5-2023

Learning-based stock trending prediction by incorporating technical indicators and social media sentiment

Zhaoxia WANG
Singapore Management University, zxwang@smu.edu.sg

Zhenda HU
Shanghai University of Finance and Economics

Fang LI
Nanyang Technological University

Seng-Beng HO
Agency for Science, Technology and Research

Erik CAMBRIA
Nanyang Technological University

Learning-Based Stock Trending Prediction by Incorporating Technical Indicators and Social Media Sentiment

Zhaoxia Wang[1*], Zhenda Hu[2], Fang Li[3], Seng-Beng Ho[4] and Erik Cambria[3]

[1]School of Computing and Information Systems, Singapore Management University, 80 Stamford Road, Singapore, 178902, Singapore.
[2]School of Information Management and Engineering, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai, 200433, China.
[3]School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore, 639798, Singapore.
[4]Social and Cognitive Computing Department, Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, Singapore, 138632, Singapore.

*Corresponding author(s). E-mail(s): zxwang@smu.edu.sg;
Contributing authors: huzhenda2020@gmail.com; asfli@ntu.edu.sg; hosb@ihpc.a-star.edu.sg; cambria@ntu.edu.sg;

Abstract

Background/Introduction: Stock trending prediction is a challenging task due to its dynamic and nonlinear characteristics. With the development of social platform and artificial intelligence (AI), incorporating timely news and social media information into stock trending models becomes possible. However, most of the existing works focus on classification or regression problems when predicting stock market trending without fully considering the effects of different influence factors in different phases.

Method: To address this gap, this research solves stock trending prediction problem utilizing both technical indicators and sentiments of the social media text as influence factors in different situations. A 3-phase hybrid model is proposed where daily sentiment values and technical indicators are considered when predicting the trends of the stocks. The proposed method leverages both traditional learning and deep learning methods as the core predictors in different phases. Accuracy and F1-score are used to evaluate the performance of the proposed method. **Results**: Incorporating the technical indicators and social media sentiments, the performance of the proposed method with different learning-based methods as core predictors are analyzed and compared in different situations. Specifically, Multi-Layer Perceptron (MLP), Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) are leveraged as the core learning predictor module, with different combinations of the degree of involvement of technical and sentiment information. The result demonstrates the effectiveness of the proposed method with an accuracy of 73.41% and F1-score of 84.19%. The result also shows that various learning-based methods perform differently for the prediction of different stocks. **Conclusion**: This research not only demonstrates the merits of the proposed method, it also shows that integrating social opinions with technical indicators is a right direction for enhancing the performance of learning-based stock market trending analysis methods.

Keywords: Stock Market Trending, Social Media Sentiment Analysis, Machine Learning, Deep Learning, Technical Indicators

1 Introduction

In recent years, stock market trending analysis has become one of the more popular research areas due to the high returns of the stock market. Stock market time series has been characterized as dynamic and largely non-linear, and stock price prediction is a challenging task [1–3]. Given the dynamic nature of the stock market, the relationship between market parameters and target price is not linear. This results in many economists' belief that stock market prediction does not seem to be viable, and this is being explained by the Efficient Market Hypothesis (EMH) and Random Walk Theory (RWT) [1, 2, 4]. EMH states that the price of a security reflects all information available and everyone has access to the information. As for RWT, it states that stock market prediction is impossible as prices are determined randomly, and hence outperforming is infeasible.

However, with the advent of modern technologies such as machine learning and Artificial Intelligence (AI), more and more researchers have started venturing into the possibilities of using AI technologies such as Machine Learning

and Deep Learning in stock market trending analysis and prediction. As early as in the 1990s, Varfis et al. [5] had tried to apply artificial neural network to financial time series tasks [31]. In addition, researchers are constantly improving the prediction models in the attempt to further enhance the performance of stock market predictions. More and more different machine learning and deep learning methods such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Long Short-term Memory Networks (LSTM) and their fusion models have been applied to stock market predictions [6–10].

Inspired by behavioral finance, researchers began to add information that can reflect investors' behavior toward the stock forecasting model. Bollen et al. [1] used an emotion tracking tool to analyze the content of tweets and used the generated emotion time series to predict the change rate of the Dow Jones Industrial Index. After that, many researchers began to use the tools that can reflect or influence the market to study the stock market based on the emotional and psychological information of participants. Furthermore, with the rise of social networks, huge amount of data is being generated every day, and there is a gaining in popularity of using these data to enhance the prediction performance [11–20].

Unlike most existing research focusing on simple classification or regression tasks for the stock trending problem, this paper not only solves the stock trending problem as a multi-label classification task to predict the trending of the stock price, but also utilizes both technical indicators and sentiment values from social platform as influence factors. Related work of this direction mostly focuses on using only technical indicators [21–23] or sentiment values [24], while this paper leverages both of them for the stock trending prediction task. Furthermore, compared with the existing work [25] which also leveraged both technical indicators and sentiment values, our paper explores the effectiveness of various learning-based algorithms when applied to different stocks, which illustrates the differences of various learning-based algorithms for real-world stock market trending.

In this research work, we propose a hybrid learning-based model to predict the stock's trend. The hybrid model is an integration of learning-based algorithms such as ANN with social media technical indicators and sentiment analysis. The results show that the performance can be improved when relevant technical indicators and social sentiment are considered.

The contributions of this paper are summarized as follows:

1. Unlike most existing research focusing on simple classification or regression tasks for the stock trending problem, this paper not only solves the stock trending problem as a multi-label classification task to predict the trending of the stock price, but also utilizes both technical indicators and sentiment values from social platform as influence factors. It is more fine-grained and more suitable for the real stock market analysis.

2. This paper proposes a hybrid learning-based model which utilizes a three-stage method to determine the final trend prediction based on two intermediate predictions. Different learning-based models are leveraged and

compared, with different combinations of usage of different technical and sentiment information.

3. Abundant experiments are conducted on one stock index Dow Jones Industrial Average (DJIA) and the five very famous stocks including Google (GOOG), Amazon (AMZN), Apple (AAPL), eBay (EBAY) and Citigroup (C) using eight learning-based algorithms. The proposed model outperforms the baseline models for predicting stock's trend, which proves the effectiveness of utilizing technical indicators and social sentiment. The results also illustrate the differences of various learning-based algorithms for the prediction of real-world stock market trend.

The rest of the paper is organized as follows: Section 2 discusses related work on stock prediction; Section 3 presents the proposed stock trending prediction methodology; Section 4 describes parameter setting in the experiments and discusses the results of the experiments; finally, Section 5 offers concluding remarks and illustrates future work.

2 Related Work

There are many internal and external factors influencing the stock price in the stock market, and the fluctuation of stock price volatility is not only affected by macro monetary policy, but also affected by macro-economic environment and emergencies. According to the different mechanisms of stock price prediction, the related work is reviewed under two different aspects: Learning-based Method for Stock Forecasting and Stock Trending Analysis with Social Media Sentiment Analysis.

2.1 Learning-based Method for Stock Forecasting

Compared with the traditional algorithm, machine learning algorithm has the capability of processing large amount of data and multi-dimensional data. Due to the better prediction performance, more and more researchers applied machine learning algorithms to stock market trending analysis and prediction.

The stock forecasting problems can be solved as regression problems to predict the values of the special stocks [15]. In recent years, with the development of deep learning technology, many stock forecasting models based on deep learning have been proposed, and promising results have been obtained [26–30]. However, there are more and more research works which treated the stock forecasting problems as trending classification tasks as described below.

As learning-based methods, SVM, ANN and Naïve Bayes (NB) are widely applied in the field of financial forecasting [31–33]. SVM is known to have capacity control of decision function, use of kernel functions and sparsity of solutions [34]. It has been applied to stock market analysis and has been verified to be effective when it is being compared with other algorithms, such as the Random Walk Model (RW), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Elman Backpropagation Neural Networks (EBNN) (Huang et al. [31]). It has been used for stock market daily price

prediction [35, 36] and Producer Price Index (PPI) prediction [35]. Although the feasibility has been proven, the research also pointed out the limitations for solving such a problem as a regression task [35].

Neural network is known to have the capability for pattern recognition [37]. Nacini et al. [32] compared feed forward Multi-Layer Perception (MLP) and Elman recurrent network by leveraging linear regression. Their experiment showed that linear regression was comparatively better in terms of predicting the direction of changes on the next day, whereas MLP displayed a lowest error in predicting the amount of value changed. This implied that neural networks adapted well to the dynamic nature of the stock market by providing the lowest error rate. From the perspective of the relationship between the stock technical indicators and the stock market, Gken et al. [38] used harmony search algorithm and genetic algorithm to select the most relevant technical indicators and applied them to ANN for stock price prediction. The experimental results showed that the mean absolute percentage error of the ANN model based on harmony search and genetic algorithm is 3.38% and 3.36% respectively, which is better than the model using only ANN algorithm.

As for NB based prediction method, it is a type of supervised learning method that learns from historic records or expert's knowledge and utilizes probabilistic approaches to find an optimal solution [39]. Huang et al. [33] utilized a set of independent data which was collected randomly from Taiwan Stock Exchange Corporation (TSEC), and 9 attributes were used to build the NB predictor. Their result showed successful prediction, with a probability of 13.46% of making a loss. This implies the possibility of using the NB based predictor for stock market prediction in obtaining good results.

Besides the traditional machine learning methods mentioned above, there are Ensemble Learning (EL) methods that have been used to forecast future trends of stock price movements [40, 41]. Random Forest (RF) can overcome overfitting problems by training multiple decision trees on different subspaces of the features at the cost of slightly increased bias. Previous experiment indicated that RF resulted in a high accuracy rate for all periods, and the longer the trading period, the higher the accuracy rate [40].

Extreme Gradient boosting (XGboost) was proposed by Chen and Guestrin [41]. It has been proven that XGboost has the characteristics of low computational complexity, fast running speed and high accuracy. For the analysis of time series data, although Gradient Boosting Decision Tree (GBDT) can effectively improve the stock prediction results, the relatively slow detection rate limits the method. In order to find a fast and high accuracy prediction method, XGboost model has been used for stock prediction, which can improve the prediction accuracy as well as the prediction speed.

With the success of learning-based methods for stock trending analysis, some researchers further enhanced the performance of the methods by considering the influence of social event (e.g., social media news sentiment analysis)

2.2 Stock Trending Analysis with Social Media Sentiment Analysis

Social media sentiment analysis is a popular research area in the Natural Language Understanding (NLU) domain that identifies and categorizes opinions that are expressed in news, articles, tweets or text [42–44]. In the field of stock market prediction, it is often used as an indicator of the public sentiment towards events and scenarios. A very popular method is to use the sentiment value as an external factor and feed it as an input that will affect the final prediction [11–14, 14–18, 25, 45].

Bharathi and Geetha [11] aimed to present the impact of Really Simple Syndication (RSS) feeds on stock market values. The approach of this article is to utilize the Sentiment Analysis result as an external factor that is used together with the Sensex-Moving Average results to produce a final-result prediction of the trend. Ichinose and Shimada [12] proposed a system that utilized Bag of Keywords from expert articles (BoK-E) to predict the trend of the next day. In the experiment conducted, it was reported that the average accuracy obtained using BoK-E was 61.8%, which is a 9.5% increase in accuracy compared to using standard Bag of Word approach. Zhang et al. [13] utilized the correlation of events from web news and public sentiments from social media and stock movement to determine the next day trend. The proposed coupled stock correlation (CMT) method (accuracy of 62.50%) performs better compared to models without stock correlation information (accuracy of 60.25%)).

In addition, Si et al. [14] proposed the use of a Semantic Stock Network (SSN) to model the relationship between stocks. It proved that the utilization of SSN has a higher capability than Correlation Stock Network (CSN) to predict the stock market. Xing et al. [46] designed a framework which captured the bi-directional interaction between movements of asset price and market sentiment for stock return fluctuation prediction. Picasso et al. [25] combined both technical and fundamental analysis using machine learning techniques for the stock market prediction problem. A high frequency trading simulation with over 80% of annualized return was conducted to exploit the prediction results. Merello et al. [47] presented a transfer learning method to estimate the amount of price change and the most performing assets, in which price fluctuations of different magnitude are treated differently through the application of different weights on samples.

There is also a gaining in popularity of using Twitter data for Sentiment Analysis [17]. In addition, Li et al. [17] also suggested that the proposed approach of using Twitter data for stock market prediction achieved a better performance when using the tweets sentiment values to predict the stock price of three days later. Gupta and Chen [48] analyzed the StockTwits tweet contents and extracted financial sentiment using a set of text featurization and machine learning algorithms. The correlation between the aggregated daily sentiment and daily stock price movement was then studied, and the effectiveness of the proposed work on stock price prediction was demonstrated through experiments on five companies (Apple, Amazon, General Electric, Microsoft,

and Target). In addition, Google Trends data was used to provide the search volume for keywords searched such that the model could determine the impact of events that might affect the stock market. Hu et al. [18] considered the use of Google Trends data in improving the performance of stock market prediction. According to the experimental results, Google Trends was capable of enhancing the accuracy in predicting the trend of the stock market.

Besides public sentiment, Khan et al. [49] also explored the effect of political situation on the stock prediction accuracy, and the experimental results showed that the sentiment feature improved the prediction accuracy of machine learning algorithms by 0-3% while political situation feature improved the prediction accuracy of algorithms by about 20%.

Unlike the work mentioned above, this paper aims to solve the stock trending problem as a triple classification task to predict the trending of the stock price, e.g., Buy or Rise (1), Sell or Drop (-1) and Hold (0). In addition, related work of this direction mostly focuses on using only technical indicators [21–23] or sentiment values [24], while this paper leverages both of them for the stock trending prediction task. Our paper explores the effectiveness of various learning-based algorithms when applied to different stocks, which illustrates the differences of various learning-based algorithms for real-world stock market trending. Not only the stock index, Dow Jones Industrial Average (DJIA) is analyzed, but also the individual stocks, such as the very famous stocks, Google (GOOG), Amazon (AMZN), Apple (AAPL), eBay (EBAY) and Citigroup (C) are analyzed by using different enhancement technologies. Our research work demonstrates the merits of the proposed method and points out the correct direction for future work in this area.

3 The Proposed Methodology

This research aims to leverage the stock market time series data to investigate the performance of different learning-based methods by incorporating technical indicators and social media sentiment analysis.

The proposed methodology consists of 3 phases, each with multiple steps. These are described in the following sub-sections.

3.1 Problem Formulation

In this research, the stock market forecasting problem is treated as a three-class classification task. It is to predict the stock market trend: Up, Hold and Down (i.e., 1, 0, -1). Specifically, given the stock prices sequences $X_T(T = 1, 2, 3, \cdots, n)$, generated technical indicator features $TI_T(T = 1, 2, 3, \cdots, n)$ and sentiment features $S_T(T = 1, 2, 3, \cdots, n)$, the task is to predict stock price trend of the next day $Y_T + 1$. It can be formulated as equation (1):

$$\begin{aligned} Y_{T+1} = F(&X_T, X_{T-1}, X_{T-2}, \cdots, X_{T-K}, TI_T, TI_{T-1}, \\ &TI_{T-2}, \cdots, TI_{T-K}, S_T, S_{T-1}, S_{T-2}, \cdots, S_{T-K}) \end{aligned} \tag{1}$$

where $F()$ represents the mapping function from input to output; K represents the size of the sliding window.

3.2 Stock Data Pre-processing

The stock index and various stocks from S&P 500 were identified and retrieved from Yahoo Finance (https://sg.finance.yahoo.com). The period for data extraction was between 1st Jan 2014 to 31th Dec 2018. Entries in the data include:

-Date: Index of each record
-Open: Price of stock at opening of trading (in USD)
-High: Highest price of stock during trading day (in USD)
-Low: Lowest price of stock during trading day (in USD)
-Close: Price of stock at closing of trading (in USD)
-Volume: Amount of stocks traded (in USD)
-Adjusted Close: Price of stock at closing adjusted with dividends (in USD)

Learning-based methods can be leveraged to analyze all the time series datasets, such as Open, High. Low, Close and Adjusted Close for the stock market data. In this paper, we illustrate the results of analyzing Adjusted Close time series for the purpose of comparing different prediction methods.

All available stock market data were downloaded for analysis, which were daily data. The trending is grouped under Buy or Rise (1) when the percentage change is above +1% and Sell or Drop (-1) when the percentage change is below -1%, else it would be grouped under Hold (0). The various learning-based methods were performed on different stocks and the results were compared.

Five stocks from the S&P 500 index were selected for performing the experiment. They were namely GOOG, AMZN, AAPL, EBAY, EBAY and C.

3.3 Stock Trending Analysis by Incorporating Technical Indicators and Social Media Sentiment Analysis

The proposed stock trending analysis by incorporating technical indicators and social media sentiment is presented and explained in detail in this section. The proposed methodology consists of 3 phases, each with multiple steps. The details of the methodology are illustrated in Fig. 1.

Phase 1: In phase 1, the 1st Intermediate Prediction is obtained using learning-based algorithms with technical indicators. The steps in this phase are as follows:

-Step 1: The model first retrieves the data either manually or automatically by using a crawler that is coded using Python.

-Step 2: The dataset then undergoes pre-processing to ensure the dataset is ready to be fed into the learning-based algorithms. In addition, technical indicators would be considered, and they can be added as part of the input dimensions. The technical indicators can be calculated using the Python library TA with the stock's Open, High, Low, Close and Volume values as inputs to the TA library.

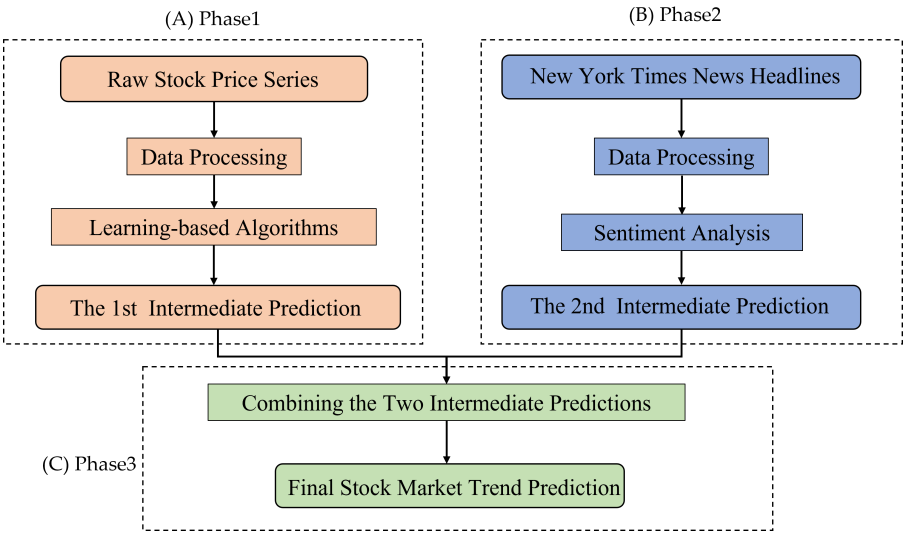

Fig. 1 Layout of the proposed method. (A) Phase 1, the first intermediate prediction, (B) Phase 2, the second intermediate prediction, (C) Phase 3, the final stock market trending prediction

-Step 3: The dataset is fed into the model as inputs and the learning-based algorithms are used to perform intermediate prediction of the trend of the next day.

Phase 2: In phase 2, the model generates a 2nd Intermediate Prediction. This intermediate prediction is the news headline. The steps in this phase are as follows:

-Step 1: In the first step, it retrieves news items that are related to the stocks from online media sources such as the New York Times. The duplicate rows and redundant information within the news are removed.

-Step 2: In this pre-processing stage, duplicated news is first removed and redundant punctuations, special characters and short words (less than 2 characters long) are then removed. Next, the New York Times News undergoes tokenization, stemming and lastly, joining the stemmed tokens back to form a stemmed sentence.

-Step 3: The pre-processed dataset then undergoes sentiment analysis to determine the daily sentiment value (polarity scores). Such scores (compound score) are then calculated using SenticNet [50], a cognitive-inspired framework for sentiment analysis. To derive the daily sentiment value for the News, the compound score (normalized, weighted composite score) of each news items within the same day is summed up and divided by the total number of news items generated on the same specific day.

Phase 3: In phase 3, the two intermediate predictions (Trend Intermediate Predictions and Daily Sentiment Values) are combined to determine the final trend prediction of the next day. The steps in this phase are as follows:

-Step 1: Once the two intermediate prediction values have been obtained, a sliding window of 3 days is applied to the two intermediate prediction datasets (Trend Intermediate Prediction and Daily Sentiment Value). The two datasets are then joined together to form a final dataset with their dates included.

In addition, the daily sentiment value of each day in the sliding window will be further pre-processed such that the impact of the Daily Sentiment Value will decrease as the days go by. The weighted daily sentiment value of Day_{t-x} on Day_t can be calculated using the following equation:

$$WeightedValue_{t-x} = \frac{w-x}{w} * Value_{t-x} \tag{2}$$

where w represents the window size.

-Step 2: After the final datasets have been generated, it is now ready to be fed into the learning-based algorithm for prediction. This final trend will then be the final prediction result of the proposed hybrid learning-based model.

4 Experiment Results and Discussion

4.1 Stock Market Data Used

One stock index and five stocks were identified to be used, namely Dow Jones Industrial Average (DJIA), Google (GOOG), Amazon (AMZN), Apple (AAPL), eBay (EBAY) and Citigroup (C). Two types of datasets are required. The first is the historical values of stocks, and the second is the relevant New York Times News headlines.

For the stocks historical values dataset, the daily data were downloaded from Yahoo! Finance. The dataset contains 7 columns, Date, Open, High, Low, Close, Volume, and Adjusted Close. The interval taken was from 1st Jan 2014 to 31st Dec 2018 five years in total.

New York Times News dataset was obtained using the New York Times Archive API. The API also allows the News to be filtered based on the stock's name and the dataset retrieved was of 5 years, from 1st Jan 2014 to 31st Dec 2018.

4.2 Experiment Parameter and Evaluation Measure

A total of eight learning-based models are used for the comparison in this research, including MLP, Decision Tree, NB, RF, Logistic Regression, XGBoost, LSTM and CNN. 80% of the dataset was selected as training set to build the model for the learning-based methods, and the remaining 20% was used as testing set to verify the performance of the learning-based methods.

For hyper-parameter tuning of the models, we used grid search and expert experience to select the parameters. To maintain fairness in the comparison of different learning-based methods, we made sure that each model used the best optimized parameters. For example, the neural network model is a 3-layer MLP model with hidden layer sizes of 30. For DT, the CART algorithm was

used for feature selection. For RF and XGboost, 100 sub models and 1000 sub models were used, respectively. For LR, L2 regularization was selected as the penalty term. For LSTM, the number of hidden layer nodes with RELU activation function was 50. For CNN, 32 filters were used, and the kernel size was 5.

For technical analysis, we used the technical analysis library in Python to generate a total of 58 features through an original stock time series dataset, and then the Recursive Feature Elimination method (RFE) was used for feature selection. Finally, five most important features were selected.

Different window sizes, n, were used for the trending prediction, it means that we use the value of the previous n days to predict the value of the $(n+1)$th day. After experimental exploration, we chose 3 as the window size.

Under data preprocessing, empty or infinite values were replaced with the value 0. In addition, the independent variable (X) was normalized from the actual value to its percentage change to obtain a smaller range of values to reduce variability, as formulated by using the following equation:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

where x_{norm} is the normalized data, X is the original data, and x_{min} and x_{max} are corresponding minimal and maximal of each data dimension.

The dependent variable (Y) would be the trending label based on n days of prediction. Finally, the performances were evaluated based on accuracy rate and $F1$-score for the triple classification task.

4.3 Performance Comparison for Stock Trending Prediction

In this section, results obtained by the proposed approach are briefly discussed and are evaluated against the *Accuracy* and $F1$-score evaluation metrics for the stock tickers: DJIA, GOOG, AMZN, AAPL, EBAY and C. For the comparison experiments, "with Technical Analysis (TA)" denotes only adding generated technical indicators to the learning-based models, similar to the methods used in some research ([21], [22]) mentioned in Section 2. "with Sentiment Analysis (SA)" denotes only adding sentiment values - i.e., New York Times news polarity - to the learning-based models. "with TA&SA" denotes the proposed method which utilizes both technical indicators and sentiment values for the stock prediction task. The results can be seen in Table 1 and Table 2.

Table 1 shows the Accuracy obtained from analyzing the stock index and individual stocks by using eight different learning-based methods. It can be observed that the baseline model with TA&SA achieves the best results in 15 out of 48 cases for all stocks while only 3 and 5 best results are achieved by Baseline and Baseline with TA, respectively. Among the one stock index and five stocks, the baseline model with TA&SA achieves the best results three times, for the stocks 'DJIA', EBAY and C, followed by the baseline model

Table 1 Accuracy of Different Learning-Based Methods for Individual Stock

Stock	Models	Baseline	with TA	with SA	with TA&SA
DJIA	MLP	73.09	64.26	72.69	64.67
	DT	61.04	60.64	72.69	64.08
	NB	66.05	57.43	72.69	72.29
	RF	69.08	70.28	72.69	72.69
	LR	72.69	71.89	72.69	72.69
	XGBoost	67.07	65.86	72.69	73.09
	LSTM	72.69	73.09	72.69	72.69
	CNN	72.69	73.09	72.69	**73.41**
GOOG	MLP	47.01	44.62	51.60	38.89
	DT	43.82	37.85	50.80	38.62
	NB	34.97	43.82	**52.40**	47.20
	RF	49.40	45.02	50.80	46.80
	LR	49.40	47.41	50.80	50.80
	XGBoost	39.44	43.82	51.20	46.80
	LSTM	49.40	49.40	50.00	49.60
	CNN	49.40	50.60	50.40	50.00
AMZN	MLP	48.40	42.80	50.20	39.55
	DT	44.40	38.80	50.20	49.00
	NB	40.84	33.20	50.20	39.49
	RF	46.80	47.60	50.20	48.19
	LR	47.20	48.40	50.20	49.00
	XGBoost	46.40	47.20	50.20	48.19
	LSTM	46.80	46.80	50.20	48.59
	CNN	46.80	46.80	**51.20**	48.59
AAPL	MLP	53.01	43.78	54.84	40.78
	DT	44.98	34.14	53.63	51.06
	NB	45.11	38.15	54.84	54.44
	RF	49.40	53.41	55.24	55.24
	LR	55.02	**56.63**	54.03	54.84
	XGBoost	42.97	50.60	54.44	55.65
	LSTM	55.02	55.02	54.84	54.84
	CNN	55.02	55.02	54.84	55.24
EBEY	MLP	58.13	54.88	60.00	46.13
	DT	43.09	42.28	60.00	46.13
	NB	46.80	49.19	60.00	**61.82**
	RF	56.10	58.54	60.00	59.59
	LR	60.16	59.76	60.00	60.41
	XGBoost	48.78	50.81	60.00	60.41
	LSTM	60.16	60.16	60.00	60.82
	CNN	60.16	60.38	60.00	60.82
C	MLP	53.63	44.76	55.06	53.94
	DT	37.90	39.52	55.06	53.94
	NB	44.95	40.73	55.06	**55.47**
	RF	49.60	54.03	55.06	54.66
	LR	54.84	52.42	55.06	**55.47**
	XGBoost	42.34	47.18	55.06	54.25
	LSTM	54.84	54.84	55.06	53.85
	CNN	54.84	53.63	55.06	54.25

Table 2 *F*1-score of Different Learning-Based Methods for Individual Stock

Stock	Models	Baseline	with TA	with SA	with TA&SA
DJIA	MLP	67.37	64.35	84.19	64.67
	DT	60.08	61.52	84.19	64.08
	NB	66.05	60.63	60.63	73.67
	RF	63.87	66.95	84.19	63.80
	LR	84.19	66.21	84.19	63.11
	XGBoost	63.90	63.60	**84.32**	64.08
	LSTM	61.20	62.97	61.20	62.06
	CNN	61.20	66.06	61.20	62.89
GOOG	MLP	38.54	42.41	41.33	38.89
	DT	43.20	37.87	40.09	38.62
	NB	34.97	42.82	42.82	41.36
	RF	43.51	37.73	40.13	39.06
	LR	38.13	39.09	39.77	38.68
	XGBoost	36.90	40.67	40.78	38.80
	LSTM	32.67	32.67	39.28	39.59
	CNN	32.67	39.66	39.18	**44.27**
AMZN	MLP	39.31	41.85	48.94	39.55
	DT	43.32	37.78	48.94	39.55
	NB	40.84	30.62	30.62	39.49
	RF	41.57	40.96	48.94	39.55
	LR	39.37	35.87	48.94	39.55
	XGBoost	45.18	41.34	48.94	**52.38**
	LSTM	29.84	29.84	48.94	39.78
	CNN	29.84	29.92	48.94	39.78
AAPL	MLP	42.50	43.00	44.91	40.78
	DT	43.29	35.67	40.79	51.06
	NB	45.11	41.06	41.06	47.74
	RF	41.21	47.10	45.60	51.06
	LR	41.37	43.18	54.57	**55.98**
	XGBoost	38.24	44.38	44.68	53.06
	LSTM	39.06	39.06	38.84	39.54
	CNN	39.06	39.06	39.72	39.74
EBEY	MLP	50.07	49.21	45.00	46.13
	DT	44.14	43.70	45.00	46.13
	NB	46.80	49.88	49.88	48.84
	RF	50.30	50.46	75.00	45.39
	LR	**75.13**	59.46	75.00	46.99
	XGBoost	46.17	46.76	75.00	47.05
	LSTM	45.20	45.20	45.00	47.07
	CNN	45.20	48.67	45.00	47.16
C	MLP	41.07	38.12	63.15	53.94
	DT	38.00	39.80	63.15	53.94
	NB	44.95	50.72	50.72	43.23
	RF	40.54	40.58	63.15	40.56
	LR	70.83	57.59	63.15	41.50
	XGBoost	37.70	38.94	**64.34**	40.39
	LSTM	38.84	38.84	39.10	38.85
	CNN	38.84	41.96	39.10	38.75

with SA. In addition, the baseline model with TA&SA manages to achieve the highest accuracy of 61.82% for the stock EBAY in all cases of the five individual stocks. The results show that the proposed approach baseline with TA&SA outperforms the other strategies in most cases.

As for different learning-based methods, NB achieves the best result for the stock GOOG and 'EBAY' while CNN achieves the best result for the stock index 'DJIA' and the stock AMZN. LR achieves the best result for AAPL while NB and XGboost both achieve the best result for the stock C.

Table 2 shows the F1-score obtained for analyzing the stock index and individual stocks by using eight different learning-based methods. Consistent with the results above from the Accuracy measure in Table 1, it can be observed that the baseline model with TA&SA achieves the highest F-score three times, followed by the baseline model with SA. It is worth mentioning that compared to the baseline model, the baseline model with SA can obtain competitive results for the stock index 'DJIA' and the stock 'C'. In addition, the baseline model with SA manages to achieve the highest F1-score of 84.19% for DIJA by five different machine learning algorithms, which shows the effectiveness of sentiment analysis.

As for different learning-based methods, LR achieves the best result for four individual stocks including GOOG and AAPL, EBAY and C, which shows that LR can be a good choice for individual stocks.

In addition, to verify the robustness and stability of the proposed method, we run the models N times (N=5 in this paper) and report the mean value of the performance as well as the variance (e.g., standard deviation) obtained from using different learning-based models. Table 3 shows the accuracies, F1-scores as well as standard deviations obtained from analyzing the stock index, DJIA. It can be seen that the standard deviations are relatively small, which indicates that the proposed method is stable. The results also indicate that the performance of the proposed method implemented using different learning-based models are slightly different but are stable as shown in Table 3. This discovery is consistent with the results obtained from analyzing different stocks in the previous section.

4.4 Result Analysis and Discussions

From the results obtained, it is discovered that the performance varies from stock to stock.

Firstly, by comparing the baseline model and the baseline model with TA, it is observed that the accuracy and $F1$-score of prediction both drop in most cases when utilizing technical indicators. However, there are also some cases where the baseline model with TA improves the accuracy and manages to generate the best accuracy compared to the other 3 approaches. The result implies that utilization of technical indicators has the potential in increasing the accuracy of prediction. However, such technical indicators must be carefully selected through an optimized feature selection algorithm to prevent it from causing the opposite effect of reducing the accuracy.

Table 3 Accuracy and $F1$-score of Different Learning-Based Methods for DJIA

	Models	Baseline	with TA	with SA	with TA&SA
Accuracy	MLP	73.09 (±0.40)	64.26 (±0.21)	72.69 (±0.34)	64.67 (±0.53)
	DT	61.04 (±0.21)	60.64 (±0.23)	72.69 (±0.34)	64.08 (±0.32)
	NB	66.05 (±0.00)	57.43 (±0.00)	72.69 (±0.00)	72.29 (±0.00)
	RF	69.08 (±0.41)	70.28 (±0.43)	72.69 (±0.34)	72.69 (±0.54)
	LR	72.69 (±0.00)	71.89 (±0.00)	72.69 (±0.00)	72.69 (±0.00)
	XGBoost	67.07 (±0.00)	65.86 (±0.00)	72.69 (±0.00)	73.09 (±0.00)
	LSTM	72.69 (±0.40)	73.09 (±0.54)	72.69 (±0.68)	72.69 (±0.75)
	CNN	72.69 (±0.35)	73.09 (±0.42)	72.69 (±0.67)	73.41 (±0.54)
$F1$-score	MLP	67.37 (±0.52)	64.35 (±0.63)	84.19 (±0.67)	64.67 (±0.54)
	DT	60.08 (±0.23)	61.52 (±0.34)	84.19 (±0.67)	64.08 (±0.33)
	NB	66.05 (±0.00)	60.63 (±0.00)	60.63 (±0.00)	73.67 (±0.00)
	RF	63.87 (±0.76)	66.95 (±0.84)	84.19 (±0.77)	63.80 (±0.68)
	LR	84.19 (±0.00)	66.21 (±0.00)	84.19 (±0.00)	63.11 (±0.00)
	XGBoost	63.90 (±0.00)	63.60 (±0.00)	84.32 (±0.00)	64.08 (±0.00)
	LSTM	61.20 (±0.65)	62.97 (±0.53)	61.20 (±0.64)	62.06 (±0.72)
	CNN	61.20 (±0.47)	66.06 (±0.56)	61.20 (±0.67)	62.89 (±0.55)

Secondly, looking at the results obtained using the baseline model and the baseline model with SA, it can be observed that the utilization of daily sentiment values from New York Times News as an external factor (Phase 2 of the proposed model) to the predicted trend is largely capable of increasing the accuracy of stock prediction. However, there are cases where slight reductions of accuracy when utilizing SA are experienced. This can be caused by reasons such as failing to capture negation in News, and an insufficient number of news items considered in the Sentiment Analysis Phase.

Thirdly, comparing the results obtained by machine learning models and deep learning models, it can be found that deep learning algorithm has no obvious advantages for stock trending prediction, which explains that for some relatively simple tasks, traditional machine learning models can also achieve competitive performances.

Lastly, from the observation of the one stock index and five stocks, the proposed baseline model with TA & SA outperforms the other three approaches in most cases. Thus, this implies that the utilization of technical indicators

together with daily sentiment values of New York Times News might have the effect of further increasing the accuracy of stock prediction.

5 Conclusion, Limitations and Future Works

In conclusion, different from EMH and RWT, where both theories emphasize the non-viability of stock market prediction, this research has demonstrated that it is possible to predict the trending of stock market by using the right methods.

The proposed method is a 3-phase hybrid prediction model where daily sentiment values and technical indicators are considered when predicting the trends of the stocks, GOOG, AMZN AAPL, EBAY and C. The 3 phases in the approach are Phase 1: Intermediate prediction using learning-based algorithms to generate the first intermediate trend prediction, Phase 2: Sentiment analysis where daily sentiment values of New York Times News are calculated, and Phase 3: Final prediction where the final trend is predicted by considering sentiment analysis as an external factor to the first intermediate trend prediction.

The performance of the model is evaluated using *Accuracy* and $F1$-score and the results show that the proposed approach managed to achieve the highest accuracy of 73.41% and the highest $F1$-score of 84.19% for DJIA. In addition, the effect of utilizing sentiment analysis and technical indicators was discussed in detail. Also, utilizing technical indicators together with sentiment analysis can be seen to further enhance the prediction performance.

It is observed that no learning-based method is capable of consistently achieving the best performance. This has been demonstrated across eight different learning-based models in this research. This suggests that the applicability of each learning-based method differs from stock to stock. In addition, this research demonstrates the merit of incorporating both technical indicators and sentiment information for stock trending analysis. Further parameter optimization and consideration of other influence factors are our ongoing work.

Declarations

- **Competing interests**
  The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- **Data availability**
  All data generated or analyzed during this study are included in this published article.
- **Authors contributions**
  Zhaoxia Wang: Conceptualization, Methodology, Supervision, Data curation, Software design, Visualization, Writing - original draft, Writing - review & editing. **Zhenda Hu**: Investigation, Formal analysis, Software testing, Visualization, Validation, Writing - review & editing. **Fang LI**:

Investigation, Data curation, Software development, Writing - original draft. **Seng-Beng HO**: Conceptualization, Methodology, Supervision, Writing - original draft, Writing - review & editing. **Erik Cambria**: Data curation, Investigation, Writing - review & editing.

References

[1] Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of computational science **2**(1), 1–8 (2011)

[2] Patel, J., Shah, S., Thakkar, P., Kotecha, K.: Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications **42**(1), 259–268 (2015)

[3] Ma, Y., Mao, R., Lin, Q., Wu, P., Cambria, E.: Multi-source aggregated classification for stock price movement prediction. Information Fusion **91**, 515–528 (2023)

[4] Maini, S.S., Govinda, K.: Stock market prediction using data mining techniques. In: 2017 International Conference on Intelligent Sustainable Systems (ICISS), pp. 654–661 (2017). IEEE

[5] Varfis, A., Versino, C.: Univariate economic time series forecasting by connectionist methods. In: 1990 International Conference on Neural Networks (ICNN), pp. 342–345 (1990). IEEE

[6] Rather, A.M., Agarwal, A., Sastry, V.: Recurrent neural network and a hybrid model for prediction of stock returns. Expert Systems with Applications **42**(6), 3234–3241 (2015)

[7] Hafezi, R., Shahrabi, J., Hadavandi, E.: A bat-neural network multi-agent system (bnnmas) for stock price prediction: Case study of dax stock price. Applied Soft Computing **29**, 196–210 (2015)

[8] Xiong, L., Lu, Y.: Hybrid arima-bpnn model for time series prediction of the chinese stock market. In: 2017 3rd International Conference on Information Management (ICIM), pp. 93–97 (2017). IEEE

[9] Lee, S.W., Um, J.Y.: Stock fluctuation prediction method and server. Google Patents. US Patent 10,185,996 (2019)

[10] Kim, K.-j.: Financial time series forecasting using support vector machines. Neurocomputing **55**(1-2), 307–319 (2003)

[11] Bharathi, S., Geetha, A.: Sentiment analysis for effective stock market prediction. International Journal of Intelligent Engineering and Systems

10(3), 146–154 (2017)

[12] Ichinose, K., Shimada, K.: Stock market prediction using keywords from expert articles. In: International Conference on Soft Computing and Data Mining, pp. 409–417 (2018). Springer

[13] Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B., Philip, S.Y.: Improving stock market prediction via heterogeneous information fusion. Knowledge-Based Systems **143**, 236–247 (2018)

[14] Si, J., Mukherjee, A., Liu, B., Pan, S.J., Li, Q., Li, H.: Exploiting social relations and sentiment for stock prediction. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1139–1145 (2014)

[15] Wang, Z., Ho, S.-B., Lin, Z.: Stock market prediction analysis by incorporating social and news opinion and sentiment. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1375–1380 (2018). IEEE

[16] Nguyen, T.H., Shirai, K., Velcin, J.: Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications **42**(24), 9603–9611 (2015)

[17] Li, B., Chan, K.C., Ou, C., Ruifeng, S.: Discovering public sentiment in social media for predicting stock movement of publicly listed companies. Information Systems **69**, 81–92 (2017)

[18] Hu, H., Tang, L., Zhang, S., Wang, H.: Predicting the direction of stock markets using optimized neural networks with google trends. Neurocomputing **285**, 188–195 (2018)

[19] Hu, Z.: Crude oil price prediction using ceemdan and lstm-attention with news sentiment index. Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles **76**, 28 (2021)

[20] Malandri, L., Xing, F.Z., Orsenigo, C., Vercellis, C., Cambria, E.: Public mood–driven asset allocation: The importance of financial sentiment in portfolio management. Cognitive Computation **10**(6), 1167–1176 (2018)

[21] Parray, I.R., Khurana, S.S., Kumar, M., Altalbe, A.A.: Time series data analysis of stock price movement using machine learning techniques. Soft Computing **24**(21), 16509–16517 (2020)

[22] Dey, P.P., Nahar, N., Hossain, B.: Forecasting stock market trend using machine learning algorithms with technical indicators. International Journal of Information Technology and Computer Science **12**(3), 32–38

(2020)

[23] Agrawal, M., Shukla, P.K., Nair, R., Nayyar, A., Masud, M.: Stock prediction based on technical indicators using deep learning model. Computers, Materials & Continua **70**(1), 287–304 (2022)

[24] Li, Y., Pan, Y.: A novel ensemble deep learning model for stock prediction based on stock prices and news. International Journal of Data Science and Analytics **13**(2), 139–149 (2022)

[25] Picasso, A., Merello, S., Ma, Y., Oneto, L., Cambria, E.: Technical analysis and sentiment embeddings for market trend prediction. Expert Systems with Applications **135**, 60–70 (2019)

[26] Fischer, T., Krauss, C.: Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research **270**(2), 654–669 (2018)

[27] Nelson, D.M., Pereira, A.C., de Oliveira, R.A.: Stock market's price movement prediction with lstm neural networks. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 1419–1426 (2017). IEEE

[28] Stoean, C., Paja, W., Stoean, R., Sandita, A.: Deep architectures for long-term stock price prediction with a heuristic-based strategy for trading simulations. PloS one **14**(10), 0223593 (2019)

[29] Kim, T., Kim, H.Y.: Forecasting stock prices with a feature fusion lstm-cnn model using different representations of the same data. PloS one **14**(2), 0212320 (2019)

[30] Sezer, O.B., Ozbayoglu, A.M.: Financial trading model with stock bar chart image time series with deep convolutional neural networks. Intelligent Automation & Soft Computing **26**(2), 323–334 (2020)

[31] Huang, W., Nakamori, Y., Wang, S.-Y.: Forecasting stock market movement direction with support vector machine. Computers & operations research **32**(10), 2513–2522 (2005)

[32] Naeini, M.P., Taremian, H., Hashemi, H.B.: Stock market value prediction using neural networks. In: 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pp. 132–136 (2010). IEEE

[33] Huang, T.-T., Chang, C.-H.: Intelligent stock selecting via bayesian naive classifiers on the hybrid use of scientific and humane attributes. In: 2008 Eighth International Conference on Intelligent Systems Design and Applications, vol. 1, pp. 617–621 (2008). IEEE

[34] Wang, Z., Jiao, R., Jiang, H.: Emotion recognition using wt-svm in human-computer interaction. Journal of New Media **2**(3), 121 (2020)

[35] Henrique, B.M., Sobreiro, V.A., Kimura, H.: Stock price prediction using support vector regression on daily and up to the minute prices. The Journal of finance and data science **4**(3), 183–201 (2018)

[36] Marković, I., Stojanović, M., Stanković, J., Stanković, M.: Stock market trend prediction using ahp and weighted kernel ls-svm. Soft Computing **21**(18), 5387–5398 (2017)

[37] Anitescu, C., Atroshchenko, E., Alajlan, N., Rabczuk, T.: Artificial neural network methods for the solution of second order boundary value problems. Computers, Materials and Continua **59**(1), 345–359 (2019)

[38] Göçken, M., Özçalıcı, M., Boru, A., Dosdoğru, A.T.: Integrating metaheuristics and artificial neural networks for improved stock price prediction. Expert Systems with Applications **44**, 320–331 (2016)

[39] Zhu, K., Zhang, N., Ying, S., Wang, X.: Within-project and cross-project software defect prediction based on improved transfer naive bayes algorithm. Computers, Materials and Continua **63**(2), 891–910 (2020)

[40] Khaidem, L., Saha, S., Dey, S.R.: Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003 (2016)

[41] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[42] Wang, Z., Chong, C.S., Lan, L., Yang, Y., Ho, S.B., Tong, J.C.: Fine-grained sentiment analysis of social media with emotion sensing. In: 2016 Future Technologies Conference (FTC), pp. 1361–1364 (2016). IEEE

[43] Wang, Z., Ho, S.-B., Cambria, E.: A review of emotion sensing: Categorization models and algorithms. Multimedia Tools and Applications **79**, 35553–35582 (2020)

[44] Xing, F.Z., Cambria, E., Welsch, R.E.: Natural language based financial forecasting: a survey. Artificial Intelligence Review **50**(1), 49–73 (2018)

[45] Hu, Z., Wang, Z., Ho, S.-B., Tan, A.-H.: Stock market trend forecasting based on multiple textual features: A deep learning method. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1002–1007 (2021). IEEE

[46] Xing, F.Z., Cambria, E., Zhang, Y.: Sentiment-aware volatility forecasting. Knowledge-Based Systems **176**, 68–76 (2019)

[47] Merello, S., Ratto, A.P., Oneto, L., Cambria, E.: Ensemble application of transfer learning and sample weighting for stock market prediction. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2019). IEEE

[48] Gupta, R., Chen, M.: Sentiment analysis for stock price prediction. In: 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 213–218 (2020). IEEE

[49] Khan, W., Malik, U., Ghazanfar, M.A., Azam, M.A., Alyoubi, K.H., Alfakeeh, A.S.: Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. Soft Computing, 1–25 (2019)

[50] Cambria, E., Liu, Q., Decherchi, S., Xing, F., , Kwok, K.: SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis. In: LREC, pp. 3829–3839 (2022)