\ 0.5 > x \geq -0.5 \\ Negative, & otherwise \end{cases} \tag{1}$$

where x is the sentimental score value. We called the first method polarity 1. In the second method, we chose the maximum value of positive, negative, and neutral percentages and considered it the polarity score; we called this method polarity 2. Then, we computed the percentage change of the closing price as the percentage between two consecutive days of the closing prices and made the same condition as denoted in (1) to determine its polarity score. When we compared polarity 1 and polarity 2 with the polarity of the daily closing price change, we noticed that polarity 1 was matched with 15,430 rows of the daily closing price change while polarity 2 was matched with 8,854 rows of the daily closing price change. Thus, we utilized the polarity of the sentimental score, *i.e.*, polarity 1 in the proposed model.

We merged all the columns of the two datasets (tweets dataset and closing price changes dataset) based on the date and stock name to get one dataset merging the tweets' polarity and the stock price information. Thus, the final dataset contains the following features: date, stock name, adjusted close, sentiment score, and polarity. We divided each dataset in a ratio of 80% : 20% for training and testing sets.

3.4. The predictive models

For the predictive model, we framed the problem of stock price prediction as a regression problem. We proposed using different ensemble learning models [31] to predict the stock prices. Ensemble learning is a machine learning approach that involves training several learners to solve the same issue. The ensemble models are broadly categorized into stacking, blending, bagging, and boosting. Stacking is an advanced ensemble learning strategy in which individual model predictions are layered and utilized as input to train the meta-model. This meta-model is then applied to the test set to make predictions. The training data set is divided into n parts. The basic model is trained for each n−1 part. Blending is a method similar to stacking in those predictions are made using a validation set from the training set. The training data set is divided into training and validation sets. Bagging is an approach that combines the findings of individual models to provide a more generalized outcome. Individual models, however, are not given the same dataset. Instead, the bootstrapping approach is used to build replacement subsets of the original dataset. In the boosting approach, each consecutive model attempts to fix the errors in the prior model. As a result, subsequent models rely on the prior model. Boosting creates a subset from the entire dataset. A basic model is trained using this subset. This model makes predictions throughout the whole dataset. Incorrect forecasts have been noticed. Then, another base model is trained to fix the prior model's mistakes.

3.5. Implementation details

We utilized the $CatBoostClassifier$ class with a loss function of "$MultiClass$" value and 200 iterations. We use the implementation of the CatBoost and ipywidgets packages. The pool is an internal data structure of CatBoost that wraps the utilized data and target values. The pool can make the training process faster. Then, we fed the model with the training dataset to fit the model. Then, the evaluation function received the true values. We utilized the $BaggingRegressor$ and $GradientBoostingRegressor$ classes with default parameter values and fitted the model with X containing the list of values of the sentiment score and Y containing the list of values of the adjusted closing prices. Stacked generalization consists of piling the result of the single estimators and using a measure to calculate the final prediction. Through stacking, the effectiveness of each predictor can be utilized by feeding its output into the last predictor. Thus, we utilized the $StackingRegressor$ with the esti mators' parameters which are $DecisionTreeRegressor$, $LinearRegression$, and $LinearSVR$. Then we fitted the model with X train and Y train parameters. We tuned the $RandomForestRegressor$ and $DecisionTreeRegressor$ classes with 25 for the random state's parameter. We utilized the $KNeighborsRegressor$ class with 3 neighbors. Finally, we utilized the linear support vector regression ($LinearSVR$) and linear regression with the default parameters.

3.6. Evaluation metrics

To evaluate the performance of the prediction models, we used the evaluation metrics as in [32] which are: The MAE measure [33] evaluates the absolute difference between actual and forecasted results which is denoted in (2). It measures how far the forecasts differed from the actual outcome.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}[y - \bar{y}] \tag{2}$$

where y is the actual value, y' is the predicted value and n represents the size of the test set.

The MSE metric is computed by averaging the square of the difference between the actual and forecasted values, as denoted in (3).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \bar{y})^2 \tag{3}$$

The RMSE metric [34] is the square root of the average of the squared variance of the real and forecasted results which id denoted in (4).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \bar{y})^2} \tag{4}$$

The MAPE metric [35] evaluates a prediction model's accuracy. It computes how accurate the anticipated value was to the actual value by averaging the absolute percentage errors of all entries in the dataset, as denoted in (5).

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y-\bar{y}}{y}\right)^2 * 100 \tag{5}$$

where R^2 is denoted in (6). It compares the residual sum of squares (SS_{res}), (7), to the total sum of squares (SS_{tot}), (8). The total sum of squares is computed by accumulating the squares of the perpendicular intervals between data points and the average line.

$$R^2 = 1 - \left[\frac{SS_{res}}{SS_{tot}}\right] \tag{6}$$

where

$$SS_{res} = Sum(y - \bar{y})^2 \tag{7}$$

$$SS_{tot} = \sum_{i=1}^{n}(y - \bar{y})^2 \tag{8}$$

4. RESULTS AND DISCUSSION

4.1. Setup

The experiments were conducted on a PC with a 64-bit Windows 11 OS with an Intel 7-core processor running at 3.20 GHz and 16 GB RAM. Scikit learns libraries were used to implement the ensemble learning and machine learning models. The utilized source code and dataset is publicly available [36].

4.2. Results

In this section, we evaluated the performance of all models using various metrics, namely, MAPE, MAE, MSE, RMSE, and R2. The comparison between the proposed ensemble learning models, ML models, and an existing work utilizing a hybrid CNN-LSTM model [25] revealed that the ensemble learning models, *e.g.*, $StackingRegressor$, $CatBoostClassifer$, and $BaggingRegressor$ perform better for most of the different stocks. Considering the MAPE metric to determine the most effective algorithm, the listed MAPE values in Table 1 show that the TSLA dataset is reduced from 0.22 for the CNN-LSTM model [25] to 0.15 for the $CatBoostRegressor$ model. A detailed results of the MAPE, MAE, and R2 metrics for the ensemble learning models and the hybrid CNN-LSTM model [25] for the TSLA, MSFT, AMZN, GOOG, AMD, and NFLX datasets are shown in Table 1.

Using TSLA, MSFT, AMZN, GOOG, and AMD datasets as examples, the bar charts in Figures 2 and 3 show the comparison between the $StackingRegressor$, $CatBoostClassifier$, $CatBoostRegressor$, *BaggingRegressor*, *GradientRegressor*, and hybrid CNN-LSTM [25] models' performance based on the RMSE metric and MSE metric. The results proved that the proposed models perform better than the other techniques with most of the datasets. The visual representation through bar charts enables clear comparison of performance differences across all tested models and datasets.

Table 1. A comparison between different models based on MAPE, MAE, and R2 metrics for many stocks

Stock Name/metrices		*StackingRegessor*	*CatBoostClassifier*	*CatBoostRegressor*	*BaggingRegressor*	*GradientBoosting*	CNN-LSTM [23]
	MAPE	0.169	0.214	0.155	0.183	0.170	0.222
TSLA	MAE	49.132	74.311	55.050	64.482	60.305	61.881
	R2	-0.646	-2.260	-0.645	-1.470	-1.149	-1.230
	MAPE	0.053	0.134	0.105	0.132	0.115	0.227
MSFT	MAE	16.511	43.007	34.366	42.600	73.245	58.670
	R2	-0.070	-6.426	-3.778	-6.172	-4.504	-10.997
	MAPE	0.187	0.164	0.184	0.174	0.178	0.293
AMZN	MAE	32.210	28.195	31.774	30.038	30.693	38.803
	R2	-29.877	-33.713	-29.301	-31.666	-30.223	-45.217
	MAPE	0.042	0.150	0.119	0.135	0.137	0.208
GOOG	MAE	6.206	21.825	17.309	19.631	19.919	42.846
	R2	-1.376	-25.792	-13.178	-18.217	-19.265	-27.258
	MAPE	0.147	0.230	0.180	0.209	0.199	0.265
AMD	MAE	21.215	31.261	25.734	28.997	27.699	27.278
	R2	-0.962	-3.177	-1.680	-2.452	-2.141	-1.895
	MAPE	0.213	0.616	0.439	0.459	0.492	2.814
NFLX	MAE	135.822	391.766	297.447	292.522	313.670	468.603
	R2	-61.593	-524.44	-270.874	-342.468	-368.860	-728.341

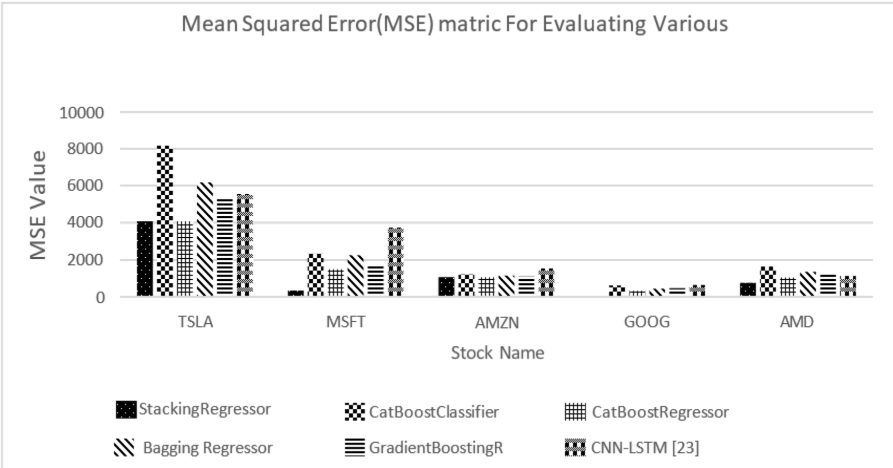


Figure 2. The MSE values for various models

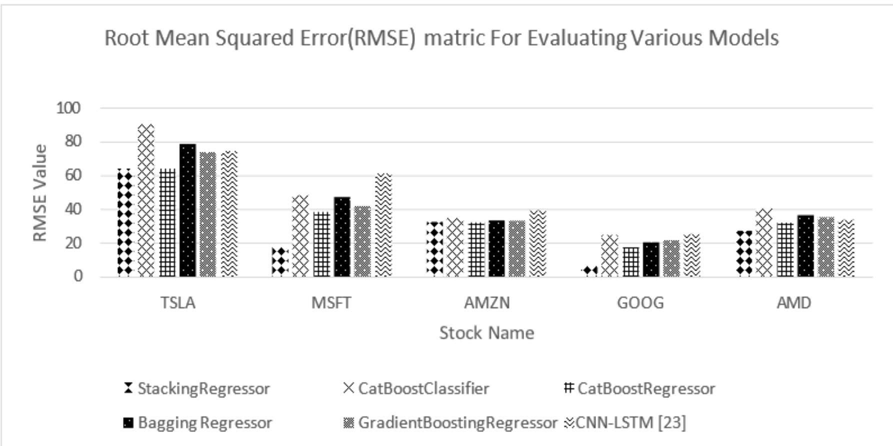


Figure 3. The RMSE values for various models

In Figure 2, the stacking regressor generally outperforms in accuracy, as indicated by lower MAPE values for a majority of the stocks. This suggests its robustness in handling financial data's complexities. The high MAPE for NFLX under the hybrid CNN-LSTM [25] model signals potential overfitting or model incompatibility with highly volatile stock data. In Figure 3, again, the stacking regressor consistently shows lower RMSE across all stocks, indicating higher prediction accuracy. In contrast, the CatBoost classifier tends to exhibit higher RMSE, particularly for TSLA, which could suggest less predictive reliability for that stock.

The detailed comparison across different stocks suggests specific models excel in certain areas; for instance, Stacking Regressor generally outperforms in accuracy, as indicated by lower MAPE values for a majority of the stocks. This suggests its robustness in handling financial data's complexities. The strikingly high MAPE for NFLX under the hybrid CNN-LSTM model [25] signals potential overfitting or model incompatibility with highly volatile stock data. The negative R2 values across models and stocks underscore the challenge of modeling stock behavior accurately, highlighting the financial market's unpredictability and the necessity for sophisticated modeling techniques that can adapt to its volatile nature.

5. CONCLUSION

In this work, we utilized ensemble learning models to predict the prices of 22 stocks based on the collected tweets from the X social media platform. We proposed merging the tweets with the stock closing price changes in one dataset to be able to predict the stock price change based on the tweets. First, we preprocessed the obtained tweets to eliminate the unstructured data. Then, we extracted the tweets' polarity with two different methods to ensure the correctness of the polarity. Next, we divided the dataset based on the stock name into 22 datasets. Then, we utilized several ensemble learning models for the predictive task. The proposed ensemble learning models were evaluated against several machine learning and deep learning models. Five different evaluation metrics were utilized, namely, MAPE, MAE, MSE, RMSE, and R^2. The experimental results outlined that the proposed ensemble learning models perform better than the state-of-the-art model and the machine learning models on average for most stocks. Furthermore, the findings of evaluation metrics showed that the stacking regressor model outperformed the other models, as it achieved the lowest MAPE value. Future research will consider merging financial sentiment analysis approaches with the proposed model.

REFERENCES

[1] R. Chopra and G. D. Sharma, "Application of artificial intelligence in stock market forecasting: A critique, review, and research agenda," *Journal of Risk and Financial Management*, vol. 14, no. 11, Nov. 2021, doi: 10.3390/jrfm14110526.

[2] Padmanayana, Varsha, and Bhavya K, "Stock market prediction using Twitter sentiment analysis," *International Journal of Scientific Research in Science and Technology*, vol. 2, no. 1, pp. 265–270, Jul. 2021, doi: 10.32628/CSEIT217475.

[3] A. M. Priyatno, W. F. R. Sudirman, and R. J. Musridho, "Feature selection using non-parametric correlations and important features on recursive feature elimination for stock price prediction," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, pp. 1906–1915, Apr. 2024, doi: 10.11591/ijece.v14i2.pp1906-1915.

[4] Y. Qiu, Z. Song, and Z. Chen, "Short-term stock trends prediction based on sentiment analysis and machine learning," *Soft Computing*, vol. 26, no. 5, pp. 2209–2224, Jan. 2022, doi: 10.1007/s00500-021-06602-7.

[5] R. J. Kuo and T. H. Chiu, "Hybrid of jellyfish and particle swarm optimization algorithm-based support vector machine for stock market trend prediction," *Applied Soft Computing*, vol. 154, Mar. 2024, doi: 10.1016/j.asoc.2024.111394.

[6] P. Chhajer, M. Shah, and A. Kshirsagar, "The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction," *Decision Analytics Journal*, vol. 2, Mar. 2022, doi: 10.1016/j.dajour.2021.100015.

[7] B. K. Meher, M. Singh, R. Birau, and A. Anand, "Forecasting stock prices of fintech companies of India using random forest with high-frequency data," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 1, Mar. 2024, doi: 10.1016/j.joitmc.2023.100180.

[8] M. A. Istiake Sunny, M. M. S. Maswood, and A. G. Alharbi, "Deep learning-based stock price prediction using LSTM and bi-directional LSTM model," in *2nd Novel Intelligent and Leading Emerging Sciences Conference, NILES 2020*, IEEE, Oct. 2020, pp. 87–92. doi: 10.1109/NILES50944.2020.9257950.

[9] G. M. Siddesh, S. R. M. Sekhar, and K. G. Srinivasa, "A long short-term memory network-based approach for predicting the trends in the S&P 500 index," *Journal of The Institution of Engineers (India): Series B*, vol. 105, no. 1, pp. 19–26, Feb. 2024, doi: 10.1007/s40031-023-00954-4.

[10] T. Swathi, N. Kasiviswanath, and A. A. Rao, "An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis," *Applied Intelligence*, vol. 52, no. 12, pp. 13675–13688, Mar. 2022, doi: 10.1007/s10489-022-03175-2.

[11] J. Yao, "Automated sentiment analysis of text data with NLTK," *Journal of Physics: Conference Series*, vol. 1187, no. 5, Apr. 2019, doi: 10.1088/1742-6596/1187/5/052020.

[12] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock closing price prediction using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 599–606, 2020, doi: 10.1016/j.procs.2020.03.326.

[13] A. H. Moghaddam, M. H. Moghaddam, and M. Esfandyari, "Stock market index prediction using artificial neural network," *Journal of Economics, Finance and Administrative Science*, vol. 21, no. 41, pp. 89–93, Dec. 2016, doi: 10.1016/j.jefas.2016.07.002.

[14] F. W. Christanto, V. G. Utomo, R. Prathivi, and C. Dewi, "The impact of financial statement integration in machine learning for stock price prediction," *International Journal of Information Technology and Computer Science*, vol. 16, no. 1, pp. 35–42, Feb. 2024, doi: 10.5815/ijitcs.2024.01.04.

[15] S. R. Thumu and G. Nellore, "Optimized ensemble support vector regression models for predicting stock prices with multiple kernels," *Acta Informatica Pragensia*, vol. 13, no. 1, pp. 24–37, Apr. 2024, doi: 10.18267/j.aip.226.

[16] Z. Khan, "Used car price evaluation using three different variants of linear regression," *International Journal of Computational and Innovative Sciences*, vol. 1, pp. 40–49, 2022.

[17] P. Sadorsky, "A random forests approach to predicting clean energy stock prices," *Journal of Risk and Financial Management*, vol. 14, no. 2, Jan. 2021, doi: 10.3390/jrfm14020048.

[18] E. Guresen, G. Kayakutlu, and T. U. Daim, "Using artificial neural network models in stock market index prediction," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10389–10397, Aug. 2011, doi: 10.1016/j.eswa.2011.02.068.

[19] N. Tripathy, S. K. Balabantaray, S. Parida, and S. K. Nayak, "Cryptocurrency fraud detection through classification techniques," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 3, pp. 2918–2926, Jun. 2024, doi: 10.11591/ijece.v14i3.pp2918-2926.

[20] P. Mehta, S. Pandya, and K. Kotecha, "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning," *PeerJ Computer Science*, vol. 7, pp. 1–21, Apr. 2021, doi: 10.7717/peerj-cs.476.

[21] N. Darapaneni *et al.*, "Stock price prediction using sentiment analysis and deep learning for Indian markets," *arXiv:2204.05783*, Apr. 2022.

[22] S. P. Chatzis, V. Siakoulis, A. Petropoulos, E. Stavroulakis, and N. Vlachogiannakis, "Forecasting stock market crisis events using deep and statistical machine learning techniques," *Expert Systems with Applications*, vol. 112, pp. 353–371, Dec. 2018, doi: 10.1016/j.eswa.2018.06.032.

[23] Y. Han, J. Kim, and D. Enke, "A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost," *Expert Systems with Applications*, vol. 211, Jan. 2023, doi: 10.1016/j.eswa.2022.118581.

[24] M. Ayitey Junior, P. Appiahene, O. Appiah, and C. N. Bombie, "Forex market forecasting using machine learning: Systematic Literature Review and meta-analysis," *Journal of Big Data*, vol. 10, no. 1, Jan. 2023, doi: 10.1186/s40537-022-00676-2.

[25] W. Abdullah and A. Salah, "A novel hybrid deep learning model for price prediction," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 3, pp. 3420–3431, Jun. 2023, doi: 10.11591/ijece.v13i3.pp3420-3431.

[26] S. Zaheer *et al.*, "A multi-parameter forecasting for stock time series data using LSTM and deep learning model," *Mathematics*, vol. 11, no. 3, Jan. 2023, doi: 10.3390/math11030590.

[27] Y. Zhao and G. Yang, "Deep learning-based integrated framework for stock price movement prediction," *Applied Soft Computing*, vol. 133, Jan. 2023, doi: 10.1016/j.asoc.2022.109921.

[28] S. Mukherjee, B. Sadhukhan, N. Sarkar, D. Roy, and S. De, "Stock market prediction using deep learning algorithms," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 82–94, Aug. 2023, doi: 10.1049/cit2.12059.

[29] O. Abiola, A. Abayomi-Alli, O. A. Tale, S. Misra, and O. Abayomi-Alli, "Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, Jan. 2023, doi: 10.1186/s43067-023-00070-9.

[30] B. A. Abdelfattah, S. M. Darwish, and S. M. Elkaffas, "Enhancing the prediction of stock market movement using neutrosophic-logic-based sentiment analysis," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 19, no. 1, pp. 116–134, Jan. 2024, doi: 10.3390/jtaer19010007.

[31] G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat, "Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis, and discussion of implications," *International Journal of Financial Studies*, vol. 11, no. 3, Jul. 2023, doi: 10.3390/ijfs11030094.

[32] M. Ali, D. M. Khan, H. M. Alshanbari, and A. A.-A. H. El-Bagoury, "Prediction of complex stock market data using an improved hybrid EMD-LSTM model," *Applied Sciences*, vol. 13, no. 3, Jan. 2023, doi: 10.3390/app13031429.

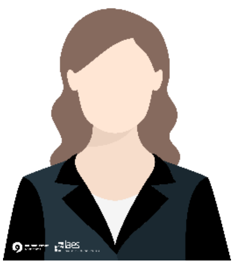
[33] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Engineering Reports*, vol. 5, no. 4, Nov. 2023, doi: 10.1002/eng2.12599.

[34] E. M. Ferrouhi and I. Bouabdallaoui, "A comparative study of ensemble learning algorithms for high-frequency trading," *Scientific African*, vol. 24, Jun. 2024, doi: 10.1016/j.sciaf.2024.e02161.

[35] A. T. Haryono, R. Sarno, and K. R. Sungkono, "Stock price forecasting in Indonesia stock exchange using deep learning: A comparative study," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 861–869, Feb. 2024, doi: 10.11591/ijece.v14i1.pp861-869.

[36] Stars-of-orion, "Stock movement prediction with sentiment analysis," *Github*. Accessed: Mar. 23, 2024. [Online]. Available: https://github.com/stars-of-orion/stock-movement-prediction-with-sentiment-analysis

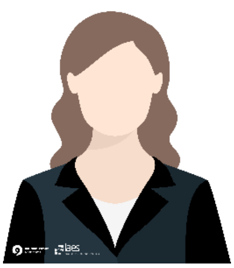
BIOGRAPHIES OF AUTHORS



Salma Elsayed holds a bachelor's degree in computers and information from Zagazig University, Egypt, 2016. She worked as a teaching assistant in the Department of Computer Science at the Faculty of Computers and Information, 6 October University, Egypt for 2 years. Her research interests include artificial intelligence, machine learning, deep learning, ensemble learning, and sentiment analysis. She can be contacted at email: salmaelsaied492@gmail.com.

Ahmad Salah received a Ph.D. degree in computer science from Hunan University, China, in 2014. He received a master's degree in CS from Ain-Shams University, Cairo, Egypt, in 2009. He is currently an associate professor of computer science at Zagazig University, Egypt. He has published more than 60 papers in international peer-reviewed journals, such as the IEEE Transactions on Parallel and Distributed Systems IEEE-ACM Transactions on Computational Biology Bioinformatics, and ACM Transactions on Parallel Computing. His current research interests are parallel computing, computational biology, and machine learning. He can be contacted at ahmad@zu.edu.eg.

Ibrahim Elhenawy received the M.Sc. and Ph.D. degrees in computer science from the State University of New York, USA in 1980 and 1983, respectively. Currently, he is a professor in the Computer Science Department, at Zagazig University. His current research interests are mathematics, networks, artificial intelligence, optimization, digital image processing, and pattern recognition. He can be contacted at henawy2000@yahoo.com.



Marwa Abdellah is a lecturer of computer science at the Computer Science Department, Faculty of Computers and Informatics, Zagazig University. She got her M.Sc. and Ph.D. degrees in computer science from the Faculty of Computers and Informatics, Zagazig University in 2013 and 2020, respectively. Her area of interest includes artificial intelligence and machine learning. She can be contacted at marwaabdella2@gmail.com.