Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 1067–1078

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach

Junaid Maqbool[a], Preeti Aggarwal[a]*, Ravreet Kaur[a], Ajay Mittal[a], Ishfaq Ali Ganaie[b]

[a]*Department of Computer Science and Engineering, UIET, Panjab University, Sector 25, Chandigarh, 160014, India,*
[b]*Panjab University, Chandigarh*

Abstract

The stock market is highly volatile as it depends on political, financial, environmental, and various internal and external factors along with historical stock data. Such information is available to people through microblogs and news and predicting stock price merely on historical data is hard. The high volatility emphasizes the importance to check the effect of external factors on the stock market. In this paper, a machine learning model is proposed where the financial news is used along with historical stock price data to predict upcoming stock prices. The paper has used three algorithms to calculate various sentiment scores and used them in different combinations to understand the impact of financial news on stock price as well the impact of each sentiment scoring algorithm. Experiments have been conducted on ten-year historical stock price data as well as financial news of four different companies from different sectors to predict the next day and next week's stock trends and accuracy metrics were checked for a period of 10, 30, and 100 days. The proposed model can achieve the highest accuracy of 0.90 for both trend and future trends for a period of 10 days. Experiments have also been performed to check the difficulty in predicting some stocks. It was found that Tata Motors an automobile company stock prediction has maximum MAPE and hence deviates more from actual prediction as compared to others.

Keywords: Financial news; Forecasting; MLP Regressor; News sentiment analysis; Stock Market; Stock prediction

1. Introduction

STOCK prediction has always been one of the challenging problems of economists, statisticians, and other financial experts as it is a vital component of a country's economy. The stock market is an area where stocks can be traded, transferred, and distributed. The stock market gives companies a chance of expanding and raising money through Initial Public Offerings (IPO) [1]. Investors can invest in stocks of various companies and can make money if they will be able to decide when to buy and when sell particular stocks. The stock market is very volatile as the prices of stocks of particular companies are dynamic and keep changing depending on the volume of shares bought and sold

* Dr Preeti Aggarwal. Tel.: +91-98721021863.
E-mail address: pree_agg@pu.ac.in

10.1016/j.procs.2023.01.086

in the market. The market is influenced by national policies, global and regional economics, as well as psychological and human factors [2], therefore external factors like social media and financial news have either positive or negative effects on stock prices. Traditionally two approaches have been used for stock price prediction viz fundamental analysis, which depends on the research of company's fundamentals like annual growth rates, previous dividend shares provided, market position, new contracts of company, revenue and expenses [3], and technical analysis which focuses on historical price data and use historical price charts to find patterns and make prediction [4]. Technical analysis is based on Dow Theory [5]. Stock price prediction has attracted researchers from various fields like computers science, economics, statistics, and operations research [6].

A large number of studies are currently active on the subject of stock prediction. Data scientists started employing machine learning algorithms to develop stock prediction models. Previous research has employed historical [7,8,9], social media [10,11], or news [12,13] data to predict the stock market using a machine learning algorithm. Machine learning models based on Artificial Neural Network (ANN) [14], Bayesian Network [15], Multi-Level Perceptron (MLP) [16], Support Vector Machines (SVM) [17], and Recurrent Neural Network-based Long Short Term Memory (LSTM) [18] have already been utilized to predict the future trend of stocks as well stock prices. Different stock markets around the world react differently to the period of crisis and other political and financial situations hence cannot be always predicted by simple trading strategies [19].

The motivation behind this study is to develop an intelligent system which will be capable of predicting the future trend of stock market based on the sentiments given in financial news articles. This study will help in limiting the risk to new investors as well also help pro traders to get extra insights about the market movement.

2. State-of-Art

The models which take the technical analysis approach mostly take prediction as a classification problem where market patterns are learned from historical stock price data and the models which work to predict exact stock prices are termed as predictive regression in economic literature [20]. The simple and naive approaches are generally unable to learn long-term trends and suffer from overfitting when applied to real-world problems.

The main aspects of literature are the following:

i. Different machine learning algorithms and different types of data such as historical price data, social media data, political and financial news data for Stock Market Prediction (SMP).
ii. Different sentiment analysis techniques have been used by researchers to classify and score texts.
iii. It is evident that public opinions and emotions as well the regional and global news have a direct effect on stock market price fluctuation.

2.1 Sentiment Analysis For Stock Prediction

The availability of millions of news articles, huge social media textual data, shopping and movie reviews have increased the importance of sentiment analyses significantly. This textual data can be mined to find opinions of individuals in various areas and to achieve the analysis of this huge data, machine learning and data mining are of extreme importance. Therefore, a lot of research is done to find the opinion of users in different fields using sentiment analysis [1]. The sentiment analysis is being explored based on rules, lexicons, and machine learning categorization methods which can classify tweets as positive or negative or neutral based on sentiment towards the field in the study. Support vector machine (SVM), Naive Bayes (NB), and Maximum Entropy (ME) were used among machine learning techniques feature scaling and word count approaches were tried for lexicon-based techniques. Bag of Words (BoW) with N-Gram achieved better performance over Part-Of-Speech linguistic annotations [21, 22]. While evaluating tweets using NB it was found the tweets are very concise and the opinions in tweets are structured and non-uniform, so can be classified into positive, negative, and neutral classes [23]. Tweets and articles regarding movie reviews will not have any effect on the stock market and hence while dealing with any specific problem the selected tweets which are influencing such research should be used. Comparative analysis of NB, ME, and SVM using unigram features and bigram features and a combination of both to classify movie review data from Twitter it was found SVM classifier performs better than other classifiers [24]. On tweets related to technology stocks like Facebook, Google, etc. Logistic Regression (LR) and neural networks with weighing

schemes Term Frequency (TF) and TF Inverse Document Frequency (TF-IDF) were applied and it was found the accuracy of classifiers does not vary much but empirical experiments showed TF-IDF outperform TF [25].
The stock market is influenced by the news as it conveys the events which have an impact on the stock market directly. The news data was also classified into upward, neutral and downward classes by using SVM on stock data and news related to concerned companies, and a direct correlation is found between stock price and news [26]. The stock-related news when auto-categorized and stock-related information is extracted has a direct relationship with the stock behavior [27]. Some sentiment scoring methods which are predefined in important libraries like NLTK [28] and NLP [29] are as:

a. VADER a rule-based general sentiment analysis method uses a combination of qualitative and quantitative methods which validate empirically a list of lexical features to calculate sentiment scores as negative, positive, neutral, and compound.
b. TextBlob predefined in NLP gives sentiment score as polarity and subjectivity where polarity classifies a statement as positive and negative giving a score as a float in the range of [-1,1], and subjectivity gives information regarding the text if it is an opinion, emotion, or factual information [30, 31].
c. FLAIR is an NLP framework that facilitates sequence labeling and text classification. The main function is to provide a unified interface for very different types of works and embedding in the document [23].

2.2 SMP Using Historical Data

The models which operate on the principle of technical analysis to predict stock price take prediction as a classification problem where the market pattern is learned from historical time series data [26]. Before the social media data and financial news data as well algorithms to score textual data were not widely available, researchers applied various machine learning algorithms on the historical stock price for predictions [28]. Machine learning models have shown that Artificial Neural Networks (ANN) can learn input-output relationships and help to make close forecasts on daily closing prices when trained on the same data [25]. A machine learning model with Particle Swarm Optimization (PSO) to optimize Least Squares (LS)-SVM using financial technical indicators like relative strength index, money flow index, etc. for stock prediction was able to overcome an over-fitting problem like in ANN [7]. Since ANN, SVM and other models were still unable to predict the chaotic fluctuations of the stock market because of the absence of memory element hence unable to remember the long-term trend. Long Short-Term Memory (LSTM) based on Recurrent Neural Network (RNN) was used to learn the stock market trend and it was able to predict the opening price while learning long-term patterns and performed better than earlier models [18]. The hybrid model of Empirical Wavelet Transform (EWT) along with LSTM and PSO performed better than other deep learning models and better than any single model used [32]. An ensemble of the Deep Q-learning model was trained to maximize the profit over time without being prone to market ups and downs and being flexible against complex stock fluctuations [19]. Attention-based bi-directional LSTM was used on the Chinese stock exchange market which considered real transaction records and used Convolutional Neural Network (CNN) to extract daily group trading vector and fed into Deep Stock-trend Neural Network (DSPNN) for prediction. The model outperformed LSTM because of the attention mechanism and bi-directional structure of DSPNN [33].

2.3 SMP Using Social Media Data and Financial News

Social media platforms allow people to share their moments, news, and opinions about anything. Financial news, political news, and social media have a direct effect on stock prices. Since all the data is available online and when used along with the historical stock prices and machine learning models like ANN the prediction accuracy and the trend of prediction improved significantly [34]. The LSTM's point to point prediction is closest to actual values and increasing the number of hidden layers didn't make any significant impact on accuracy, after implementing sentiment analysis on tweets it was found tweets affect stock price more when polarizing news about a company float in media sources [35]. Incorporating stock time series data with financial news and using neural networks it was found stock price change has a strong relationship with financial news articles. Time series prediction models like Auto-Regressive Integrated Moving Average (ARIMA), RNN, and Facebook Prophet were used along with financial news data. RNN performed better and a correlation between news textual data and stock price direction was found [36]. Positive correlations exist between stock trends of companies that belong to the same sector and the effect of political news range from 10% - 20% [38]. Combing sentiments of social media and financial news, the highest accuracy decreases but overall accuracies of most classifiers increase. The spam tweet reduction and feature selection have a positive impact on the performance of classifiers. The prediction accuracy of individual classifiers increases when the voting ensemble method on an ensemble of predictions of individual classifiers was used. The highest prediction accuracy of 83.22% was achieved [1]. The data used is general news or social media without giving focus on the particular news about the specific industry or social media activities which are specific to the stock company under consideration.

The remaining part of this paper is organized as follows: Materials and methods used in this study is explained in Section 3. In Section 4, the proposed methodology is described. Section 5 provides the experimental results and discussions and then in section 6, the conclusion, and suggestions for future work in this field are given.

3. Materials and Methods

The motivation for this study is to harness and understand the effect of financial news on the stock price trend and to predict the future stock trend based on the financial news sentiments by training model using financial news sentiments and historical stock data. There are various ways to calculate the sentiments of the text. This paper has explored three types of sentiment scores along with MLP Regressor to predict the change in stock trend and to understand the effect of each sentiment analyzer independently as well in combination with other sentiment scores. The three sentiment scoring methods are Valance Aware Dictionary and Sentiment Reasoner (VADER) [39] from Natural Language Toolkit (NLTK) [21], TextBlob [22], and Flair [23] both from NLP [24]. This paper also uses the parameter *label* which uses keywords to check if any news is about or related to the company whose stock price predictions are under consideration. This paper predicts the *trend* on the upcoming day using the previous day's price and sentiment scores and then comparing the prices of the next day with the previous day, *future trend* by comparing the stock prices of one day with the price after *n* days. The model has been tested on four different companies from varied sectors to evaluate the model and check the validity of the model in a variety of fields. The experiments were conducted on G*oogle Colab*† under default settings with NLP and NLTK libraries.

The major contributions of this research are:

- Proposing an MLP-Regressor model with financial news sentiments and historical stock price data to predict the future trend of stock price and check the accuracy metrics for each case and check the consistency of the model.
- Different sentiment scores are used along with MLP-Regressor to check the effect of each sentiment scoring method on the stock price prediction and using the sentiment scoring algorithms independently as well in combinations and to compare which combination of sentiment scores has more impact on the stock price movement.
- Different stock companies are tested against the mentioned model to check the sectors and companies which are most influenced by financial news and to check the stocks which are easily predicted and the ones which are difficult to predict.

4. Research Methodology

This section will describe all the steps taken in the proposed framework for stock prediction. The proposed framework includes the following steps also shown in Fig 1. Financial news is put through various sentiment analyzing algorithms to calculate sentiment scores and then merged with historical stock data which is used as an input to MLP Regressor to predict the stock price-output (Fig 1). Further, the detailed methodology along with data collection is explained.

† Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser,

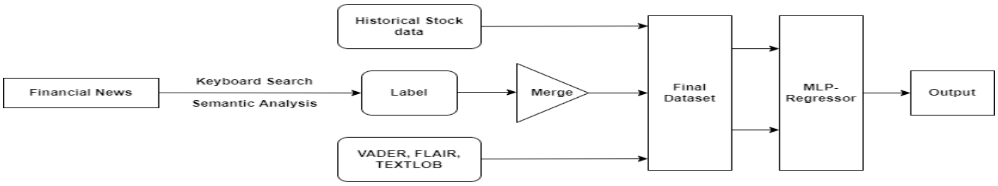

Fig 1. Flowchart of the proposed method

4.1 Data Collection

Historical stock data is available on Yahoo Finance. The stock historical data is downloaded as per the requirement of the period from Yahoo Finance. These .csv format files have values Open, High, Low, Close, Volume, and Adjusted Close for each date the stock market was open. The features represent the Opening stock price, the highest price of the stock on that day, the lowest stock price, closing price, volume of shares traded, and Adjusted Closing price which represents the Closing price of the stock after paying dividends to investors respectively for a specific date [40]. An example of Stock historical price data of HDFC bank is presented in Table 1.

Table 1. Historical HDFC Bank stock price data from yahoo

Date	Open	High	Low	Close	Adj close	Volume
04-01-2010	13.28	13.35	13.16	13.35	12.47	2117000
05-01-2010	13.28	13.40	13.25	13.40	12.52	1906000
06-01-2010	13.25	13.51	13.25	13.46	12.58	1293000
07-01-2010	13.36	13.69	13.35	13.68	12.78	2944000

Table 2: List of companies' understudy

Name of Company	Company sector	Data from	Data to
Reliance	Telecom	Jan 2010	May 2020
Tata Motors	Automobile	Jan 2010	Jan 2020
Tata Steel	Metal	Jan 2010	Jan 2020
HDFC	Banking	Jan 2010	Jan 2020

The historical time series stock data of four companies from different sectors (Table 2) was obtained and used in this study.

These four different companies from varying sectors were chosen to check the generalizability of the proposed model. The companies were taken into consideration to check which sector or company is most influenced by the financial news articles and which sector is easier to predict. Indian Financial news for the time range of historic data used was also collected and historical price data was merged with the financial news data by adding all the news on a specific date from financial news data to the historical data. Headlines and a short description of news for each date were concatenated together. All financial news headlines were added without segregating news of different sectors or companies. To understand the effect of related news and general financial news some keywords related to the company were used to check if that news is regarding a particular company and stored as a new feature *label* which equals *zero* when news is not related to the specific company and *one* otherwise.

4.2 Sentiment Analysis of Financial News

Sentiment analysis of financial news articles is done using three different algorithms to calculate sentiment scores from each algorithm. The algorithms used include VADER [20,39] from NLTK [21] which give scores like positive, negative, neutral, and compound. VADER checks Polarity, the intensity of the emotion by checking how intensely the statement is positive, negative, or neutral [20]. TextBlob [34] from NLP [24] scores the subjectivity and polarity of the financial articles and then Flair from NLP [23] scores each sentence in the textual data and returns an array of the score of each statement.

In this paper, the score of individual statements is summed to provide a feature score sum that represents the complete Flair score of financial news. The above three sentiment scoring sets are used individually and then in combination with each other to find the best and most effective sentiment scores. The algorithms and the sentiment scores under them and their significance are listed in Table 3.

Table 3. Features from different algorithms

Algorithm/ Library	Sentiment score	Range	Significance
VADER	Positive	[0-1]	The proportion of textual data that fall in the positive category
	Negative	[0-1]	The proportion of textual data that fall in the Negative category
	Neutral	[0-1]	The proportion of textual data that fall in the Neutral category
	Compound	[-1,1]	Calculates the sum of all lexicon ratings which have been normalized between [-1,1]
TextBlob	Subjectivity	[0,1]	Subjectivity tells us the extent to which a statement is subjective or objective where 0.0 represents very objective and 1.0 represents highly subjective. The higher subjectivity means text contains personal opinions rather than factual information.
	Polarity	[-1,1]	Calculates the sentiment of a statement where -1 represents a negative statement and +1 is a positive statement.
FLAIR	score/ score_sum	[-1,1]	A state-of-the-art NLP model applied to text generally individual statements to calculate positive or negative comments using a pre-trained model. *Score_sum* is calculated by summing all the individual sentence scores of each textual section.

4.3 Feature Extraction

This proposed model takes an input from various combinations of features. Besides the sentiment scores from the given three algorithms and closing price and previous day closing price from the historical stock price data which are normalized between 0 and 1. The other two features used include *Trend* and *Future Trend.*

Trend and **Future Trend** are extracted by subtracting stock closing price on two dates. Trend and future trends have nominal values of positive, negative, and neutral. The criteria for calculating these values are given in the following equation.

$$trend = \begin{cases} Positive\ if\ C_d - C_{d-1} > 0 \\ Neutral\ if\ C_d - C_{d-1} = 0 \\ Negative\ if\ C_d - C_{d-1} < 0 \end{cases} \quad (1)$$

The trend represents the stock movement on the next day of trade. C_d is the closing price on day *d* and C_{d-1} is the stock price on day *d-1*.

The future trend represents a change in stock movement after *n* days. In this research, *n* is taken as five because the stock market is open five days a week, so *future trend* generally gives us a stock trend movement of one week. The future trend is calculated by subtracting the stock price of day d from day *d+n*. This gives the future trend on the day *(d+n)*. If the difference is positive means the trend will be positive after *n* days and if negative, then it will be negative showing the expected movement of stock price. The below equation can be used to calculate future trend.

$$future\ trend = \begin{cases} Positive\ if\ C_{d+n} - C_d > 0 \\ Neutral\ if\ C_{d+n} - C_d = 0 \\ Negative\ if\ C_{d+n} - C_d < 0 \end{cases} \quad (2)$$

where C_{d+n} is the stock price on the day *(d+n)* and C_d is the stock price *n* days before *(d+n)*.

One more feature used is the *label* which represents if the news is regarding any particular company or not by simply checking the news for certain keywords specific to each company. The label is **true (1)** when a particular date has a piece of news specific to the company under consideration otherwise **false (0)**.

4.4 Final Dataset

A final dataset is complied with historical time series stock price data, financial news sentiments calculated by the discussed algorithm for each day, label, trend, and future trend. All of these features make up the final dataset with some rows shown in Table 4. The combination of different features at once was input to the model along with the previous day's closing price to predict the closing price of the stock market. The accuracy metrics are compared to understand the impact of various sentiment scores on the stock price prediction.

The final datasets of each company is available for research purposes upon request.

Table 4. A view of the final dataset of HDFC stock price with sentiment scores

Date	Close	Label	Subjectivity	Polarity	Compound	Negative	Neutral	Positive	Score_sum
04-01-2010	13.346	1	0.392845	0.026326	0.9349	0.01	0.898	0.091	0.653
05-01-2010	13.403	1	0.348687	0.05759	0.8074	0.076	0.806	0.118	-1
06-01-2010	13.463	1	0.323457	0.083025	0.9756	0.018	0.896	0.086	0.608
07-01-2010	13.678	1	0.281247	0.063513	0.9896	0.038	0.861	0.101	-1.07
08-01-2010	13.719	1	0.368182	-0.01205	0.8969	0.07	0.839	0.091	-1.03
11-01-2010	13.51	1	0.355764	0.013681	0.7351	0.061	0.856	0.083	-1
12-01-2010	12.812	1	0.34876	0.05221	0.9906	0.02	0.899	0.081	-0.973

4.5 Applying Machine Learning Classifier

In this research, an MLP-Regressor [25] is used to predict the stock closing price. Different sets of features for predicting the stock closing price are taken and then used one set of features with different companies data to understand the generalizability of the model. The model used is shown in Fig 2.
Financial news articles are processed and then the sentiment is computed and along with these sentiments the historical stock price is input to the MLP-Regressor to calculate the output from the model.

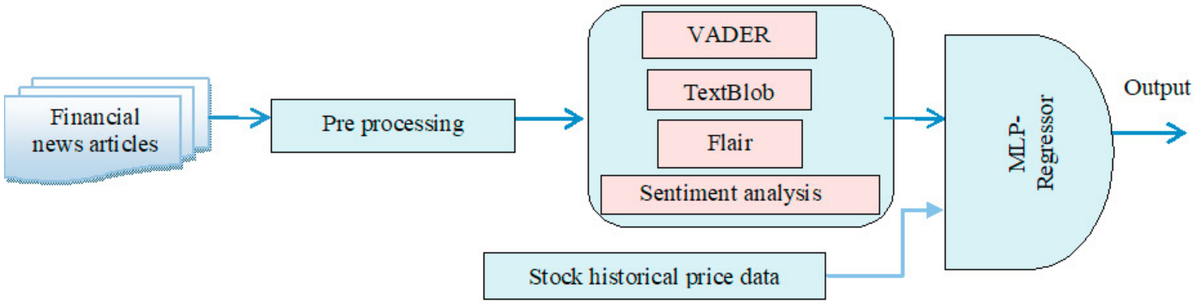

Fig. 2. MLP- Regressor along with sentiment analysis model

4.6 Performance Metrics For The Proposed Study

The metrics which are used to check the performance of the model and evaluate the ability of the model to predict the stock price are discussed in this section. This paper evaluates the effectiveness of the model to predict the stock market on the next day called a *trend.* And also, the ability to predict *future trends* which shows how well the model predicts the stock movement in the future. While predicting the stock price, two things are evaluated, stock price movement and stock price. The effectiveness of stock movement can be checked from trends and future trends. And the ability of the model to predict stock price can be checked by comparing predicted stock price with actual stock price and in this paper, Mean Absolute Percentage Error (MAPE) is used to check the error in stock price prediction without considering stock direction. The accuracy metrics used in this paper are defined as follows.

Accuracy is a popular metric used in classification problems. It represents the number ratio of correctly classified values to a total number of classified values [19] and can be calculated by (3). If the trend is negative, it represents the decision to sell the stocks, and if the trend is positive or neutral the decision is to buy/hold (1/0). And hence accuracy is checked for binary classification viz buy/hold or sell (-1). The values used to calculate Accuracy as in (3), Precision in (4), Recall as in (5), and F1-Score as in (6) use values as given in Table 5.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

This is to be noted the two values are clubbed together for an upward or static trend as 0/1.
Precision represents the ratio of actual positive among a total number of positively classified.

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

Recall is a ratio of actual positives among all real positives

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

F1-score is the harmonic mean of recall and precision. It is a function of both and hence represents the overall scores of a model.

$$F1 - Score = \frac{2*precision*recall}{precision+recall} \quad (6)$$

MAPE: Average relative error of forecast in percentage hence one of the most popular prediction accuracy metrics

$$MAPE = \frac{100}{n}\sum_{j=1}^{n}\left|(y_j - y'_j)/y_j\right| \quad (7)$$

where y_j is the original value and y'_j is the corresponding predicted value. Accuracy, Precision, Recall, and F1-Score have been used to check the metrics of trend as well future trend which explains how well the model can predict the immediate stock movement and future stock movement.

Table 5. Trend values for performance metric calculation

Original Trend	Predicted Trend	Classification group
0/1	0/1	True Positive (TP)
0/1	-1	False Negative (FN)
-1	0/1	False Positive (FP)
-1	-1	True Negative (TN)

5. Experimental Results and Discussion

The model MLP-Regressor is tested with a different combination of sentiments discussed in Table 3. In this paper, the impact of each sentiment analysing algorithm was tested along with the MLP-Regressor. The trend, future trend accuracy (Table 7), and MAPE (Table 6) were checked for all combinations of sentiment feature sets. The performance metrics were recorded for 10 days, 30 days, and 100 days. The effectiveness of models was tested for every combination of HDFC data to understand the impact of all feature set combinations. The graphs for all combinations are plotted to understand the way every model behaves in stock forecast prediction. The prediction metrics for trend and F-trend (future trend) of HDFC data on various models for 10, 30, and 100 days are presented in Table 6.

Fig. (3-10) each represent the graph of prediction of HDFC using MLP Regressor with a different set of features. Table 7 shows that the highest prediction accuracy for the next-day trend can be seen in FLAIR (F), VADER + TextBlob *(V+T),* and *F+T* models which is equal to 0.90 which is very high compared to state of art model which have highest accuracy of 83.22% [1]. It is also observed recall for future trends almost remains highest in every model except in *V+T*. Whereas checking the accuracy metrics for 30 days in Table 7, it is observed maximum accuracy of 0.73 is observed in next day prediction with *F+T* and *V+T+F*. TextBlob (T) and Flair (F) each individually show the same accuracy for future trend and the highest MAPE is observed in TextBlob when checked for a complete test set. In Fig. 4 the uneven spikes in prediction price when using TextBlob justify the high MAPE (Table 6).

Trend and Future trend accuracy when checked for 100 days (Table 7) to understand the long-term applicability of the model it is observed the model is able to predict future trends with the highest accuracy of 0.75 when using Flair sentiments individually and hence shows the effect of FLAIR sentiment in case of stock prediction is of considerable factor. While the next-day trend was predicted best by the *F+T* model. From each model, it can be observed future trend prediction is more predictive with sentiment scores than the next-day trend. MAPE shows how closely the stock price can be predicted to the original price and it is the mean percentage of relative error and is checked for the complete test set.

It is observed *V+T+F* and *V+T+F+Lable (L)* later show less MAPE than *V+T+F* proving the effectivity of the label in a positive direction for a stock price forecast. In Fig.. 4,6,8,9, and 10 it is checked using TextBlob sentiment causes uneven prediction and hence affects the model to predict the amount of change in stock price causing the increase in MAPE. From Figs. 3 and 5 it is observed that VADER and Flair can follow the trend more closely. Figs. 3-10 shows how closely some combinations of features can follow the trend and how combinations with TextBlob and some models with more features have exaggerated fluctuations and more complex than the original plot.

Table 6. MAPE for each model

Sentiment Combination	VADER	TextBlob	FLAIR	VADER +TextBlob	VADER+ FLAIR	FLAIR +TextBlob	VADER + TextBlob +FLAIR	VADER + TextBlob +FLAIR +Label
MAPE	1.77	2.32	1.48	1.64	1.74	1.55	2.00	1.83

To understand the generalizability of the model and to understand which stocks are more affected by financial news sentiments, feature combinations from Table 3 viz Positive, Negative, Neutral from VADER and score_sum from Flair were used to check the prediction metrics for Reliance (Fig. 11), Tata Steel (Fig. 12), Tata Motors (Fig. 13) and HDFC (Fig. 14) for 10, 30 and 100 days (Table 7). From Table 8, it can be checked Tata Motors is having the highest MAPE among all four companies taken into consideration and in Fig 12 all these uneven spikes can be seen.

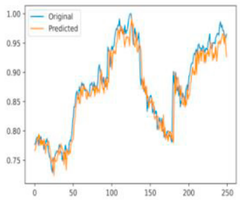

Fig 3. HDFC with VADER plot

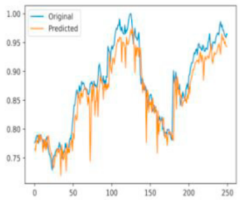

Fig 4. HDFC with TextBlob

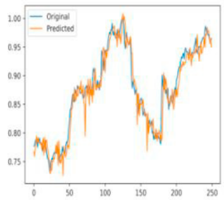

Fig 5. HDFC with FLAIR

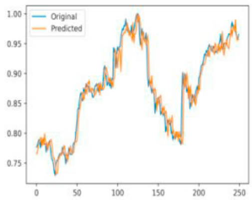

Fig 6. HDFC with VADER and TextBlob

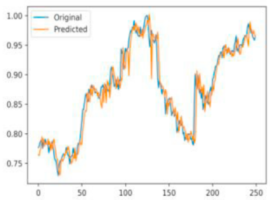

Fig 8. HDFC with FLAIR and TextBlob

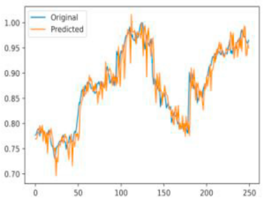

Fig 9. HDFC with Vader, flair, and TextBlob

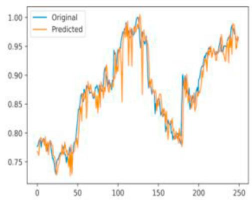

Fig 10. HDFC with flair, Vader, TextBlob, and label

Table 7. HDFC data along with for 10, 30 and 100 days' trend

Sentiment combinations and days		Precision		Recall		Accuracy		F1-Score	
		Trend	F-Trend	Trend	F-Trend	Trend	F-Trend	Trend	F-Trend
VADER	10	0.857	0.875	0.857	1.0	0.8	0.9	0.857	0.933
	30	0.714	0.714	0.625	0.666	0.66	0.7	0.666	0.689
	100	0.56	0.77	0.55	0.76	0.53	0.71	0.56	0.76
TextBlob	10	0.83	0.7	0.714	1.0	0.7	0.7	0.769	0.823
	30	0.66	0.705	0.625	0.8	0.63	0.73	0.645	0.75
	100	0.6	0.77	0.55	0.76	0.56	0.71	0.57	0.76
FLAIR	10	0.875	0.875	1.0	1.0	0.9	0.9	0.933	0.933
	30	0.705	0.733	0.75	0.73	0.7	0.73	0.72	0.73
	100	0.58	0.81	0.62	0.77	0.56	0.75	0.60	0.79
VADER + TextBlob	10	1.0	0.75	0.857	0.857	0.9	0.7	0.92	0.79
	30	0.73	0.71	0.687	0.66	0.7	0.7	0.70	0.68
	100	0.62	0.77	0.62	0.74	0.6	0.7	0.62	0.75
VADER + FLAIR	10	0.6	0.77	0.42	1.0	0.4	0.8	0.5	0.875
	30	0.61	0.66	0.5	0.66	0.56	0.66	0.55	0.66
	100	0.58	0.76	0.57	0.76	0.55	0.7	0.57	0.76
FLAIR + TextBlob	10	1.0	0.77	0.85	1.0	0.9	0.8	0.92	0.875
	30	0.78	0.63	0.68	0.8	0.73	0.66	0.73	0.70
	100	0.66	0.72	0.59	0.74	0.62	0.66	0.62	0.73
VADER + TextBlob + FLAIR	10	0.8	0.77	0.57	1.0	0.6	0.8	0.66	0.875
	30	0.714	0.73	0.625	0.73	0.66	0.73	0.66	0.73
	100	0.55	0.77	0.53	0.71	0.52	0.69	0.54	0.74
VADER + TextBlob + FLAIR + Label	10	1.0	0.77	0.57	1.0	0.7	0.8	0.72	0.875
	30	0.85	0.8	0.75	0.8	0.8	0.8	0.79	0.8
	100	0.63	0.76	0.61	0.73	0.6	0.69	0.622	0.74

Table 8. Accuracy metrics of Reliance, Tata Motors, and Tata Steel

Company	MAPE	No. Of Days	Precision		Recall		Accuracy		F1-Score	
			Trend	F-Trend	Trend	F-Trend	Trend	F-Trend	Trend	F-Trend
Reliance	2.57	10	0.75	1.0	0.75	0.714	**0.8**	**0.8**	0.75	0.83
		30	0.66	0.61	0.66	0.68	0.66	0.6	0.66	0.64
		100	0.53	0.57	0.55	0.61	0.58	0.65	0.54	0.59
Tata Motors	**4.71**	10	0.75	0.88	0.5	0.88	0.6	0.8	0.6	0.88
		30	0.46	0.68	0.4	0.72	0.46	0.63	0.42	0.70
		100	0.59	0.71	0.53	0.73	0.57	0.71	0.56	0.72
Tata Steel	2.55	10	0.25	0.83	0.33	0.71	0.5	0.7	0.28	0.76
		30	0.3	0.55	0.3	0.55	0.53	0.73	0.3	0.55
		100	0.55	0.69	0.51	0.65	0.61	0.67	0.53	0.67
HDFC	1.61	10	0.77	0.7	1.0	1.0	0.8	0.7	0.87	0.82
		30	0.68	0.62	0.68	0.66	0.66	0.63	0.68	0.64
		100	0.61	0.73	0.64	0.74	0.59	0.67	0.63	0.74

In Fig. 11 and 13, it is observed that Reliance and Tata Steel are more closely following the original price plot showing the better relationship between sentiments and price fluctuation of stocks of these companies. Reliance is able to achieve 0.8 accuracies in the next day trend as well future trend for seven days (Table 7). Tata Motors MAPE as well predicted price and original price plot (Fig. 12) and also HDFC plot (Fig. 14) show predicted price does not follow original price closely and the correlation of financial news in stronger with Telecom Company Reliance, Metal company Tata Steel as compared to Banking sector company HDFC and automobile company Tata Motors.

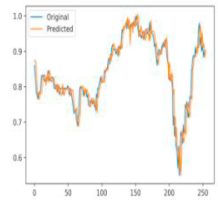

Fig 11. Reliance stock prediction

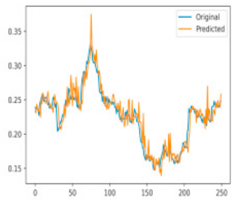

Fig 12. Tata Motors stock prediction

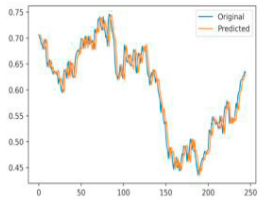

Fig 13. Tata Steel stock prediction

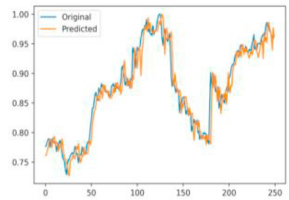

Fig 14. HDFC stock prediction

MAPE for various combinations of sentiments from Table 4 is plotted in Fig. 15 and it is observed for TextBlob MAPE is maximum and causes an uneven shift in prediction prices, *V+T+F* shows the second highest MAPE while when adding *Label* to *V+T+F* the MAPE decreases by 0.17. MAPE for various companies as in Table 8 is pictorially represented in Fig. 16. It can be observed for Tata Motors, MAPE is maximum hence the prediction prices will have more deviation from the real price and HDFC have the lowest MAPE. The proposed model is able to predict stock price trends even to an accuracy of 0.9 and which is comparable to the state-of-art models for stock price prediction. The model can also predict the future trend even up to the same accuracy in some cases and in most cases recall remains highest and shows the model can predict actual positive among all real positives at a rate of 100% (Recall) in most cases. The model was tested with a wide range of sentiment score combinations and explores the best features to be used in sentiments of financial for stock price prediction.

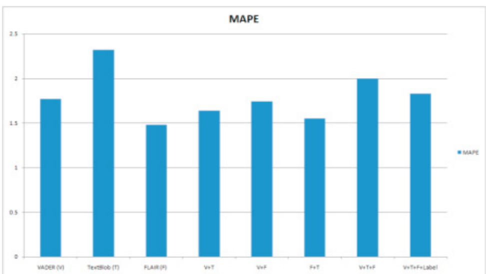

Fig. 15. MAPE for different combinations of sentiments used for HDFC

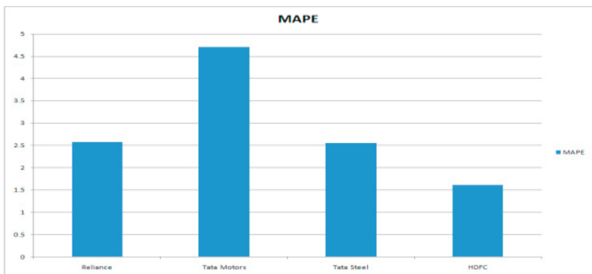

Fig. 16. MAPE for different companies using the same features

The model when used with different companies also shows the effectiveness of the model in general and hence can be used with most stock companies and financial news to forecast the stock trend.

6. Conclusion and Future Work

The proposed study presents a framework for stock price prediction using historical stock data and sentiments of financial news. The sentiment analysis is done using NLP and NLTK libraries to calculate sentiment scores which depict whether the text is positive, negative or neutral. Different combinations of sentiment scores were tested again in the machine learning models with HDFC stock data to understand the effect of each set on the stock price fluctuation. The effect of financial news sentiments is checked for 10 days, 30 days, and 100 days of stock prediction. The accuracy of next-day stock price trend and future stock movements was observed, and each set of sentiments was tested for accuracy as well MAPE to understand the effect of sentiment score on stock prices. Then a set of sentiment scores was used on four different companies to check the generalizability of the model and to see stocks of which companies are most influenced by financial news. It was also observed that using *label* helps reduce MAPE and hence fit the trend and predict stock price more accurately depicting the higher correlation between stock price and financial news about this specific company. While using TextBlob sentiments MAPE increases and uneven exaggeration in stock price change were observed while in the case of the model using Flair sentiments the MAPE was lowest. Higher MAPE in the case of Tata Motors shows less correlation between stock price of automobile companies like Tata Motors and financial news. This paper concludes using financial news sentiments along with MLP-Regressor can predict the stock price to an accuracy of 0.90 and shows a high correlation between stock price and financial news. There are more models to be used in future studies for stock market prediction and an ensemble of some such models can also be employed. LSTM, as it has a memory element, can be used to remember the trend and sentiments can be added to increase the efficiency of the model to predict the rate of change in stock price. Along with financial news, social media posts, political and geopolitical news can also be employed to predict stock prices more precisely and accurately. Furthermore, the inclusion of both market data and textual data from online sources may improve the prediction accuracies.

References

[1] 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.*, no. 0123456789, 2020, doi: 10.1007/s12652-020-01839-w.
[2] X. Pang, Y. Zhou, P. Wang, W. Lin, and V. Chang, “An innovative neural network approach for stock market prediction,” *J. Supercomput.*, vol. 76, no. 3, pp. 2098–2118, 2020, doi: 10.1007/s11227-017-2228-y.
[3] I. K. Nti, A. Felix Adekoya, · Benjamin, and A. Weyori, “A systematic review of fundamental and technical analysis of stock market predictions,” *Artif. Intell. Rev.*, vol. 53, pp. 3007–3057, 123AD, doi: 10.1007/s10462-019-09754-z.
[4] J. L. Wang and S. H. Chan, “Stock market trading rule discovery using pattern recognition and technical analysis,” *Expert Syst. Appl.*, vol. 33, no. 2, pp. 304–315, Aug. 2007, doi: 10.1016/J.ESWA.2006.05.002.
[5] R. Singh and S. Srivastava, “Stock prediction using deep learning,” *Multimed. Tools Appl.*, vol. 76, no. 18, pp. 18569–18584, 2017, doi: 10.1007/s11042-016-4159-7.
[6] J. Zhang, S. Cui, Y. Xu, Q. Li, and T. Li, “A novel data-driven stock price trend prediction system,” *Expert Syst. Appl.*, vol. 97, pp. 60–69, 2018, doi: 10.1016/j.eswa.2017.12.026.
[7] O. Hegazy, O. S. Soliman, and M. A. Salam, “97F44F6Cdaa778B97Efe930F6534B3Edc92F,” vol. 4, no. 12, pp. 17–23, 2013.
[8] 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.
[9] P. Ou and H. Wang, “Prediction of Stock Market Index Movement by Ten Data Mining Techniques,” *Mod. Appl. Sci.*, vol. 3, no. 12, pp. 28–42, 2009, doi: 10.5539/mas.v3n12p28.
[10]P. Chakraborty, U. S. Pria, M. R. A. H. Rony, and M. A. Majumdar, “Predicting stock movement using sentiment analysis of Twitter feed,” *2017 6th Int. Conf. Informatics, Electron. Vis. 2017 7th Int. Symp. Comput. Med. Heal. Technol. ICIEV-ISCMHT 2017*, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/ICIEV.2017.8338584.
[11]S. K. Khatri and A. Srivastava, “Using sentimental analysis in prediction of stock market investment,” *2016 5th Int. Conf. Reliab. Infocom Technol. Optim. ICRITO 2016 Trends Futur. Dir.*, pp. 566–569, 2016, doi: 10.1109/ICRITO.2016.7785019.
[12]D. Lien Minh, A. Sadeghi-Niaraki, H. D. Huy, K. Min, and H. Moon, “Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network,” *IEEE Access*, vol. 6, no. September, pp. 55392–55404, 2018, doi: 10.1109/ACCESS.2018.2868970.
[13]M. R. Vargas, C. E. M. Dos Anjos, G. L. G. Bichara, and A. G. Evsukoff, “Deep Leaming for Stock Market Prediction Using Technical Indicators and Financial News Articles,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, pp. 1–8, 2018, doi: 10.1109/IJCNN.2018.8489208.
[14]E. Chong, C. Han, and F. C. Park, “Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies,” *Expert Syst. Appl.*, vol. 83, no. September, pp. 187–205, 2017, doi: 10.1016/j.eswa.2017.04.030.

[15]L. S. Malagrino, N. T. Roman, and A. M. Monteiro, “Forecasting stock market index daily direction: A Bayesian Network approach,” *Expert Syst. Appl.*, vol. 105, pp. 11–22, 2018, doi: 10.1016/j.eswa.2018.03.039.
[16]V. Turchenko, P. Beraldi, F. De Simone, and L. Grandinetti, “Short-term stock price prediction using MLP in moving simulation mode,” *Proc. 6th IEEE Int. Conf. Intell. Data Acquis. Adv. Comput. Syst. Technol. Appl. IDAACS’2011*, vol. 2, no. September, pp. 666–671, 2011, doi: 10.1109/IDAACS.2011.6072853.
[17]C. Y. Yeh, C. W. Huang, and S. J. Lee, “A multiple-kernel support vector regression approach for stock market price forecasting,” *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2177–2186, 2011, doi: 10.1016/j.eswa.2010.08.004.
[18]A. Moghar and M. Hamiche, “Stock Market Prediction Using LSTM Recurrent Neural Network,” *Procedia Comput. Sci.*, vol. 170, pp. 1168–1173, 2020, doi: 10.1016/j.procs.2020.03.049.
[19]S. Carta, A. Ferreira, A. S. Podda, D. Reforgiato Recupero, and A. Sanna, “Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting,” *Expert Syst. Appl.*, vol. 164, no. July 2020, p. 113820, 2021, doi: 10.1016/j.eswa.2020.113820.
[20]D. E. Rapach, J. K. Strauss, and G. Zhou, “International stock return predictability: What is the role of the united states?,” *J. Finance*, vol. 68, no. 4, pp. 1633–1662, 2013, doi: 10.1111/jofi.12041.
[21]A. Alsaeedi and M. Z. Khan, “A study on sentiment analysis techniques of Twitter data,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 361–374, 2019, doi: 10.14569/ijacsa.2019.0100248.
[22]W. Khan, U. Malik, M. A. Ghazanfar, M. A. Azam, K. H. Alyoubi, and A. S. Alfakeeh, “Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis,” *Soft Comput.*, vol. 24, no. 15, pp. 11019–11043, 2020, doi: 10.1007/s00500-019-04347-y.
[23]M. N. M. Ibrahim and M. Z. M. Yusoff, “Twitter sentiment classification using Naive Bayes based on trainer perception,” *2015 IEEE Conf. e-Learning, e-Management e-Services, IC3e 2015*, pp. 187–189, 2016, doi: 10.1109/IC3e.2015.7403510.
[24]R. Joshi and R. Tekchandani, “Comparative analysis of twitter data using supervised classifiers,” *Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016*, vol. 2016, 2016, doi: 10.1109/INVENTIVE.2016.7830089.
[25]M. Qasem, R. Thulasiram, and P. Thulasiram, “Twitter sentiment classification using machine learning techniques for stock markets,” *2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015*, pp. 834–840, 2015, doi: 10.1109/ICACCI.2015.7275714.
[26]M. Dang and D. Duong, “Improvement methods for stock market prediction using financial news articles,” *NICS 2016 - Proc. 2016 3rd Natl. Found. Sci. Technol. Dev. Conf. Inf. Comput. Sci.*, pp. 125–129, 2016, doi: 10.1109/NICS.2016.7725636.
[27]M. Tirea and V. Negru, “Text Mining News System - Quantifying Certain Phenomena Effect on the Stock Market Behavior,” *Proc. - 17th Int. Symp. Symb. Numer. Algorithms Sci. Comput. SYNASC 2015*, pp. 391–398, 2016, doi: 10.1109/SYNASC.2015.65.
[28]E. Loper and S. Bird, “NLTK: The Natural Language Toolkit,” 2002, doi: 10.3115/1225403.1225421.
[29]R. Bose, P. S. Aithal, and S. Roy, “Sentiment analysis on the basis of tweeter comments of application of drugs by customary language toolkit and textblob opinions of distinct countries,” *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 7, pp. 3684–3696, 2020, doi: 10.30534/ijeter/2020/129872020.
[30]A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, “FLAIR: An easy-to-use framework for state-of-the-art NLP,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Demonstr. Sess.*, pp. 54–59, 2019
[31]S. Biswas, P. Das, R. Bose, and S. Roy, “Examining the Effects of Pandemics on Stock Market Trends through Sentiment Analysis,” *J. Xidian Univ.*, vol. 14, no. 6, 2020, doi: 10.37896/jxu14.6/138.
[32]H. Liu and Z. Long, “An improved deep learning model for predicting stock market price time series,” *Digit. Signal Process. A Rev. J.*, vol. 102, p. 102741, 2020, doi: 10.1016/j.dsp.2020.102741.
[33]J. Long, Z. Chen, W. He, T. Wu, and J. Ren, “An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market,” *Appl. Soft Comput. J.*, vol. 91, p. 106205, 2020, doi: 10.1016/j.asoc.2020.106205.
[34]I. K. Nti, A. F. Adekoya, and B. A. Weyori, “Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana,” *Appl. Comput. Syst.*, vol. 25, no. 1, pp. 33–42, 2020, doi: 10.2478/acss-2020-0004.
[35]R. Pimprikar, S. Ramachandra, and K. Senthilkuma, “Use of Machine Learning Algorithms and Twitter Sentiment Analysis for Stock Market Prediction,” *Int. J. Pure Appl. Math.*, vol. 115, no. 6, pp. 521–526, 2017.
[36]S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu, “Stock price prediction using news sentiment analysis,” *Proc. - 5th IEEE Int. Conf. Big Data Serv. Appl. BigDataService 2019, Work. Big Data Water Resour. Environ. Hydraul. Eng. Work. Medical, Heal. Using Big Data Technol.*, pp. 205–208, 2019, doi: 10.1109/BigDataService.2019.00035.
[37]D. Shah, H. Isah, and F. Zulkernine, “Predicting the Effects of News Sentiments on the Stock Market,” *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 4705–4708, 2019, doi: 10.1109/BigData.2018.8621884.
[38]A. Bhardwaj, Y. Narayan, Vanraj, Pawan, and M. Dutta, “Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty,” *Procedia Comput. Sci.*, vol. 70, pp. 85–91, 2015, doi: 10.1016/j.procs.2015.10.043.
[39]C. J. Hutto and E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” *Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014*, pp. 216–225, 2014.
[40]M. M. Mostafa, “Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait,” *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6302–6309, 2010, doi: 10.1016/j.eswa.2010.02.091.