Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 10, 2025

AN ENSEMBLE MODELLING APPROACH FOR STOCK MARKET TREND PREDICTION USING SENTIMENT ANALYSIS

Fatima Akram[1], **Komal Bashir**[*2]

[1]*Lahore College for Women University, Lahore*

[*2]komal.bashir@gmail.com

DOI:https://doi.org/10.5281/zenodo.17365843

Keywords
Deep Learning
Financial Market
Pakistan Stock Exchange
Stock Market
Stock Market Trend Prediction
Stock Market Prediction

Article History
Received: 11 July 2025
Accepted: 21 September 2025
Published: 16 October 2025

Corresponding Author: *
Komal Bashir

Abstract *The swift advancement of the economy prompts individuals to increasingly invest in the stock market; yet, many stay reluctant owing to uncertainties regarding next-day profits. Due to the volatility of stock markets caused by many factors, such as demonetization, next-day predictions are crucial for investors. Stock market prediction mostly relies on a company's historical data, and many studies have utilized deep learning for stock market forecasting, with models demonstrating enhanced performance. Nonetheless, deep learning models require a large set of input variables to enhance performance. As the number of variables increases, the quantity of parameters also escalates, which leads to the issue of overfitting. In this research, LSTM and RF are merged with SA to resolve this issue. Four worldwide stock indices–S&P 500, KOSPI 200, SSE, and KSE 100–are used with 43 technical indicators for the experimental validation of the selected models. The RMSE of the proposed SMTP model for the S&P500, KOSPI200, SSE, and KSE100 are 16.9, 5.7, 44.11, and 33.19, respectively, whereas the forecast accuracies for these indices are 71.21, 95.64, 83.22, and 98.86. This demonstrates a significant level of accuracy, indicating that the model can facilitate educated investment decisions and serve as a valuable resource for investors.*

INTRODUCTION

The thing that fascinates the most at the current time is investing one's capital in the Stock Market (SM). The two main strategies that investors use for deciding on the SM are investing more money and gaining maximum profit with minimum risks [1]. Due to the noisy environment in the SM, its prediction becomes a difficult yet challenging task [2]. Prediction of the stock market is difficult because it is influenced by many factors like market news, quarterly earnings, etc., as the SM fluctuates because of different reasons like war, flood, technological change, dollar price many others. If the price is high, it might be possible that the next day the Financial Market (FM) price will be low, and fluctuations occur frequently.

The SM is a marketplace where investors can purchase and sell ownership in publicly traded companies. These companies raise capital by issuing stocks to investors, and these stocks represent a portion of ownership in the company [3]. SM provides a platform for investors to buy such stocks that are expected to increase in value over time and earn a huge return on their investments [4],[5]. They can also get a return from shares that pay dividends, which are a portion of the company's profits paid to its shareholders [6], [7].

SM plays a crucial role in the country's economy as it helps companies to raise their capital and helps investors with a way to earn maximum profit on their investments. Due to fluctuations in the market,

it results in changes that affect economic conditions, which is the main reason to call it a barometer of the overall health of the country's economy [8].

New people who want to invest in the stock market feel hesitant about their investments. Stock market prediction for the next day with maximum accuracy is the main issue for investors. Financial behavior includes decisions related to investments, which are influenced by human emotions. Accuracy will be increased when sentiment analysis is merged with ensemble modeling.

SM plays a challenging role in the worldwide economy. It also performs several crucial tasks that are listed below.

Raising capital: It helps businesses to generate money, and when these stocks are bought by investors, they get a share in the company's control and future earnings [9].

Providing the liquidity: SM provides investors with a platform where they can buy and sell stocks of publicly traded companies, thereby providing liquidity. Due to liquidity, investors can enter or exit the market more rapidly [10].

Valuation Establishment: For determining the value of a business, SM offers a method that is based upon companies' market demand, its financial performance, and its growth aspects. With the support of valuation companies plan for their future development and attract new investors [11].

Investment facility: SM makes it possible for investors to invest in a variety of businesses across various regions and industries. Diversification made it possible for investors to manage risk and create assets for the long term [12].

Indicator for business: To know about the overall economy of the country, it can be evaluated through SM, as it serves as a standard. An increase in prices of shares shows strong consumer demand and economic expansion, while a decrease in stock prices can indicate a weakened economy of that country [13].

1.1 Factors affecting SM

SM's are affected by a variety of factors, both internal and external [14]. Some of the most important factors include:

Economic growth: When the economy is growing, businesses are doing well, and investors are more likely to buy stocks. This is because they expect businesses to make more profits, which will lead to higher stock prices [2].

Interest rates: Changes in interest rates can also impact the stock market. When interest rates are low, it is cheaper for businesses to borrow money, which can lead to higher profits. This can lead to higher stock prices [15], [16].

Political stability: Political stability is another important factor that can impact the stock market. When there is political instability, investors may be less willing to invest in the stock market. This is because they are worried about the possibility of government intervention or changes in laws and regulations that could harm businesses [17].

Foreign investment: Foreign investment is also an important factor that can impact the stock market. When foreign investors buy stocks, it can lead to higher stock prices. This is because foreign investors bring new money into the market, which can increase demand for stocks [18].

Company news and performance: The news and performance of individual companies can also impact the stock market. When a company has strong earnings or a new product launch, it can cause the stock price to go up. On the other hand, if it has a decline in earnings or a product recall, it can cause the stock price to go down [19].

Global events: Global events, such as wars, natural disasters, and political upheavals, can also impact the stock market. These events can cause investors to become more risk-averse, which can lead to lower stock prices [17].

Financial markets are important to the global economy because they offer investment possibilities,

companies with capital, and a reliable economic indicator.

1.2 Pakistan SM

The Pakistan stock market is a growing market with a lot of potential. The market has been on a bull run in recent years, with the KSE-100 Index, which tracks the performance of the top 100 companies on the Pakistan Stock Exchange (PSX), more than doubling in value since 2016 [20].

The Pakistan stock market is a good investment option for investors who are looking for exposure to the Pakistani economy. The PSX is a liquid, diversified, and growing market, which offers several advantages to investors [21]. However, investors should be aware of the risks involved in investing in the Pakistan stock market, including the risks of political instability, economic uncertainty, and foreign investment volatility.

1.3 Indexes of SM of Pakistan

The Pakistan stock market has several indexes that track the performance of different sectors of the market. Some of the most important indices include:

KSE 100 Index: The KSE 100 Index is the most widely followed index of the Pakistan stock market. It tracks the performance of the top 100 companies listed on the Pakistan Stock Exchange (PSX) [22].

KSE 30 Index: The KSE 30 Index tracks the performance of the top 30 companies listed on the PSX [13]. It is a narrower index than the KSE 100 Index, and is often used by investors who want to track the performance of the largest and most liquid companies on the PSX.

PSX Dividend 20 Index: The PSX Dividend 20 Index tracks the performance of the top 20 companies listed on the PSX that pay dividends. It is a good index for investors who are looking for companies that generate income through dividends [23].

KMI 30 Index: The KMI 30 Index tracks the performance of the top 30 companies listed on the PSX that are considered to be "market leaders" [24]. It is a good index for investors who want to track the performance of the best-performing companies on the PSX.

These are just a few of the many indices that track the performance of the Pakistan stock market. Investors can use these indexes to track the performance of different sectors of the market and to identify companies that are performing well [8].

1.4 SM Trend Prediction (SMTP)

Predicting the trend of SM helps investors to make more informed decisions about selling, buying, or holding the stocks. They can gain insight into future movements by analyzing market trends and adjusting their investment strategies accordingly [25],[26].

Two main approaches for predicting the stock market are fundamental and technical analysis. Investigating quantitative data, such as stock price and portfolio, and qualitative data, such as connected firms' profiles and strategies, can be done as a part of fundamental analysis [3]. In technical analysis, by examining trends in the past and present stocks, the analysts forecast the future of stocks. Such approaches are very helpful in studying market liquidity [3].

1.5 Sentiment Analysis

Sentiment analysis (SA) is a subfield of natural language processing (NLP) that deals with the identification and extraction of opinions and emotions from text [27]. SA is often used to classify text as positive, negative, or neutral [3]. SA is a powerful tool that can be used to understand the opinions and emotions of people. It can also be used to identify specific emotions, such as happiness, sadness, anger, or fear [22].

SA has a wide range of applications, including:

- **Market research:** SA can be used to track customer sentiment towards a product or service [28],[29]. This information can be used to improve the product or service, or to develop new products or services [29],[30].
- **Social media monitoring:** SA can be used to monitor social media for mentions of a company or product. This information can be used to identify potential problems or to gauge customer satisfaction [15].

- **Customer service:** SA can be used to identify customer feedback in customer service tickets [10]. This information can be used to improve the customer service process.

- **Political analysis:** SA can be used to track public opinion on political issues. This information can be used to inform political campaigns or to gauge the public's support for a particular policy [31].

1.6 Existing Work for SMTP

Stock market prediction has been under consideration for a couple of decades. It was observed that many issues arise during the stock market's next-day prediction with maximum accuracy.

This study involves an HFS-based LSTM model to predict SM and uses a dataset from Pakistan. Different metrics involving RMSE, MAE, and MSE were evaluated on companies' datasets like HBL, Netsol Tech, etc. A small quantity of technical stock price characteristics was used. For United Bank, RMSE= 18.52552, MAE= 11.6681, and MSE= 210.6903 [32]. This study was proposed for the Egyptian Exchange to forecast the closing prices of the day by using an equilibrium Optimizer with Support Vector Regression (EO-SVR). It was considered an optimal model due to its superior outcomes. CPU time=1.1231 was calculated [19].

For evaluating the financial markets, soft computing was widely accepted by studying different published articles that focus on neural and neuro-fuzzy techniques. While forecasting input data, performance measures, and predicting methodology were used for classifying. While defining the hidden layers and the structure of the model, difficulties arise [33].

Deep Learning (DL) models were used for intraday prediction. CNN and RNN were focused. Results show that CNN was far better than RNN, as it helps in analyzing the sentiments from texts, and RNN was used for modelling complex temporal aspects and getting the context information. A hybrid model was proposed, named SI-RCNN, in which RCNN was used to predict the intraday directional movements. CNN was used for financial news gathering, and LSTM for technical indicators. Results show an accuracy of above 50% but it does not lead to profitable predictions of the next day [34]. Over the 1980 to 2010 period, twelve emerging countries were examined for the connection between stock market development and exchange rates. It was analyzed in the results that six main economies have significant long-run relationships. Stock market development in China, Pakistan, Mexico, and Venezuela was negatively affected by the volatility of the exchange rates. While the Philippines and South Africa were positively affected [35].

Artificial Neural Network (ANN) and Support Vector Machine (SVM) were used as classification techniques. For the input of the proposed models, ten technical indicators were used to predict the directional movement of the Istanbul Stock Exchange (ISE). From the results, it was analyzed that ANN was significantly better than SVM [36]. For predicting the Brazilian Stock Market, technical indicators were induced into an LSTM to forecast the direction of the stock prices. It was concluded that LSTM outperformed Multi-level Perceptron (MLP) [37].

A Regression and LSTM-based ML were focused on predicting the stock market. For regression, R-square=0.86625 was calculated. For training, MSE=0.00106, RMSE=0.03, and for testing, MSE=0.00875, and RMSE=0.09 were observed. The LSTM model offers more accuracy than regression [38].

Historical data of close prices of the S&P 500 index were used in different models of CNN, LSTM, and MLP. In the results, it was shown that CNN outperformed the other two [39]. For predicting the stock prices, deep learning models such as CNN, LSTM, MLP, and RNN were used. It was observed that deep learning models are better than non-linear models [25],[40].

To measure the predictive performance of DL models, high-frequency data was used. Three traditional artificial neural networks, such as extreme learning machine, radial basis function neural network, and back propagation neural network, were compared with DL models. And it was found that DL models had shown better results [41]. For extracting the qualitative company information, several primary studies were observed that implement text mining techniques on data. The data was utilized for the prediction of the future behavior

of stock prices based on how the news impacts the company in a good or bad way [23].

In the last decade, a study was proposed for stock market prediction in which the architecture of a Generative Adversarial Network (GAN) with MLP was used to classify the input data, and Long Short-Term Memory (LSTM) was used for processing the data. This study was used for forecasting of closing of the stock market. In GAN, MAE=3.0401, RMSE=4.1026, MAPE=0.0137, and AR=0.7554. The GAN model achieved the best result in distributing the real stock data [42]. A study was conducted to improve the forecasting accuracy of the financial expert system neural network (NN), SVM, and decision trees were used. An intelligent decision-making tool was used for effective decision-making on the stocks. Exclusion of the other online data sources was the major drawback of the proposed system [43].

Random Forest (RM) and Artificial Neural Network (ANN) were used to predict the next day's stock price. ANN outperforms RM with RMSE= 0.42, MAPE=0.77, and MBE=0.013 [44]. A CNN-BiLSTM-AM was considered for predicting the next day's stock exchange based on the historical data. Closing, highest, opening, and lowest prices were used as inputs. While processing the input CNN was utilized, and for making the feature extraction, learn BiLSTM was used. It has the most accurate results among the other models [28].

Stock prices of different companies were observed, and many technical indicators were implemented on K Nearest Neighbor (KNN) for forecasting the future movements. Comparing experimental results to machine learning methods, they performed better [45].

Bidirectional Encoder Representation from Transformers (BERT) is used for sentiment analysis of stock market news, and it achieved an accuracy of 87.3% accuracy [46]. Non-linear regression and the K-nearest neighbor method were used to predict stock prices, and they were applied to six sample companies that were registered on the Jordanian stock exchange. According to the results, the prediction results were close to the actual ones, and it shows that KNN is robust with simple error ratio [38].

A system that consists Hidden Markov Model (HMM) was developed to predict the behavior of the Tehran Stock Exchange (TSE) to increase the precision. The dataset was collected for 3 different industries from 2011 to 2014. Industries include Shiraz Petrochemical Company, Jaber Ebne Hayyan Pharmaceutical, and Shargh Cement Company. Results show maximum accuracy 82.37%, F1 measure=79.37% and precision= 78.57% for Jaber Ebne Hayyan Pharmaceutical. While other industries show an accuracy between 69% 82% only [47].

Brazilian stock indices were predicted by using an ANN and the Adaptive Exponential Smoothing Method (ESM). Results show that both models produce similar results in predicting index returns. While ANN outperforms ESM in the forecast of market movement [48]. Random RM and SVM were used in predicting the stock market. Pre-processing of data was done before data analysis. Them ML algorithms were applied to the data. RM improved the accuracy of forecasting [29],[49].

A deep learning technique, the Multi-Channel Convolutional Neural Network (CNN), was proposed. For optimizing CNN parameters Genetic Algorithm (GA) was used. Standard CNN and ANN were compared with the model. ANN achieved an accuracy of 58.62%, the standard CNNs accuracy was 70.16%, and the GA-optimized CNN accuracy was 73.74% [50]. An optimized Deep-ConvLSTM model was proposed, and it was trained by the Rider-MBO algorithm. MSE=7.2487 and RMSE=2.6923 were calculated from the six forms of livestock market data. Hence, it was considered an effective model for forecasting the stock market [30].

SVM and Radial Basis Function (RBF) were used for the stock market prediction. SVM doesn't have a problem with overfitting. A dataset from IBM was used [14]. The Multi-Model Generative Adversarial Network Hybrid Prediction Algorithm (MMGAN-HPA) is used for improved performance of stock market prediction. For the dataset of the TCS model, MAE=0.00263344, MSE=0.00003490, correlation=0.995985974, it gives better performance [51].

An equilibrium Optimizer with Support Vector Regression (EO-SVR) was proposed for the Egyptian Exchange to forecast the closing prices of the day. It

was considered an optimal model due to its superior outcomes. CPU time=1.1231 was calculated [19].

2. Materials and Methods

The methodology consists of several steps, including dataset, data preprocessing, feature extraction, model training, and evaluation. Figure 1 shows the proposed model of SMTP. It consists of multiple phases.

2.1 Dataset Acquisition: Datasets are acquired from different websites that publicly provide financial data, such as Yahoo Finance, Google Finance, Quandl, scs trade, PSX, and Dawn News.

Stock Data

Daily SM data of KSE 100, S&P500, SSE, and KOSPI200 is collected from the official website of PSX, Yahoo Finance, and SCS Trade. The dataset contains date, high, low, open, close, volume, and change.

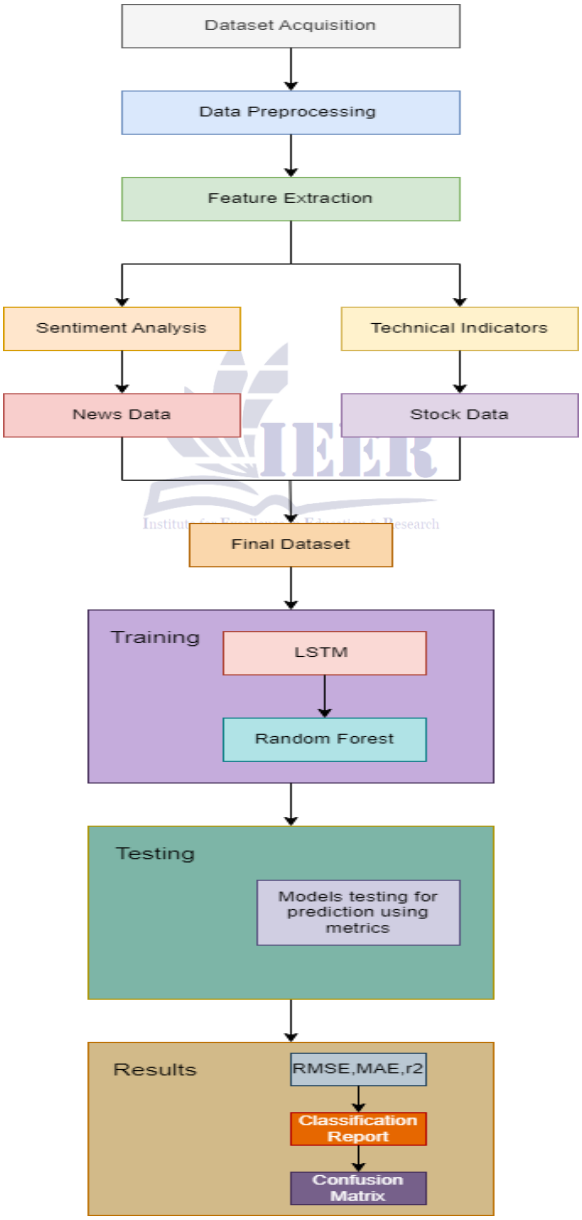

Figure 1: Proposed Mode

The KSE100 10-year daily dataset is collected for the period 1/2/12 to 6/6/23, involving 2733 trading days. KOSPI200: 11 years of daily trends are c ollected for the period 4/3/02 to 5/18/15. 11 years of data of S&P500 for 8/1/02 to 4/23/15, and 10 10-year dataset for SSE 10/13/03 to 7/21/15.The dataset contains 6 columns. Table 1 describes the dataset columns, and Figure 2 gives a screenshot of the dataset.

Table 1: Description of Dataset Features

Features	Description
Date	The date specified for each trading day is in the format of year-month-day
Open	Price related to stock at market open
High	Maximum daily prices
Low	Minimum daily prices
Close	The stock price when the stock market is closed for a specific trading day
Volume	Number of traded shares

	Date	Open	High	Low	Close	Volume
0	6/6/2023	41771.03	42026.94	41751.20	41923.46	122360032
1	6/5/2023	41390.47	41681.68	41375.57	41667.94	81400128
2	6/2/2023	41264.42	41403.43	41237.41	41352.98	67578672
3	6/1/2023	41387.24	41481.53	41226.94	41266.77	49182896
4	5/31/2023	41724.85	41852.23	41265.03	41330.56	90799056

Figure 1: Snapshot of KSE100 Dataset

The datasets used in this research work belong to 4 different countries, SMs as America, China, Korea, and Pakistan. Every dataset contains approximately 3000 entries with 6 columns named as date, open, high, low, close, and volume. News data is collected from Dawn News.

Financial News Articles/Headlines

Financial news is collected from Dawn News to date. API keys are used to extract data from the website for performing sentiment analysis. To make any financial decision, investors deeply look into companies' profiles, news, trading history, etc.

Data Preprocessing: Real-time stock data can involve missing values due to closed weekends. News data includes noisy data like missing data, punctuations, tokenization, etc. The main task in the preprocessing of data is data cleaning, reduction, and transformation.

Datasets must be cleaned during the preprocessing step; during this phase, we apply several cleaning and filtering techniques on the news data, such as removing links, identifiers, and deleting words that contain fewer than 3 characters, and stock data is preprocessed to check missing values and filter empty words. Features are extracted from the datasets.

Tokenization

Tokenization is the step that takes the sentence and breaks it down into words. Using white spaces and punctuation symbols as delimiters, the text is broken down into single words or tokens. We must first define the words that make up a string of characters before processing the natural language. As a result, tokenization is the most fundamental step in the NLP process (text data). This is significant since the text's meaning can be easily deduced by examining the words in the text.

Lower Casing

Lower casing is the step of changing to the lower case of a word in the English language. Words like BOOK and book have the same meaning, but in the vector space model, they are represented as two separate words if they are not transformed to lowercase, which results in more dimensions. Lowercasing is one of the most basic and effective text preprocessing techniques, which also greatly improves desired output consistency.

Stopword, Punctuation Removal, and Stemming

The common words in any natural language are stopwords. These stopwords cannot add much value to the context of the document when interpreting text data and constructing NLP models. Stop words include articles, prepositions, conjunctions, and certain pronouns. Some common words in a document are "the", "is", "in", "about", "where", "at", "to", "", and so on. The concept is to simply eliminate the terms that appear in all of the corpus documents. The punctuation in the text adds very little or no value to the information. When punctuation is added to any word, it becomes difficult to distinguish it from other words. So, in text preprocessing, we also eliminate the punctuation.

2.3 Feature Extraction: In news and stock data detection is done using deep learning is used; feature extraction is the process of extracting important features from the news and stock data that can be used to train a deep learning model. The goal of feature extraction is to represent the data in a way that captures its important characteristics and patterns, making it easier for the model to learn to predict next-day SM**.** The dataset is scaled and split into training and testing. Null values are evaluated and removed shown in Figure 3. The features list is gathered from the dataset in Figure 4.


```
Date      -       0.0
Open      -       0.0
High      -       0.0
Low       -       0.0
Close     -       0.0
Volume    -       0.0
```

Figure 3: Checking Null values

```
FEATURE LIST
['Date', 'High', 'Low', 'Open', 'Close', 'Volume']
```

	Date	High	Low	Open	Close	Volume	Prediction
2727	2012-01-06	11267.01	11075.94	11214.25	11125.35	34596452	11125.35
2728	2012-01-05	11368.79	11143.70	11366.69	11187.88	24557542	11187.88
2729	2012-01-04	11463.51	11334.85	11433.99	11361.97	39976120	11361.97
2730	2012-01-03	11412.97	11253.63	11258.77	11402.04	53124168	11402.04
2731	2012-01-02	11347.66	11193.63	11339.90	11282.01	29301396	11282.01

Figure 4: Feature List

2.4 Sentiment Analysis: SA is performed using Vader, which is a tool that combines rule-based and lexicon methods. It classifies the textual data consisting of emotions into negative, positive, and neutral. It uses techniques to determine the sentiment of a piece of text. The lexicons are a list of words that have been assigned positive, negative, or neutral sentiment scores. The rules are used to take into account the context of the words in the text.

Technical Indicators

TI is implemented on the stock dataset to predict more accurate values. They are mathematical tools that are used to analyze historical price data in an attempt to predict future price movements. By using them, investors can identify trends and patterns in the market, which can help them make better investment decisions. Table 2 contains list of technical indicators.

Table 2: List of Technical Indicators

TI	Significance
O, H, L, C	Nominal daily open, highest, lowest, and closing price
Volume	Daily trading volume
SD	Degree of variation in a trading-price series
UB & LB	Upper and lower Bollinger bands
FI	Force index
FI5	Average force index
C%, V%	Closing price and volume change
NVI	Cumulative negative volume index
SEMV	Average ease of movement
TSI	True Strength Index
MFR	Money-flow ratio
MFI	Money-flow index
UVI & LVI	Upper and lower vortex indicator
KST	Know sure thing
KST9	Average knows a sure thing.
DPO20	Detrend price oscillator
DX	Directional index
ADX7, ADX14	Average directional index

SMA(5,10,20)	Simple moving average
EMA(6,10,14)	Exponential moving average
MACD(6,12)	Moving average convergence divergence
RSI(10,14)	Relative strength index
CCI(20)	Commodity channel index
H–L	True range index (high–low)
H–Cp	True range index (high–previous close)
L–Cp	True range index (low–previous close)
TR	True range
ATR14	Average true range
ROC12	Price rate of change
Williams%R	Williams percentage range
OBV	On-balance volume

2.5 Final Dataset: The final dataset is the combination of stock data and news data, which can be used to train DL models to predict the stock prices. The stock data provides information about the historical prices of stocks, while the news data provides information about current events that may affect stock prices. By combining these two datasets, DL models can learn to predict how stock prices will change in the future. This information can be used by investors to make better investment decisions.

2.6 Training: Models are trained on the dataset provided. The performance of the model depends on the quality of the dataset on which it is trained.

LSTM model

The LSTM model is used for predicting the next day's stock market. Due to recurrent and gate mechanisms, it is efficient to learn long-range temporal patterns. The gate mechanisms in an LSTM model are used to control how much information is remembered and how much is forgotten. The forget gate determines how much of the previous state is forgotten, the input gate determines how much of the current input is incorporated into the state, and the output gate determines how much of the state is output. They can be used to predict both continuous and categorical variables. They are relatively complex to train, but they can be very effective for tasks that require learning long-range dependencies.

2.7 RF model: The RF model is used for predicting the next day's stock market. It prevents overfitting even if many variables are used. In terms of performance, it is also an advantageous model. It can handle a large number of features, able to learn complex relationships between target variables and features. The predictions from each decision tree in an RF model are then averaged to produce a final prediction. This helps to improve the accuracy of the model. They can be used to predict both continuous and categorical variables. They are relatively easy to interpret, which can be helpful for investors who

want to understand why the model made a particular prediction.

2.8 Testing: Models are tested on the dataset provided to ensure that they are working as intended. The dataset should be representative of the data that the model will be used on in the real world. If the model does not perform well on the test dataset, it may not be able to generalize to new data. In this case, the model may need to be retrained or adjusted.
RF and LSTM models are trained to make predictions for the next day's SM. Data is scaled, a feature list is made, and splitting data is split into training and testing 75% by 25%. Feed the data into the trained models to obtain the predicted values.
Models are tested for the dataset; metrics are calculated for evaluating models' performances.

Model testing for prediction
Model's performances are evaluated by using metrics to measure how well they perform on a given task. These metrics can be used to compare different models or to track the performance of a model over time.

Results
Performance is evaluated by using metrics involving RMSE, MAE, MAPE, R2, Precision, recall, support, F1, and a confusion matrix is calculated.

2.8 Algorithm for Proposed Model

Algorithm: Stock Market Prediction using SA

Step 1: Importing Libraries
- pandas, - re, - nltk, - sklearn, - matplotlib.pyplot, - numpy

Step 2. Load the dataset from a CSV file:
- data = pd.read_csv("path/to/dataset.csv")

Step 3. Extracting news and Sentiment Analysis
- extract news= API key for Dawn news
- date= end_date
- dawn_news_sentiment()

Step 4. Preprocessing and feature extraction

```
def preprocess_text(text):
    # Implement your text preprocessing steps here (e.g., remove stopwords, punctuation, etc.)
    return text
def extract_features(texts,data):
    vectorizer = TfidfVectorizer()
    features = vectorizer.fit_transform(texts)
      return features
```

Step 5. Training the LSTM model

```
def build_rnn_model(input_shape, num_classes):
    model = Sequential()

model.add(Embedding(input_dim=input_shape[0], output_dim=input_shape[1]))
    model.add(LSTM(128))
    model.add(Dense(num_classes, activation='softmax'))
    model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
    return model
```

Step 6. Training RF model

```
def build_rf_model
    rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
    rf_model.fit(train x, train y)
    accuracy = accuracy_score(tested values, pred values)
      return model
```

Step 7. Getting actual and predicted values

```
Actual values = [...]
Predicted values = [...]
Actual_values = [preprocess_stock_data_closing prices]
Predicted_values = [values_calculated]
```

Step 8. Test the models

```
test_data = [...]
test_data = [preprocess_stock_data]
test_features = extract_features(test_data)
lstm_predictions = lstm_model.predict(test_features)
rf_predictions = rf_model.predict(test_features)
dt_predictions = dt_model.predict(test_features.toarray())
```

Step 9. Results and Metrics evaluation

```
mse = mean_squared_error(test data, pred data)
   mae = mean_absolute_error(test data, pred data)
      mape = np.mean(np.abs((test_data – pred_data) / test_data)) * 100
```

3. Results and Discussion

The proposed SMTP model is designed by applying SA, LSTM, and RF. After completion of the necessary processing and training of the dataset, all the models are assessed. In this work, all the models are evaluated in various ways by checking their accuracy, confusion matrix, recall, precision, F1-score, and other metrics.

3.1 Performance Evaluation of KSE100

For the KSE100 dataset, the following metrics are calculated:

Accuracy

It is the ratio of the number of correct predictions to the total number of predictions made by the model. A higher accuracy value indicates better performance.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision

It is the ratio of the true positive predictions to the total number of positive predictions made by the model. A higher precision value indicates fewer false positive predictions.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall

It is the ratio of the true positive predictions to the total number of actual positive instances in the dataset. A higher recall value indicates fewer false negative predictions.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

It is the harmonic mean of precision and recall. It provides a balanced measure between precision and recall and is a commonly used metric for imbalanced datasets.

$$\text{F1} - \text{score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Table 3: Proposed Model Performance Metrics for KSE100

Accuracy	Precision	Recall	F1 score
98.86%	70%	78%	95%

Correlation Heatmap

A correlation heat map is a visual representation of the correlation between different variables in a dataset. It uses a color-coded matrix to display the strength and direction of the relationships between pairs of variables. Positive correlations are typically represented by warmer colors (e.g., red), indicating that the variables move together in the same direction. Negative correlations are represented by cooler colors (e.g., blue), indicating that the variables move in opposite directions. The intensity of the colors reflects the magnitude of the correlation, with darker shades representing stronger correlations. Correlation heat maps are useful for identifying patterns, dependencies, and potential relationships between variables in complex data sets. Figure 5 shows Correlation Heatmap of KSE100

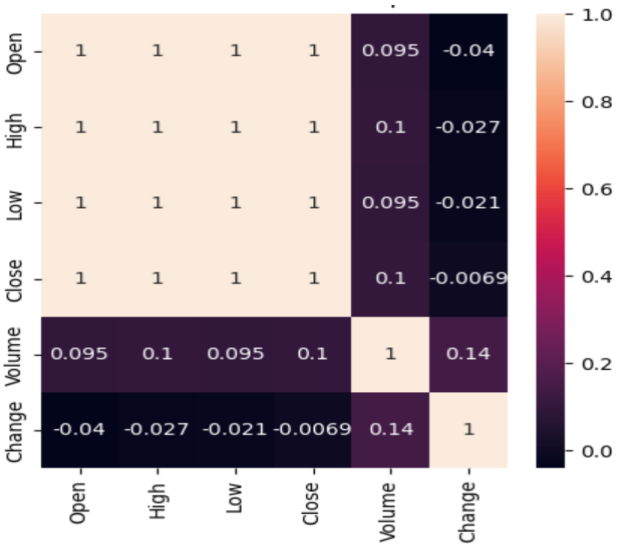

Figure 5: Correlation Heatmap of KSE100

3.2 Predicted values of KSE100

For the KSE100 dataset, predicted values are as presented in Figure 6.

	Actual	Predicted
0	27083.12	26632.890625
1	27115.20	26593.597656
2	27309.02	26662.570312
3	27176.26	26626.458984
4	26892.23	26509.796875

Figure 6: Predicted values

Augmented Dickey Fuller Test

The Augmented Dickey Fuller (ADF) test is a statistical test used to determine whether a time series is stationary or not. A stationary time series is one whose statistical properties do not change over time. Figure 7 shows Dickey-Fuller Test for KSE100.

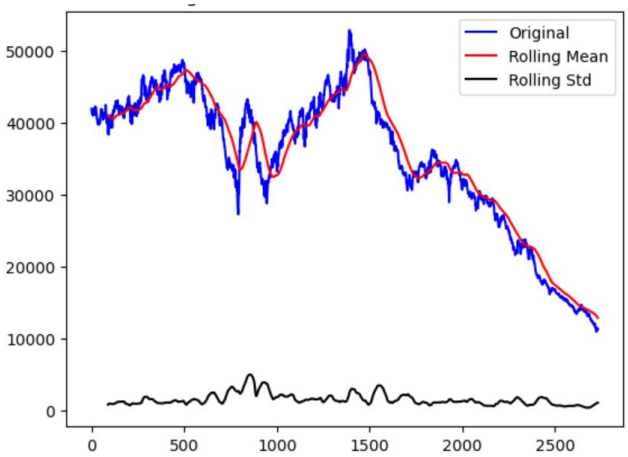

Figure 7: Dickey-Fuller Test for KSE100

3.3 Metrics Evaluation

RMSE

Root Mean Square Error is a measure of the average error between predicted and actual values. It is calculated by taking the square root of the mean squared error. The lower the RMSE, the better the model is at predicting the actual values.

MAE

Mean Absolute Error is another measure of the average error between predicted and actual values. It is calculated by taking the absolute value of the mean difference between the predicted and actual values. The lower the MAE, the better the model is at predicting the actual values.

MAPE

Mean Absolute Percentage Error (MAPE) is a measure of the average error between predicted and actual values, expressed as a percentage of the actual value. It is calculated by taking the absolute value of the mean difference between the predicted and actual values, dividing by the actual value, and multiplying by 100%. The lower the MAPE, the better the model is at predicting the actual values.

R2

R-squared is a measure of the goodness of fit of a model. It is calculated by taking the square of the correlation coefficient between the predicted and actual values. The higher the R2, the better the model is at predicting the actual values.

BACC

Bias-corrected accuracy is a measure of the accuracy of a model. It is calculated by taking the mean of the true positive rate and the true negative rate. The more accurate the model, the higher the BACC will be. All of the above metrics are summarized in Table 4.

Figure 8 gives training and validation loss graph for KSE 100.

Table 1: Metrics Evaluated for KSE100

RMSE	MAE	MAPE	R2	BACC
33.19	1.01	0.75	0.99	0.80

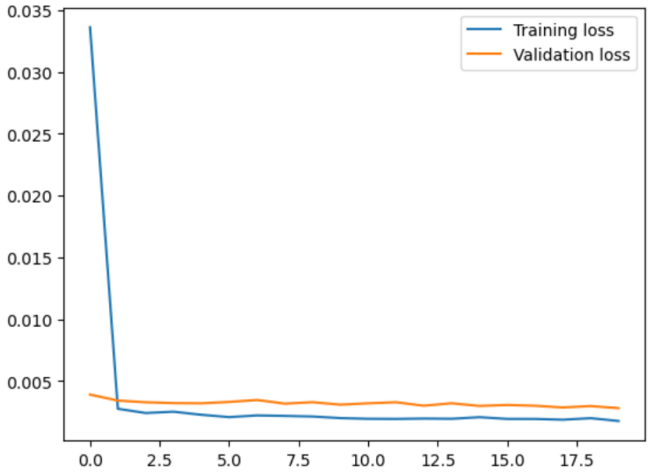

Figure 8: Training and Validation Loss Performance

3.4 Performance Evaluation for KOSPI200

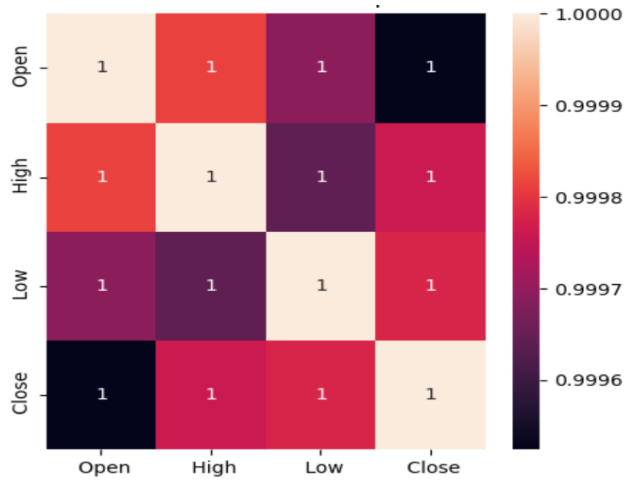

Figure 9: Correlation Heatmap of KOSPI200

	Actual	Predicted
0	0.289172	0.274374
1	0.284125	0.274473
2	0.277029	0.274189
3	0.279748	0.273516
4	0.285017	0.273392

Figure 10: Predicted Values of KOSPI200

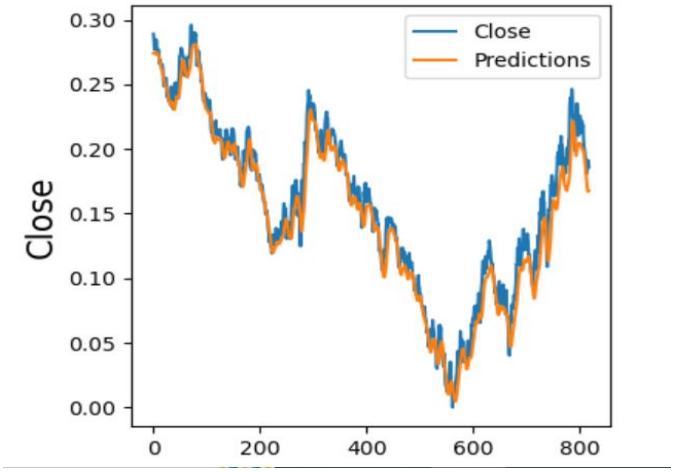

Figure 11: Predicted Graph

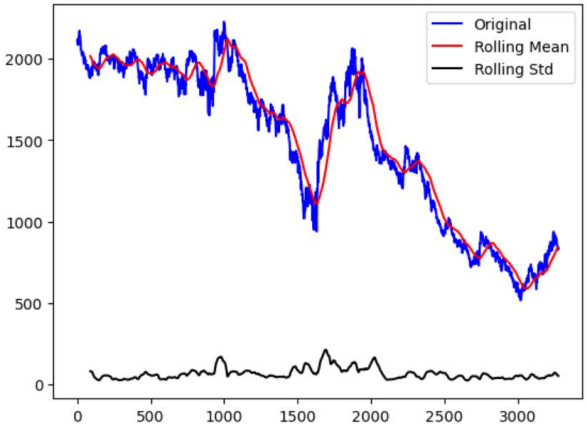

Figure 12: ADF of KOSPI200

3.5 Performance Evaluation for S&P500

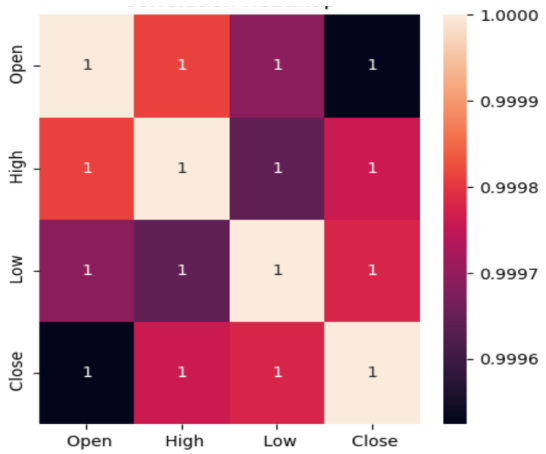

Figure 13: S&P500 Correlation Heatmap

	Actual	Predicted
0	0.653839	0.577459
1	0.587066	0.586963
2	0.559839	0.592315
3	0.507805	0.593701
4	0.541807	0.585600

Figure 14: Predicted Values

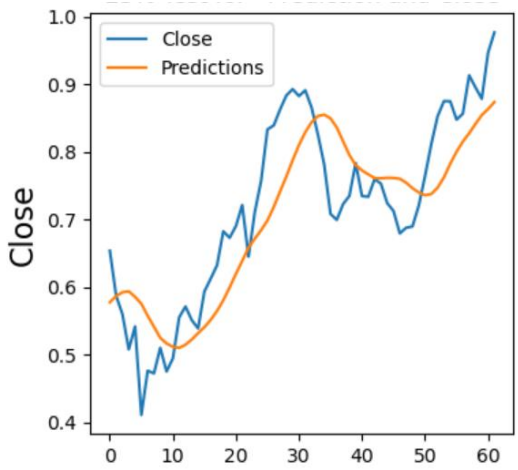

Figure 2: Predicted Values Graph

3.5 Performance Evaluation for SSE

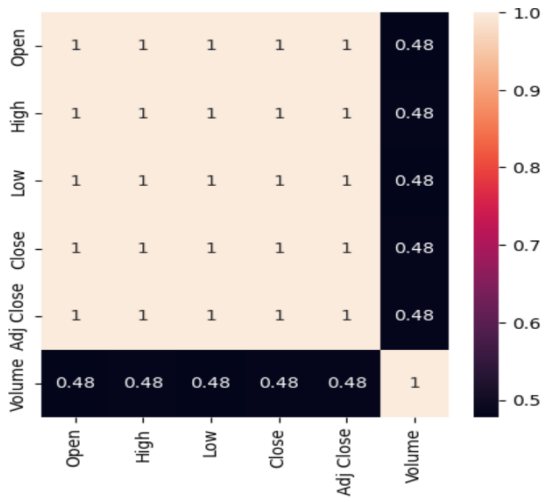

Figure 3: Correlation Heatmap of SSE

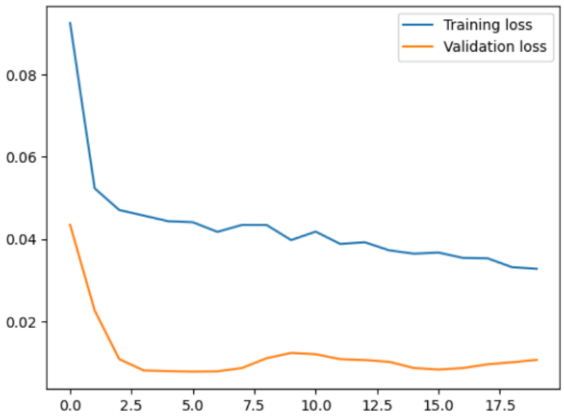

Figure 17: Training and Validation Loss for SSE

3.6 Comparing Performances for all datasets

In Table 4, all datasets performances are calculated and evaluated.

Table 2: Comparing Performances for all datasets

Attributes	S&P500	KOSPI200	KSE100	SSE	Average
RMSE	16.9	5.7	33.19	44.11	24.975
MAE	13.0	3.4	1.01	27.0	11.1025
MAPE	0.55	0.77	0.75	0.78	0.7125
R2	0.99	0.97	0.99	0.96	0.9977
BACC	0.78	0.70	0.80	0.81	0.7725
Precision	0.25	0.60	0.70	0.39	0.485
Recall	0.26	0.65	0.78	0.40	0.5175
F1-score	0.30	0.85	0.95	0.38	0.62
Accuracy	71.21%	95.64%	98.86%	83.22%	87.23%

4. Conclusion

This study presents a research contribution by concentrating on next-day stock market predictions to overcome uncertainty and facilitate informed decision-making. We suggested a model that integrates technical indicators and financial news data, utilizing long short-term memory (LSTM) and random forest (RF) to effectively capture temporal dynamics and manage extensive feature sets without overfitting. The model, trained and assessed on data from the United States, China, Korea, and Pakistan with a distinct training and test set division, attained an accuracy of 98.86% for KSE100, demonstrating its practical applicability for forecasting next-day stock market performance. These findings suggest that the methodology may serve as a significant asset for investors. Future endeavors will broaden the framework to encompass additional markets and timeframes, enhance the incorporation of financial news data, and evaluate practical deployment factors, including transaction costs, risk management, and resilience during market regime transitions.

REFERENCES

[1] D. Kumar, P. K. Sarangi, and R. Verma, "A systematic review of stock market prediction using machine learning and statistical techniques," Mater. Today Proc., vol. 49, pp. 3187–3191, 2020, doi: 10.1016/j.matpr.2020.11.399.

[2] D. Shah, H. Isah, and F. Zulkernine, "Stock market analysis: A review and taxonomy of prediction techniques," Int. J. Financ. Stud., vol. 7, no. 2, 2019, doi: 10.3390/ijfs7020026.

[3] A. Thakkar and K. Chaudhari, "A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions," Expert Syst. Appl., vol. 177, no. Oct. 2020, p. 114800, 2021, doi: 10.1016/j.eswa.2021.114800.

[4] C. R. Kumar and S. Manikandan, "SEMP-TA: A Novel Stock Market Prediction Approach Based on Stacking Ensemble Machine Learning for Effective Trend Analysis," IEEE Access, 2025.

[5] R. Chiong, Z. Fan, Z. Hu, and S. Dhakal, "A novel ensemble learning approach for stock market prediction based on sentiment analysis and the sliding window method," IEEE Trans. Comput. Soc. Syst., vol. 10, no. 5, pp. 2613–2623, 2022.

[6] A. Picasso, S. Merello, Y. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," Expert Syst. Appl., vol. 135, pp. 60–70, 2019

[7] A. Singh, P. Gupta, and N. Thakur, "An empirical research and comprehensive analysis of stock market prediction using machine learning and deep learning techniques," IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012098.

[8] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh, "Stock market prediction using machine learning classifiers and social media, news," J. Ambient Intell. Humaniz. Comput., vol. 13, no. 7, pp. 3433–3456, 2022, doi: 10.1007/s12652-020-01839-w.

[9] W. Mehmood, R. Mohd-Rashid, N. Che-Yahya, and C. Z. Ong, "Determinants of heterogeneity in investors' opinions on IPO valuation: evidence from the Pakistan stock market," Rev. Behav. Financ., vol. 13, no. 5, pp. 631–646, 2021, doi: 10.1108/RBF-04-2020-0078.

[10] R. Ren, D. D. Wu, and D. D. Wu, "Forecasting stock market movement direction using sentiment analysis and support vector machine," IEEE Syst. J., vol. 13, no. 1, pp. 760–770, 2019, doi: 10.1109/JSYST.2018.2794462.

[11] X. Zhang et al., "Improving stock market prediction via heterogeneous information fusion," Knowledge-Based Syst., vol. 143, pp. 236–247, 2018, doi: 10.1016/j.knosys.2017.12.025.

[12] O. Bustos and A. Pomares-Quimbaya, "Stock market movement forecast: A Systematic review," Expert Syst. Appl., vol. 156, p. 113464, 2020, doi: 10.1016/j.eswa.2020.113464.

[13] E. Hoseinzade and S. Haratizadeh, "CNNpred: CNN-based stock market prediction using a diverse set of variables," Expert Syst. Appl., vol. 129, pp. 273–285, 2019, doi: 10.1016/j.eswa.2019.03.029.

[14] A. Subasi, F. Amir, K. Bagedo, A. Shams, and A. Sarirete, "Stock Market Prediction Using Machine Learning," Procedia Comput. Sci., vol. 194, pp. 173–179, 2021, doi: 10.1016/j.procs.2021.10.071.

[15] K. A. Althelaya, S. A. Mohammed, and E. S. M. El-Alfy, "Combining deep learning and multiresolution analysis for stock market forecasting," IEEE Access, vol. 9, pp. 13099–13111, 2021, doi: 10.1109/ACCESS.2021.3051872.

[16] W. T. Ziemba, S. Lleo, and M. Zhitlukhin, "STOCK MARKET CRASHES: Predictable and Unpredictable and What To Do About Them," Stock Mark. Crashes Predict. Unpredictable What To Do About Them, vol. 18, no. 6, pp. 1–308, 2017, doi: 10.1080/14697688.2018.1464792.

[17] H. Moian Nydal, "Stock Market Prediction with Deep Reinforcement Learning," no. Ictck, pp. 11–12, 2015. [Online]. Available: [http://hallvardnydal.github.io/new_posts/2015-07-21-deep_q/] (http://hallvardnydal.github.io/new_posts/2015-07-21-deep_q/)

[18] C. Demir, "Macroeconomic determinants of stock market fluctuations: The case of BIST-100," Economies, vol. 7, no. 1, 2019, doi: 10.3390/economies7010008.

[19] E. H. Houssein, M. Dirar, L. Abualigah, and W. M. Mohamed, An efficient equilibrium optimizer with support vector regression for stock market prediction, vol. 34, no. 4. Springer London, 2022, doi: 10.1007/s00521-021-06580-9.

[20] W. Jiang, "Applications of deep learning in stock market prediction: Recent progress," Expert Syst. Appl., vol. 184, no. Mar. 2020, p. 115537, 2021, doi: 10.1016/j.eswa.2021.115537.

[21] K. Rashid, Y. Bin Tariq, and M. U. Rehman, "Behavioural errors and stock market investment decisions: recent evidence from Pakistan," vol. 7, no. 2, pp. 129–145, 2022, doi: 10.1108/AJAR-07-2020-0065.

[22] N. Rouf et al., "Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions," Electron., vol. 10, no. 21, 2021, doi: 10.3390/electronics10212717.

[23] A. Nikfarjam, E. Emadzadeh, and S. Muthaiyah, "Text mining approaches for stock market prediction," in 2010 2nd Int. Conf. Comput. Autom. Eng. (ICCAE), vol. 4, pp. 256–260, 2010, doi: 10.1109/ICCAE.2010.5451705.

[24] D. Cui and D. Curry, "Prediction in marketing using the support vector machine," Mark. Sci., vol. 24, no. 4, pp. 595–615, 2005, doi: 10.1287/mksc.1050.0123.

[25] M. Hiransha, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "NSE Stock Market Prediction Using Deep-Learning Models," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1351–1362, 2018, doi: 10.1016/j.procs.2018.05.050.

[26] V. Ingle and S. Deshmukh, "Ensemble deep learning framework for stock market data prediction (EDLF-DP)," Glob. Transitions Proc., vol. 2, no. 1, pp. 47–66, 2021, doi: 10.1016/j.gltp.2021.01.008.

[27] W. Lu, J. Li, J. Wang, and L. Qin, "A CNN-BiLSTM-AM method for stock price prediction," Neural Comput. Appl., vol. 33, no. 10, pp. 4741–4753, 2020, doi: 10.1007/s00521-020-05532-z.

[28] M. Leippold, Q. Wang, and W. Zhou, "Machine learning in the Chinese stock market," J. financ. econ., vol. 145, no. 2, pp. 64–82, 2022, doi: 10.1016/j.jfineco.2021.08.017.

[29] A. Subasi, F. Amir, K. Bagedo, A. Shams, and A. Sarirete, "Stock Market Prediction Using Machine Learning," Procedia Comput. Sci., vol. 194, no. 4, pp. 173–179, 2021, doi: 10.1016/j.procs.2021.10.071.

[30] A. Kelotra and P. Pandey, "Stock Market Prediction Using Optimized Deep-ConvLSTM Model," Big Data, vol. 8, no. 1, pp. 5–24, 2020, doi: 10.1089/big.2018.0143.

[31] D. Li and J. Qian, "Text sentiment analysis based on long short-term memory," in 2016 1st IEEE Int. Conf. Comput. Commun. Internet (ICCCI), pp. 471–475, 2016, doi: 10.1109/CCI.2016.7778967.

[32] I. Javid, R. Ghazali, I. Syed, M. Zulqarnain, and N. A. Husaini, "Study on the Pakistan stock market using a new stock crisis prediction method," PLoS One, vol. 17, no. 10, pp. 1–24, Oct. 2022, doi: 10.1371/journal.pone.0275022.

[33] G. S. Atsalakis and K. P. Valavanis, "Surveying stock market foreasting techniques - Part II: Soft computing methods," Expert Syst. Appl., vol. 36, no. 3, pt. 2, pp. 5932–5941, 2009, doi: 10.1016/j.eswa.2008.07.006.

[34] M. R. Vargas, B. S. L. P. De Lima, and A. G. Evsukoff, "07995302(1)," 2017.

[35] M. Hajilee and O. M. Al Nasser, "Exchange rate volatility and stock market development in emerging economies," J. Post Keynes. Econ., vol. 37, no. 1, pp. 163–180, 2014, doi: 10.2753/PKE0160-3477370110.

[36] Y. Kara, M. Acar Boyacioglu, and Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," Expert Syst. Appl., vol. 38, no. 5, pp. 5311–5319, 2011, doi: 10.1016/j.eswa.2010.10.027.

[37] D. M. Q. Nelson, A. C. M. Pereira, and R. A. De Oliveira, "Stock market's price movement prediction with LSTM neural networks," in Proc. Int. Jt. Conf. Neural Networks (IJCNN), 2017, vol. 2017-May, no. Dcc, pp. 1419–1426, doi: 10.1109/IJCNN.2017.7966019.

[38] I. Parmar et al., "Stock Market Prediction Using Machine Learning," in ICSCCC 2018 - 1st Int. Conf. Secur. Cyber Comput. Commun., pp. 574–576, 2018, doi: 10.1109/ICSCCC.2018.8703332.

[39] L. Di Persio and O. Honchar, "Artificial neural networks architectures for stock price prediction: Comparisons and applications," Int. J. Circuits, Syst. Signal Process., vol. 10, pp. 403–413, 2016.

[40] I. K. Nti, A. F. Adekoya, and B. A. Weyori, A systematic review of fundamental and technical analysis of stock market predictions, vol. 53, no. 4. Springer Netherlands, 2020, doi: 10.1007/s10462-019-09754-z.

[41] L. Chen, Z. Qiao, M. Wang, C. Wang, R. Du, and H. E. Stanley, "Which Artificial Intelligence Algorithm Better Predicts the Chinese Stock Market?," IEEE Access, vol. 6, pp. 48625–48633, 2018, doi: 10.1109/ACCESS.2018.2859809.

[42] K. Zhang, G. Zhong, J. Dong, S. Wang, and Y. Wang, "Stock Market Prediction Based on Generative Adversarial Network," Procedia Comput. Sci., vol. 147, pp. 400–406, 2019, doi: 10.1016/j.procs.2019.01.256.

[43] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," Expert Syst. Appl., vol. 79, pp. 153–163, 2017, doi: 10.1016/j.eswa.2017.02.041.

[44] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock Closing Price Prediction using Machine Learning Techniques," Procedia Comput. Sci., vol. 167, pp. 599–606, 2020, doi: 10.1016/j.procs.2020.03.326.

[45] M. Ananthi and K. Vijayakumar, "Stock market analysis using candlestick regression and market trend prediction (CKRM)," J. Ambient Intell. Humaniz. Comput., vol. 12, no. 5, pp. 4819–4826, 2020, doi: 10.1007/s12652-020-01892-5.

[46] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019—2019 Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol., vol. 1, pp. 4171–4186, 2019.

[47] K. Alkhatib, H. Najadat, I. Hmeidi, and M. K. A. Shatnawi, "Stock Price Prediction Using K-Nearest Neighbor Algorithm," Int. J. Business, Humanit. Technol., vol. 3, no. 3, pp. 32–44, 2013. [Online]. Available: https://www.ijbhtnet.com/journals/Vol_3_No_3_March_2013/4.pdf

[48] E. L. de Faria, M. P. Albuquerque, J. L. Gonzalez, J. T. P. Cavalcante, and M. P. Albuquerque, "Predicting the Brazilian stock market through neural networks and adaptive exponential smoothing methods," Expert Syst. Appl., vol. 36, no. 10, pp. 12506–12509, 2009, doi: 10.1016/j.eswa.2009.04.032.

[49] Z. Fathali, Z. Kodia, and L. Ben Said, "Stock Market Prediction of NIFTY 50 Index Applying Machine Learning Techniques," Appl. Artif. Intell., vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2111134.

[50] H. Chung and K.-s. Shin, "Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction," Neural Comput. Appl., vol. 32, no. 12, pp. 7897–7914, 2020, doi: 10.1007/s00521-019-04236-3.

[51] S. R. Polamuri, D. K. Srinivas, and D. A. Krishna Mohan, "Multi-Model Generative Adversarial Network Hybrid Prediction Algorithm (MMGAN-HPA) for stock market prices prediction," J. King Saud Univ. – Comput. Inf. Sci., vol. 34, no. 9, pp. 7433–7444, 2021, doi: 10.1016/j.jksuci.2021.07.001.