Learning-Based Stock Market Trending Analysis by Incorporating Social Media Sentiment Analysis

Zhaoxia Wang (✉ zxwang@smu.edu.sg)
Singapore Management University https://orcid.org/0000-0001-7674-5488

Zhenda HU
Shanghai University of Finance and Economics

Fang LI
Nanyang Technological University

Seng-Beng HO
Institute of High Performance Computing

Research Article

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

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-181424/v1

Noname manuscript No.
(will be inserted by the editor)

Learning-Based Stock Market Trending Analysis by Incorporating Social Media Sentiment Analysis

Zhaoxia WANG[1]☒ · Zhenda HU[2] · Fang LI[3] · Seng-Beng HO[4]☒

Received: date / Accepted: date

Abstract
Stock market trending analysis is one of the key research topics in financial analysis. Various theories once highlighted the non-viability of stock market prediction. With the advent of machine learning and Artificial Intelligence (AI), more and more efforts have been devoted to this research area, and predicting the stock market has been demonstrated to be possible. Learning-based methods have been popularly studied for stock price prediction. However, due to the dynamic nature of the stock market and its non-linearity, stock market prediction is still one of the most difficult tasks. With the rise of social networks, huge amount of data is being generated every day and there is a gaining in popularity of incorporating these data into prediction model in the effort to enhance the prediction performance. Therefore, this paper explores the possibilities of the viability of learning-based stock market trending prediction by incorporating social media sentiment analysis. Six machine learning methods including Multi-Layer Perception, Support Vector Machine, Naïve Bayes, Random Forest, Logistic Regression and Extreme Gradient Boosting are selected as the baseline model. The result indicates the possibilities of successful stock market trending prediction and the performance of different learning-based methods is discussed. It is discovered that the distribution of the value of stocks may affect the prediction performance of the methods involved. This research not only demonstrates the merits and weaknesses of different learning-based methods, but also points out that incorporating social opinion is a right direction for improving the performance of stock market trending prediction.

Keywords Social Media Sentiment Analysis · Machine Learning · Technical Analysis · Stock Market Trending

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 (Bollen et al. 2011; Patel et al. 2015). 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)(Bollen et al. 2011; Patel et al. 2015; Dutta and Rohit 2017). EMH states that the price of a security reflects all information

☒ Zhaoxia WANG
zxwang@smu.edu.sg

☒ Seng-Beng HO
hosb@ihpc.a-star.edu.sg

[1] School of Computing and Information Systems, Singapore Management University, Singapore

[2] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai

[3] School of Computer Science and Engineering, Nanyang Technological University, Singapore

[4] Social and Cognitive Computing Department, Institute of High Performance Computing(IHPC), Agency for Science, Technology and Research (A*STAR), Singapore

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 researches 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. (1990) 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 (Rather et al. 2015; Hafezi et al. 2015; Li et al. 2017; Lee et al. 2019; Kim 2003).

Inspired by behavioral finance, researchers began to add information that can reflect investors' behavior to the stock forecasting model. Bollen et al. (2011) used the 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 from 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 (Bharathi and Geetha 2017; Ichinose and Shimada 2018; Zhang et al. 2018; Si et al. 2014; Wang et al. 2018; Nguyen et al. 2015; Li et al. 2017; Hu et al. 2018;).

In this research work, we proposed a hybrid Machine Learning model to predict the stock's trend. The hybrid model is an integration of Machine Learning Algorithms such as Artificial Neural Network (ANN) with Social Media Sentiment Analysis and Technical Indicators. The results show that the performance can be improved when relevant social sentiment and Technical Indicators are considered.

The contributions of this paper are summarized as follows: 1. This paper solves the stock trending problem as a typical classification problem to predict the trending of the stock price. 2. This paper proposes a hybrid Machine Learning model integrating Machine Learning Algorithms with Social Media Sentiment Analysis and Technical Indicators. The model utilized a three-stage method to determine the final trend prediction based on intermediate predictions. 3. Abundant experiments were conducted on six stocks data including Dow Jones Industrial Average (DJIA), Google (GOOG), Amazon (AMZN), Apple (AAPL), eBay (EBAY) and Citigroup (C). The proposed model outperformed several baseline models for predicting stock's trend, which proved the effectiveness of relevant social sentiment and Technical Indicators.

The rest of the paper is organized as follows: "Related Work" section discusses the related work about stock prediction and "The Proposed Methodology" presents the proposed stock trending prediction methodology. "Experiment and Discussion" shows parameter setting in the experiments and discusses the results of the experiments. "Conclusion, Limitations and Future Work" section presents the conclusion obtained from the experiments and 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 macroeconomic environment and emergencies. According to the different mechanisms of stock price prediction, the related work is reviewed under two different aspects as follows.

2.1 Stock Forecasting Based on Stock Price

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.

As learning-based methods, Support Vector Machine (SVM), Neural Network and Naïve Bayes (NB) are widely applied in the field of financial forecasting (Huang et al. 2005; Nacini et al. 2010; Huang et al. 2008). Support Vector Machine (SVM) is known to have capacity control of decision function, use of kernel functions and sparsity of solutions (Wang et al. 2020b). 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. 2005). It has been used for stock market daily price prediction (Henrique et al. 2018; Marković et al. 2017) and Producer Price Index (PPI) prediction (Tang et al. 2018). Although the feasibility was proved, the research also pointed out the limitations for solving such problem as a regression task (Henrique et al. 2018). And Neural network is known to have the capability for pattern recognition (Anitescu et al. 2019). Nacini et al. (2010) 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, Göçken et al. (2016) used harmony search algorithm and genetic algorithm to select the most relevant technical indicators and applied them to the artificial neural network for stock price prediction. The experimental results show 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 only using ANN algorithm. As for Naïve Bayes (NB) based prediction method, it's a type of supervised learning method that learns from historic records or expert's knowledge and utilizes probabilistic approaches to find an optimal solution (Zhu et al. 2020). Huang et al. (2008) 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 getting good results.

Besides the traditional machine learning methods mentioned above, there are some Ensemble Learning (EL) methods used to forecast future trends of stock price movements (Khaidem et al. 2016; Chen and Guestrin 2016). 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. The previous experiment indicated that RF resulted in high accuracy rate for all periods, and the longer the trading period, the higher the accuracy rate (Khaidem et al. 2016). XGboost was proposed by Chen and Guestrin (2016). It was proved 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 is used for stock prediction, which can improve the prediction accuracy as well as the prediction speed.

In recent years, with the development of deep learning technology, many stock forecasting models based on deep learning have been proposed. Fischer et al. (2018) studied the application of LSTM in financial market forecasting by using stack LSTM and bidirectional LSTM to forecast the SP 500 index. By comparing with deep network, random forest and logistic regression model, the empirical results show that LSTM has higher forecasting accuracy. Nelson et al. (2017) used LSTM model to analyze five stocks from the Brazilian stock market and compared the forecasting results of LSTM, MLP, random forest and pseudo-random with statistical testing and trading strategies, and proved that the accuracy of LSTM is higher. Stoean et al. (2019) used LSTM and CNN to build stock prediction models respectively and established trading strategies according to the prediction results. Kim et al. (2019) combined LSTM and CNN model to predict stock data from two perspectives of time series and stock image. Image data is also used for stock market forecasting. Sezer and Ozbayoglu (2020) directly used 2-D stock bar chart images to trained a deep Convolutional Neural Network (CNN) for stock trading model and obtained promising results.

2.2 Stock Forecasting 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 (Wang et al. 2016; Wang et al. 2020a). In the field of stock market prediction, it is often used as an indicator of the public sentiment towards events and scenarios. There are several ways to incorporate Sentiment Analysis into stock market prediction. The very popular method is to feed the sentiment value as an input and another method is to use it as an external factor that will affect the final prediction (Bharathi and Geetha 2017; Ichinose and Shimada 2018; Zhang et al. 2018; Si et al. 2014; Wang et al. 2018; Nguyen et al. 2015; Li et al. 2017; Hu et al. 2018;).

Bharathi and Geetha (2017) 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 (2018) 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. (2018) 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 (62.50%) performs better compared to models without stock correlation information (60.25%)).

In addition, Si et al. (2014) proposed the use of a Semantic Stock Network (SSN) to model the relationship between stocks. It proves that the utilization of SSN has a higher capability than Correlation Stock Network (CSN) to predict the stock market. And Nguyen et al. (2015) incorporated aspect-based Sentiment Analysis into SVM for stock market prediction and showed the effectiveness of aspect-based methodology. Based on the experimental results, it was observed that the proposed approach achieved 9.83% better accuracy compared to method that only uses historical prices and is 3.03% better than Human Sentiment methodology.

There is also a gaining in popularity of using Twitter data for Sentiment Analysis (Li et al. 2017). In addition, Li et al. (2017) also suggested that the proposed approach of using Twitter data for stock market prediction achieved a better performance when using Tweets' sentiment values to predict the stock price of three days later. Gupta and Chen (2020) 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 is then studied. And the effectiveness of the proposed work on stock price prediction is demonstrated through experiments on five companies (Apple, Amazon, General Electric, Microsoft, and Target). In addition, Google Trends data is used to provide the search volume for keywords searched such that the model can determine the impact of events that might affect the stock market. Hu et al. (2018) considered the use of Google Trends data in improving the performance of stock market prediction. According to the experimental results, Google Trends is capable of enhancing the accuracy in predicting the trend of the stock market. This paper solved the stock trending prediction problem as a typical regression problem to predict the price of the stocks. Considering that different events may affect public sentiments and emotions differently, the paper proposed a learning-based method which incorporated social and news opinion and sentiment analysis to predict stock price. Besides public sentiment, Khan et al. (2019) also explored the effect of political situation on the stock prediction accuracy. And the experimental results show that the sentiment feature improves the prediction accuracy of machine learning algorithms by 0–3% while political situation feature improves the prediction accuracy of algorithms by about 20%.

Different from the work done above, this paper aims to solve the stock trending problem as a typical classification problem to predict the trending of the stock price, e.g., Buy or Rise (1), Sell or Drop (-1) and Hold (0). The research work and findings of this research not only demonstrate the merits of the proposed method, but also point out the correct direction for future work in this area.

3 The Proposed Methodology

The performance of various learning-based methods has been demonstrated by different researchers using different stock market datasets. This research aims to leverage the same stock market time series data to investigate the performance of different learning-based methods by incorporating social media sentiment analysis.

In this section we propose a new stock market data analysis method to investigate and compare different learning-based methods in a unique way which considers data correlation analysis between different stocks. The proposed methodology consists of 3 phases, each with multiple steps.

3.1 Stock Data Pre-processing

Various stocks from SP 500 were identified and retrieved from Yahoo Finance (https://sg.finance.yahoo.com). The period for data extraction was between 1st Jan 2000 to 26th 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 under Hold (0). The learning-based methods were performed on different stocks and the results were compared.

The 5 stocks from SP 500 index were selected for performing the experiment. They were namely Alphabet Inc Class A (“GOOGL”), Amazon.com Inc. (“AMZN”), Apple Inc. (“AAPL”), Citigroup Inc. (“C”) and “EBAY” Inc. (“EBAY”).

3.2 Stock Trending Analysis by Incorporating Social Media Analysis

The proposed Stock Trending Analysis by Incorporating Social Media Analysis method used to perform stock trending will be presented and explained in detail. 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 Machine Learning Algorithms.

-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 Machine Learning Algorithms. In addition, Technical Indicator would be considered, and it 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 as inputs to the ta library.

-Step 3: The dataset is fed into the model as inputs and the Machine Learning 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 daily sentiment values of the public derived from performing Sentiment Analysis on the News’ headline.

-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 be 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). The sentiment scores (compound score) is then calculated using the vaderSentiment library. 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.

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

In addition, the Daily Sentiment Value of each sliding window day 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:

$$WeightedValue_{t-x} = \frac{w-x}{w} * Value_{t-x} \qquad (1)$$

where w represents the window size.

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

4 Experiment Results and Discussion

4.1 Stock Market Data Used

Six stocks were identified to be used, namely Dow Jones Industrial Average (DJIA), Google (GOOG), Amazon (AMZN), Apple (AAPL), eBay (EBAY) and Citigroup (C). For the six stocks, two type 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 six stocks daily data were downloaded from Yahoo! Finance. The

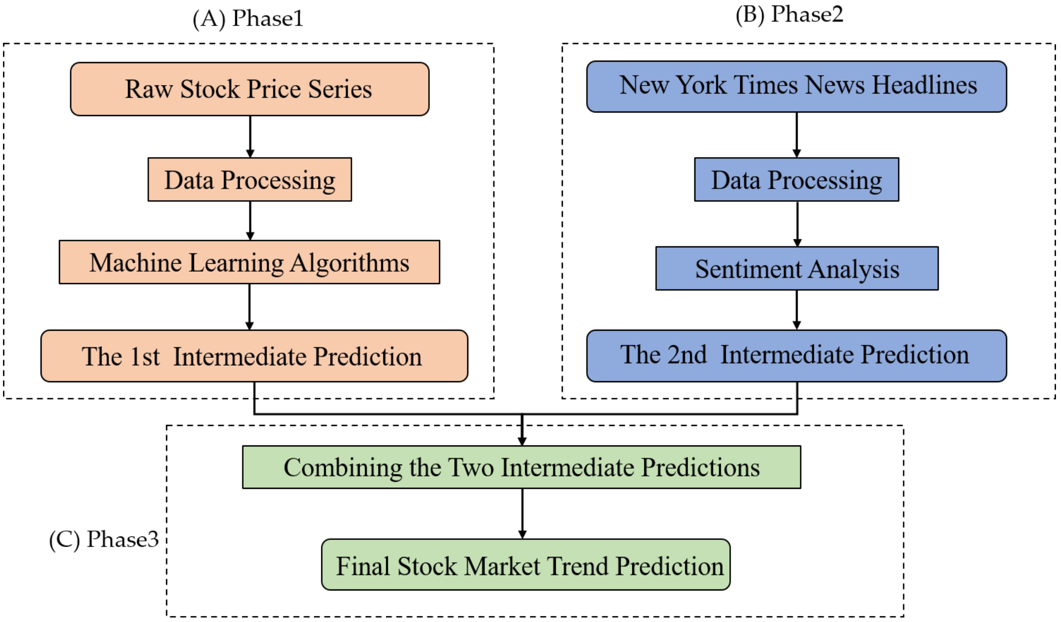
(A) Phase1
Raw Stock Price Series
Data Processing
Machine Learning Algorithms
The 1st Intermediate Prediction
(B) Phase2
New York Times News Headlines
Data Processing
Sentiment Analysis
The 2nd Intermediate Prediction
(C) Phase3
Combining the Two Intermediate Predictions
Final Stock Market Trend Prediction

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

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.

For the New York Times dataset, the 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 Parameter setting

The experiment is designed as a trending prediction problem. A total of 6 learning-based methods are used for comparison in this research. It comprises SVM, neural networks, Naïve Bayes based method, Random Forest, Logistic Regression and XGBoost model. For SVM, we select RBF as the kernel function. For neural networks method, we used a 3-layer MLP model and the hidden layer sizes were all set as 300. For Random Forest and XGboost, 100 sub models and 1000 sub models were chosen, respectively. In addition, 80% of the dataset was selected as training set, which was used to build the model for the learning-based methods, and the remaining 20% was used for testing, which was used to verify the performance of the learning-based methods.

For Technical Analysis, this paper uses 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. The Logistic Regression model was set as the estimator considering the time cost. Finally, five most important features including 'volume_cmf', 'volume_mfi', 'volatility_kcp', 'volatility_dcw' and 'volatility_ui' were selected. 'cmf' means Chaikin Money Flow. Different window sizes, n, were used for the trending prediction. For example, when the window size n is 10, it means that we use the value of the previous 10 days to predict the value of the 11th 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, 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:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

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 performance was evaluated based on accuracy rate and F-score. Accuracy rate measures the total number of correct predictions over total number of predictions. F-score measures the precision and recall rate comprehensively in classification task. Precision measures the number of correct predictions in the total number of positive predictions and Recall measures the number of positive predictions in the total true positive samples. The formula for the balanced F-score, $F1$, is as follows:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

4.3 Comparison of Performance for Individual Stock

In this section, results obtained using the proposed approach are briefly discussed and are evaluated against the Accuracy and F-score evaluation metric for the stock tickers: GOOG, AMZN, AAPL, C and EBAY. Base Line with Technical Analysis means adding generated technical indexes to the original stock time series dataset. Base Line with Sentiment Analysis means adding New York Times news polarity to original stock time series dataset. Base Line with Sentiment Analysis and Technical Analysis means using both of them. The experiments can be seen in Table 1 and Table 2.

Table 1 shows the Accuracy obtained for analyzing the individual stock by using six different learning-based methods. It can be observed that Base Line with Sentiment Analysis and Technical Analysis achieved the best results in 24 out of 30 cases for all stocks while only two best results were achieved both by Base Line and Base Line with Technical Analysis. Compared with Base Line, Base Line with Technical Analysis achieved better results only in 8 out of 30 cases and Base Line with Sentiment Analysis achieved better results in 24 out of 30 cases. And Base Line with Sentiment Analysis and Technical Analysis managed to achieve the highest accuracy of 68.16% for the stock ‘EBAY’ in all cases. The results show that Base Line approach and Base Line with Technical Analysis have the worst prediction accuracy while the proposed approach Base Line with Sentiment Analysis and Technical Analysis outperforms the other approaches in most cases.

As for different learning-based methods, SVM achieved the best result for the stock ‘GOOG’ while four models all achieved the best result for the stock ‘AMZN’. Random Forest achieved the best result for ‘AAPL’ while SVM and LR both achieved the best result for the stocks ‘EBAY’ and ‘C’.

Table 2 shows the F-score obtained for analyzing the individual stocks by using six different learning-based methods. It can be observed that Base Line with Sentiment Analysis and Technical Analysis achieved the best results in 15 out of 30 cases for all stocks while other three methods achieved the best results in no more than 10 cases. Compared with Base Line, Base Line with Technical Analysis achieved better results in 14 out of 30 cases and Base Line with Sentiment Analysis achieved better results in 20 out of 30 cases. And Base Line with Sentiment Analysis and Technical Analysis managed to achieve the highest F-score of 80.87% for the stock ‘C’ in all cases.

As for different learning-based methods, SVM and LR both achieved the best result for four stocks ‘GOOG’, ‘AAPL’, ‘EBAY’ and ‘C’ while SVM achieved the best result for the stock ‘AMZN’, which shows that SVM can be a good choice for the five stocks.

4.4 Discussions

From the results obtained, it is discovered that the performance of the proposed methods varies between stocks.

Firstly, by comparing the Base Line approach and Base Line with Technical Indicator approach, it is observed that the accuracy and F-score of prediction both drop in most case when utilizing Technical Indicators. However, there are also some cases where the Base Line with Technical Indicator approach 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.

Secondly, looking at the results obtained using Base Line approach and Base Line with Sentiment Analysis approach, it can be observed that the utilization of daily sentiment values from New York Time 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 drops of accuracy when utilizing Sentiment Analysis are experienced. This can be caused by reasons such as failing to capture negation in News, and insufficient number News items considered in the Sentiment Analysis Phase.

Lastly, from the observation of the 6 stocks, the proposed approach of Base Line with Sentiment Analysis and Technical Analysis outperforms the 3 other 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.

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

Stock	Learning-Based Methods	Base Line	Base Line with Technical Analysis	Base Line with Sentiment Analysis	Base Line with Sentiment Analysis and Technical Analysis(**proposed**)
GOOG	MLP	56.97	56.18	57.83	**59.44**
	SVM	61.75	61.75	**61.85**	**61.85**
	Naïve Bayes	**58.17**	44.22	57.83	57.83
	Random Forest	52.59	56.97	56.22	**58.23**
	Logistic Regression	61.75	58.57	61.85	**61.85**
	XGBoost	51.79	56.57	57.03	**57.03**
AMZN	MLP	55.20	49.20	**58.23**	51.81
	SVM	56.80	56.80	58.23	**59.84**
	Naïve Bayes	56.00	47.60	58.23	**59.84**
	Random Forest	56.80	57.20	58.23	**59.04**
	Logistic Regression	57.20	56.80	58.23	**59.84**
	XGBoost	50.00	55.20	**58.23**	57.43
AAPL	MLP	58.63	49.00	60.73	**61.94**
	SVM	63.05	63.05	61.54	**63.56**
	Naïve Bayes	58.23	54.62	61.13	**63.56**
	Random Forest	61.85	58.63	59.51	**63.97**
	Logistic Regression	63.05	62.65	63.16	**63.56**
	XGBoost	56.63	**61.85**	59.11	60.32
EBEY	MLP	**65.85**	61.79	60.82	63.67
	SVM	67.48	67.48	67.35	**68.16**
	Naïve Bayes	61.79	61.79	**66.53**	**66.53**
	Random Forest	58.13	63.01	61.63	**64.49**
	Logistic Regression	67.48	65.85	67.35	**68.16**
	XGBoost	58.13	60.57	63.27	**66.12**
C	MLP	63.31	55.24	66.26	**67.07**
	SVM	67.74	67.74	**67.89**	**67.89**
	Naïve Bayes	64.92	57.66	**65.45**	**65.45**
	Random Forest	56.45	**63.71**	59.76	59.76
	Logistic Regression	67.74	66.53	**67.89**	**67.89**
	XGBoost	55.24	55.65	**60.16**	**60.16**

5 Conclusion, Limitations and Future Works

In conclusion, different from EMH and RWT, where both theories emphasize the non-viability of stock market prediction, the research in this paper 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 stocks, GOOG, AMZN AAPL, EBAY and C. The 3 phases in the approach are Phase 1: Intermediate Prediction using Machine Learning Algorithm 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 against Accuracy and results have shown that the proposed approach managed to achieve the highest accuracy of 72.98% and the highest F-score of 84.11% for DJIA. In addition, the effect of utilizing Sentiment Analysis and Technical Indicator was also discussed in detail. Also, utilizing Technical Indicator together with Sentiment Analysis can be seen to further increase the prediction accuracy.

Due to the limited amount of News retrieved, the effect of utilizing Sentiment Analysis may be limited and thus not fully reflected in the results. This is also the limitation of this research work.

It was observed that no learning-based method is capable of consistently achieving the best accuracy across the 6 different approaches. This suggests that the applicability of each learning-based method differs among stocks. In the future, combining different deep learning-based methods, such as LSTM, CNN, and transfer learning methods will be attempted, and the detailed method and discovery will be reported.

Acknowledgements The author would like to thank the FYP students, Jun Hao NG, and Adrian TAN from NTU, and Senior Engineer, Chee Seng CHONG from IHPC, A*STAR for their contributions, support and assistance.

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

Stock	Learning-Based Methods	Base Line	Base Line with Technical Analysis	Base Line with Sentiment Analysis	Base Line with Sentiment Analysis and Technical Analysis(proposed)
GOOG	MLP	46.58	51.18	50.95	**55.17**
	SVM	76.35	76.35	**76.43**	**76.43**
	Naïve Bayes	45.42	39.04	**53.69**	**53.69**
	Random Forest	48.64	**56.88**	51.3	53.69
	Logistic Regression	76.35	59.14	**76.43**	**76.43**
	XGBoost	50.7	**57.08**	55.39	55.39
AMZN	MLP	47.77	48.96	**55.21**	51.61
	SVM	**72.45**	**72.45**	55.21	55.25
	Naïve Bayes	45.58	46.05	55.21	**55.25**
	Random Forest	53.29	**55.41**	55.21	55.15
	Logistic Regression	42.05	46.06	55.21	**55.25**
	XGBoost	48.44	**54.95**	55.21	55.35
AAPL	MLP	52.9	49.78	56.28	**60.62**
	SVM	**77.34**	**77.34**	48.5	61.23
	Naïve Bayes	54.24	55.39	57.17	**61.23**
	Random Forest	58.82	58.15	55.67	**62.31**
	Logistic Regression	**77.34**	49.95	50.01	61.23
	XGBoost	55.45	**62.17**	58.38	59.68
EBEY	MLP	56.14	**61.38**	54.91	60.87
	SVM	**80.58**	**80.58**	80.49	61.49
	Naïve Bayes	56.78	**62.63**	55.84	61.83
	Random Forest	56.21	60.64	58.25	**62.23**
	Logistic Regression	**80.58**	56.14	80.49	62.29
	XGBoost	57.6	60.5	59.44	**63.74**
C	MLP	**57.54**	55.58	54.8	55.22
	SVM	80.77	80.77	**80.87**	**80.87**
	Naïve Bayes	**57.76**	57.73	55.01	55.01
	Random Forest	52.24	58.57	55.61	**55.94**
	Logistic Regression	80.77	57.26	**80.87**	**80.87**
	XGBoost	53.42	54.19	**54.79**	**54.79**

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

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, Data curation, Writing - original draft, Writing - review & editing.

References

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

Bharathi S, Geetha A (2017) Sentiment analysis for effective stock market prediction. International Journal of Intelligent Engineering and Systems 10(3):146–154

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

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

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

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

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

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

Henrique BM, Sobreiro VA, Kimura H (2018) 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

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

Huang TT, Chang CH (2008) 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, IEEE, vol 1, pp 617–621

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

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

Khaidem L, Saha S, Dey SR (2016) Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:160500003

Khan W, Malik U, Ghazanfar MA, Azam MA, Alyoubi KH, Alfakeeh AS (2019) Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. Soft Computing pp 1–25

Kim Kj (2003) Financial time series forecasting using support vector machines. Neurocomputing 55(1-2):307–319

Kim T, Kim HY (2019) Forecasting stock prices with a feature fusion lstm-cnn model using different representations of the same data. PloS one 14(2):e0212320

Lee SW, Um JY (2019) Stock fluctuation prediction method and server. US Patent 10,185,996

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

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

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

Naeini MP, Taremian H, Hashemi HB (2010) Stock market value prediction using neural networks. In: 2010 international conference on computer information systems and industrial management applications (CISIM), IEEE, pp 132–136

Nelson DM, Pereira AC, de Oliveira RA (2017) Stock market's price movement prediction with lstm neural networks. In: 2017 International joint conference on neural networks (IJCNN), IEEE, pp 1419–1426

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

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

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

Sezer OB, Ozbayoglu AM (2020) Financial trading model with stock bar chart image time series with deep convolutional neural networks. Intelligent Automation Soft Computing 26(2):323–334

Si J, Mukherjee A, Liu B, Pan SJ, Li Q, Li H (2014) 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

Stoean C, Paja W, Stoean R, Sandita A (2019) Deep architectures for long-term stock price prediction with a heuristic-based strategy for trading simulations. PloS one 14(10):e0223593

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

Wang Z, Chong CS, Lan L, Yang Y, Ho SB, Tong JC (2016) Fine-grained sentiment analysis of social media with emotion sensing. In: 2016 Future Technologies Conference (FTC), IEEE, pp 1361–1364

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

Wang Z, Ho SB, Cambria E (2020a) A review of emotion sensing: Categorization models and algorithms. Multimedia Tools and Applications pp 1–30

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

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

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

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

Figures

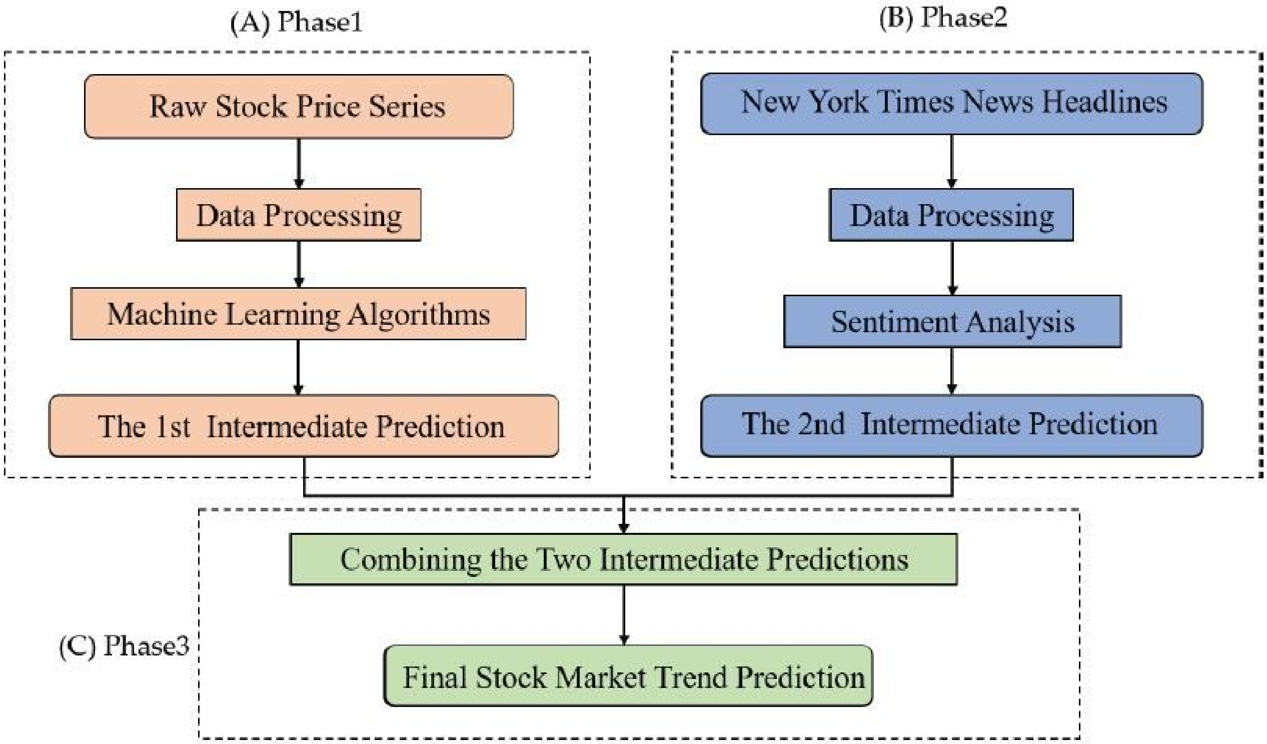
(A) Phase1
Raw Stock Price Series
Data Processing
Machine Learning Algorithms
The 1st Intermediate Prediction
(B) Phase2
New York Times News Headlines
Data Processing
Sentiment Analysis
The 2nd Intermediate Prediction
(C) Phase3
Combining the Two Intermediate Predictions
Final Stock Market Trend Prediction

Figure 1

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