ISSN: 2633-4828 Vol. 5 No.4, December, 2023

International Journal of Applied Engineering & Technology

STOCK PRICE PREDICTION USING MACHINE LEARNING

[1]Rakshit Rana, [2]Shiva Singh, [3]Shivam Pundir, [4]Swati Sharma and [5]Preksha Pratap
[1,2, 3,4,5] Department of Information Technology,
Meerut Institute of Engineering & Technology, Meerut (U.P)

ABSTRACT

Traditionally, investors relied on the analysis of stock prices, stock indicators, and associated news to forecast market movements. Consequently, the significance of news in influencing stock prices became evident. Previous research in this domain primarily concentrated on categorizing market news as positive, negative, or neutral and illustrating their impact on stock prices. Alternatively, it focused on historical price patterns to predict future movements. In this study, we introduce an automated trading system that integrates mathematical functions, machine learning techniques, and external factors, such as sentiment analysis of news articles, to enhance stock prediction accuracy and generate profitable trading signals. Our specific objective is to forecast the price or trend of a specific stock for the upcoming end-of-day session, taking into account the initial trading hours. To attain this objective, we conducted experiments involving the training of conventional machine learning algorithms and the development of multiple deep learning models, all of which factored in the significance of relevant news. Our experimentation resulted in the highest accuracy of 82.91% being achieved using Support Vector Machine (SVM) for Apple Inc. (AAPL) stock.

Keywords- Machine, Support Vector, Learning, Efficient Market Hypothesis, Stock Market.

1. INTRODUCTION

The economic marketplace represents a dynamic and multifaceted device in which individuals can interact within overover the counter buying and promoting of currencies, stocks, equities, and derivatives through digital structures facilitated with overover the counter brokers. The inventory marketplace, specifically, gives buyers over-the-counter opportunity every day accumulateover the counter shares in publicly traded companies either via formal exchanges or share markets. This avenue has enabled people day-to-day doubtlessly boom over-the-counterir wealth and enjoy monetary prosperity by usingover the counter making an investment pretty small initial sums of cash, with comparatively lower dangers in comparison daily beginning a brand new commercial enterprise or pursuing a high-earnings career (Invesevery daypedia, July 2008). Numerous day-to-day affect stock markets, leading day-to-day uncertainty and excessive stages of volatility. even as people can execute buying and selling orders and put up over the counterm day-to-day over-the-counter marketplace, automatic trading structures (ATS), driven via computer applications, regularly outperform humans in phrases of order execution velocity and precision. over-the-counterover the counter, it remains important to evaluate and manipulate over-the-counter performance of ATS by usingover the counter implementing hazard strategies and safety measures guided with overover the counter human judgment. growing an effective ATS includes thinking about every daydifferent facdayeveryday, which includes overover the counter of trading strategy, over-the-counter incorporation of complex maover-the-countermatical capabilities that mirror a particular stock's country, over-the-counter utility of device every day knoweveryday algorithms for predicting future inventory values, and over-the-counterover the counter of pertinent information every dayassociated with over-the-counter analyzed inventory.

Numerous studies have explored over-the-counter prediction of stock charge traits, by and large inside every day timeframes. those research have worried over-the-counter construction of models that combine numerous data resources, along with information articles, Twitter records, Google, and Wikipedia information. The amalgamation over-the-counter outside daily with inventory prices and technical indicators has tested over-the-counterir affect on stock charge moves (Narayan et al. 2023),(Mall et al. 2023).

The stock market is famend for its inherent volatility, a result of over-the-counter diverse outside day-to-day influencing its conduct, over-the-counter dynamic nature of over-the-counter market, and over the counter

complexity of its multidimensional records. those traits render over the counter assignment of predicting inventory developments/costs in particular day-to-day, even if using superior deep every day knoweveryday models (Singh, Aishwarya 2019). those outside daily may be categorized indaily essential day-to-day, technical everyday, and market sentiments, as delineated under:

- Deliver and demand. for example, if investors tend dayeveryday this inventory more than selling it, this could affect over-the-counter price possibly by way of rising for overover the counter call for might be more than over-the-counter deliver. stock rate Prediction over-the-counter gadget studying.(Narayan et al. 2023),(Mall et al. 2024).
- Stock expenses can have sudden moves every dayeveryday a unmarried information which continues a stock artificially excessive or low. as a result, investors can't are expecting day-to-day manifest with a inventory on a foundation. that is referred dayeveryday market sentiment daily and that overover the counter consist of organization news, economic system, and global activities. (Chaturvedi et al. 2021)
- Global economic system. The float of money and transactions is day-to-day on the economic system of over the counter investors that's tormented by over the counter economy of over-the-counter country.
- Stock historical charges. every inventory has a range which tick facts moves inside, while searching indaily chart patterns and conduct of traders.
- Public sentiments and social media. A tweet from a president or a piece of writing release affects over-the-counter rate of over-the-counter associated inventory(s). as an instance, an unofficial resignation of a CEO on twitter.
- Herbal disasters. for instance, over the counter “haiti earthquake” that killed round 316,000 human beings affected over-the-counter S&P index through taking place 6.6% after 18 buying and selling days.
- Earnings consistent with proportion (EPS) is a essential issue that impacts inventory rate. traders have a tendency everyday buy shares with high EPS in view that over the countery understand that overover the counter gain substantial profits. The call for on this inventory, over-the-counter organization management, over the counter market area dominance and over-the-counter cyclical enterprise performance result in over the counter motion of over the counter stock fee.
- Inflation and deflation are technical day-to-day. Inflation manner higher purchase rate and as a consequenceover the counter higher interest quotes. this may result in a lower of inventory charge. at the contrary, deflation manner lower purchase costs and for that reason decrease earnings and hobby fee.

The difficult interaction over the counter diverse every dayrs, along oover-the-counterrs, exerts a sizable have an effect on on rate movements, over the counterreby contributing day-to-day over the counter complexity of stock prediction. Researchers usually function underneath the belief that market prediction does not adhere daily random conduct (Schumaker, R. et al. 2009). overover the counter, numerous guides have delved inevery day this domain with over-the-counter goal of improving over the counter accuracy of future rate predictions. as an instance, Mark L. et al. (1994) performed a examine examining over the counter effect of public data said by Dow Jones and determined an instantaneous correlation among posted information articles and inventory marketplace sports.

Information releases every dayassociated with a company's operations generally tend day-to-day generate assumptions amongst investors that finally affect rate actions. effective information tends everyday power traders day-to-day, ensuing in an growth in stock expenses. Conversely, poor information prompts traders daily promote, over-the-counterreby pushing inventory charges decrease. at overover the counter over-the-counter a clean connection among information and investors' moves, it's noteworthy that handiest a restrained variety of research comprise over the counter news component inday-to-day over-the-counterir price motion predictions.

An expansion of system day-to-day algorithms can be carried out every day inventory market information for over the counter reason of forecasting destiny inventory price moves. in this observe, we employed numerous over-the-countertic intelligence strategies using both market and news facts. The structure of this paper is organized as follows: section 2 presents a complete overview of over-the-counterover the counter literature on inventory marketplace prediction. segment three outlines over the counter facts collection method, statistics cleaning tactics, and over the counter layout of device studying models. section four presents over the counter results of our experiments, and over-the-counter end, in segment five, we draw conclusions, summarize over the counter paper's findings, and endorse ability avenues for future research.

2. RELATED WORK AND BACKGROUND

In preliminary research related to stock market Prediction, proposed by Fama, E. F. (1970).The Efficient Market Hypothesis (EMH) and Horn, J.C., & Parker, G. G. (1967) proposed a randomPrinciple of walking. These principles proposed that marketPrices are influenced by information other thanHistorical prices and therefore cannot be market prices prediction.The EMH theory suggests that price a Stocks completely depend on market information And thus any new information will lead to price Change in response to new releases Information This theory also claims that stocks Traders where always traded at their fair value Cannot buy or sell shares at special price Undervalued or inflated and therefore the only way

A trader can increase his profit danger. EMH discusses three different variations which Affects the market price: Weak form, where only Historical data is considered, semi-robust form, which additionally incorporates existing public data Historical data and strong form, that go Next to include private data. The EMH states that Any price movement is either the result of a new Published information or random moves that will do Prevent predictive models from succeeding. Horne, J.C. By the random walk hypothesis, & Parker, G.G. (1967) state that stock prices

Randomly changed and it argues past value The movement is independent of the current movement. This is slightly different from the EMH because it focuses Short term patterns of stock market. Based on the above two hypotheses by Horne, J.C. et al. (1967) and Fama, E. F. (1970), Stock The market will follow a random walk and Such movements cannot be accurately predicted More than 50%. In opposition to these theories, many recent ones The study shows that stock market prices Movements are somewhat predictable. This The study relies on two different types of finance Analysis to predict stock market prices:

- Fundamental analysis: It depends on the health of the company and includes

Qualitative and quantitative factors such as interest rate, return on assets, income, expenditure and cost of earnings among others. The objective of this analysis is to check the long term sustainability and strength of the company for long term investment purpose.

- Technical analysis: It is based on time series data. Traders analyze historical price movements and chart patterns and consider time as a crucial parameter in the prediction.Technical analysis can rely on three main keys: stock prices movement although many times the movement seems to be random, historical trends which are assumed to repeat as time passes, and all relevant information about a stock.

In the most recent studies, various machines Learning techniques have been used to predict stocks Prices machine learning proved to be good A tool used in due price forecasting functions Techniques used to analyze the data in that drawing Generalized pattern. Discrete machine learning Models and risk strategies are implemented The function of stock market forecasting attempts to predict Mainly price direction for different time periods Frames and will use different features which Affects the market price.

Arevalo, A. et al. (2016) used four key features as input to a deep neural network (DNN) model. These features can be considered as technical analysis features for the stock market as they are based on mathematical calculations as described below:

- Log return: a finance term that represents the logarithmic difference between the close price at time t and close price at time t-1
- Pseudo-log-return: the logarithmic difference between average prices of consecutive minutes
- Trend Indicator: a linear model applied on 1- minute tick data to generate a linear equation with a certain slope. A negative slope implies a decrease in the price while a positive slope implies an increase and a slope close to zero implies that the price is almost stable.

Arévalo, A. et al. 2016 formalize the input data as follows: the time feature which is included in the inputs as minutes and hours parameters, and a variable window size (n) which is used for the other inputs. Thus, the input file will include last n pseudo-log-return, last n standard deviations and last n trend indicators. The output of the model was "next one-minute pseudo-log-ret. Then after having the input data file ready, it was given to a DNN with one input layer, five hidden layers and one output layer. The data was fragmented into training and testing data. The model was trained during 50 epochs with different window sizes and the results show that window size 3 can show the best performance of the model with accuracy 66% and 0.07 MSE.

Weng, B. et al. (2017) attempted to predict one- day ahead price movement using disparate sources of data, where combining data from online sources with prices and indicators can enhance the prediction of the stock market state. This study was tested on Apple Inc. (APPL) stock information gathered over 3 years with multiple inputs and different output targets. The target was a binary value (0 or 1) which represent a fall or rise of variation between prices. Four datasets were gathered from disparate sources: first dataset includes the public information available at yahoo finance online for stock prices; second dataset includes number of unique page visits to Wikipedia per visitor per day; third dataset includes count of data published on google related to a company on a specific date; forth dataset includes three technical indicators (Stochastic Oscillator, Larry William, Relative Strength index) that represent the variation of stock price over time. Additional features were generated from the four datasets to provide a meaningful parameter for the model. Twenty features were selected as input. A common observation was drawn, that for any target, all the datasets were represented by at least one feature. Different AI techniques: Artificial Neural Network (ANN), Support Vector Machines (SVM) and Decision Trees (DT) were applied to predict stock price movement and compared to each other. After the evaluation on the three different models listed above, the output comparing open price of day i+1 to open price of day i achieves the best prediction accuracy with around 85% using SVM model.

Schumaker, R. P. et al. (2009) tried to predict direction of the price movement based on financial news. The study was done in 2009 as market prediction was and still facing difficulties due to the ill-defined parameters. In order to use the financial news articles in the prediction model, news should be represented as numerical value. Several techniques have been known to analyze articles related to certain stock to label these articles with sentiments or use them as vectors for the input features. These techniques could be bag of words, noun phrases, named entities and proper nouns. Proper noun technique is a combination of noun phrases and named entities. The proposed technique outperformed other techniques based on a comparison study.

AZFin Text is another system built by (Schumaker, R. P. et al 2009) that predicts price changes after 20 minutes of news release. The main component of this system is the financial news articles collected from yahoo finance and represented as noun phrases; all the collected noun phrases are represented as vector of binary values indicating the presence or absence of a phrase in the article. The second main component of this system is the stock price data collected in one- minute time frame. Then, the final major task after collecting the data and formalizing the inputs was building and training the AI model. To finalize the input of the model, stock price quotation at the same minute news was released, have been added to the input matrix, in addition to that +20 minutes price which will be the output of the system. The data was then fed to different models. Support Vector Regression (SVR) model was built to predict the price after 20 minutes of news release. Only the data during market time was included leaving 1 hour for opening of the market to show the effect of news released during the

closure of the market. Moreover, a new constraint was added to the model where only one article could be used for 20 minutes. If two articles were released during the same 20-minute period, both will be discarded. The results show that the average directional accuracy established was 71.18%.

It is evident that released news and published articles affect the market. Most of the existing studies analyzing news rely on shallow features such as bag-of-words, named entities and noun phrases. A newer representation was introduced by

(Ding, X. et al. 2014) which represents news as structured events to predict the daily stock price movement. Unlike the previous approaches, this representation can show the relation between events since representing phrases as vectors or bag of words cannot show the actor, action, and the actor which the action was applied on, thus trivial representations cannot show the relation between event and stock. To evaluate the performance of this new representation, news articles data were collected from Reuters and Bloomberg, in addition to the daily close prices of S&P index.

Two different models were created to test the representation: a linear SVM model that has a news document as input and a +1 or -1 output indicating an increase or decrease in price for different time frames (1 day, 1 week and 1 month). A nonlinear deep neural network model is also implemented to learn the hidden relationships between events.

The input functions for both linear and nonlinear models were the same: bag-of-word functions that use a trivial TFIDF representation after removing stop words and event elements represented by different combinations of tuples

($o1$, P, $o2$, $o1$+ P, P +$o2$, $o1$+ P + $o2$) where $o1$ is the first object to the left of the extracted sentence above and $o2$ is the closest object to the right and P represents the verb. This function representation is used to reduce the sparsity of representation next to verb classes.

Different scenarios were used to evaluate the models. When comparing the results of the models with bag-of-words representation, structured events performed better. From another point of view, when comparing models, DNN performed better than SVM due to its ability to learn hidden relationships. Additionally, it was differentiated by the different time frames used (1 day, 1 week, 1 month); the shorter the frame, the better the results. Thus, DNN with structured event features was the best model for daily prediction with an accuracy of around 60%.

As the recent machine learning based studies above show, stock price movements can be predicted with more than 50% accuracy, contradicting EMH and random walk theory using different time frames, functions and models. In the next section, we describe our proposed prediction models in detail and highlight their superior performance over existing models.

3. Proposed method

The proposed approach is divided into several steps, each of which is detailed in this section as follows: Section 3.1 describes the information sources we used to build our system. Section 3.2 presents the treatment of data sources. Section

3.3 represents the manner in which news and prices were matched. Section 3.4 presents the input functions. Section 3.5 shows the data normalization method and part 3.6 discusses the proposed models.

3.1 Data Sources

For our study, we need two sources of information: (1) news sentiment and (2) historical prices. Ten years of tick data and news were collected from Reuters platform from Jan-01-2008 to 2017-12-31 for five different stocks AAPL for shared apple company, GOOGL for google stock, AMZN for Amazon stock, FB for Facebook stock. So a tick is a measure of the minimum price movement up or down. In many cases, a one second time frame includes many ticks reaching 20 to 30 ticks.

Tick data was collected to include the following details: open bid, close bid, high bid, and low bid, in addition to the timestamp. This high-frequency data is collected to perform intraday short-term prediction. Our model requires at least one tick to be released every hour because we group our data every hour. This huge data requires some pre-processing that takes into account the large volume of data (7 trading hours * 3600 = 25200 ticks per day) and the difference in the interval between tick data. Ticket data may have multiple prices released in the same second and some tickets may be missing in other seconds. In addition to tick data, we collected sentiment reports. News data includes stock symbol, release date and time, source, news headline, sentiment (0 for neutral news, 1 for positive news, and -1 for negative news), negative sentiment polarity, positive sentiment polarity, and neutral sentiment polarity. News polarity is based on the number of positive/negative words in a news article.

3.2 Data pre-processing

Due to the huge amount of Tick data and to facilitate data manipulation, we imported our data into a MySQL database where the data is sorted when queried.

The first step was to replace the missing ticks. Tick data has different time intervals in the data collected between ticks. This is because data is not recorded for a period of time. For example, a second may have four prices recorded, and subsequent seconds may not have a single price recorded. To fill in the missing whistles, we look for the nearest ticket dates to fill in our missing seconds. After importing the data into our database and filling in the missing checkboxes, we group our data into a one minute time interval where we get the last received tick for each minute recorded in our data. We then store the pure one-minute data into a new table (no weekends, no off-market tickets).

3.3 Alignment of messages with tick data

Unlike other approaches that filter out messages outside trading hours and messages issued in the same interval, we have created different scenarios to handle these cases. When generating our data, we give the user the option to choose one of the following three reporting scenarios:

1) The last sentiment received on the given day based on the time to be used: for example, if we want to get the sentiment for 01-03-2010 at 14:00, we will get the last sentiment received on 01-03-2010 before 14:00 and adopt sifter. If there is no sentiment, we consider the sentiment to be neutral.

2) The latest sentiment in the selected time interval: if we group our data into an hourly time frame, we check the last sentiment published during that hour and consider it dominant, and if no news is published, we consider the sentiment neutral.

3) Overall average for the day in the selected interval: if more than one sentiment is released during the time frame, we calculate the average for positive (a_p), negative (a_n) and neutral (a_{nu}) messages (i.e.

$$a_p = sum(positive\ news) \quad / \quad count(all\ news)$$

In the case of the same sentiments, we add up the polarity of sentiments (polarity of positive sentiment, polarity of negative sentiment, polarity of neutral sentiment features) and check which of these features has the highest sum and consider it as the dominant sentiment. In the case of the same polarity, we consider a neutral sentiment. In this scenario, we will apply the above formulas to the weekend data for the Monday sentiment label.

For tick data, the data properties were generated from our one-minute and tick-based database tables in an hourly interval. As such, the input to the machine learning algorithm will be hourly data functions with one sentiment function based on one of the above scenarios and the output of the trained model will be the final price of the day.

3.4 Function generation

Different window sizes were tested in our models, i.e. how many hours do you want to go back when you want to train the models. This will generate our input data in the following format (window size * properties).

The functions used in our models are as follows:

- Maximum: Maximum price received during the selected hour
- Minimum: Minimum price received during the selected hour
- Average: Average price received during the selected hour
- Standard Deviation: Standard deviation of prices received during the selected hour
- Pseudo log return: Logarithmic difference between average prices of two consecutive hours.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \frac{\sum p_i}{count(p_i)})^2}$$

$$\ln\left(\frac{p_t}{p_{t-1}}\right)$$

where p_t *is the average price at time* t

- Trend Indicator: The slope of a linear model applied to the relevant hourly tick data, which gives an idea of the trend over the past period of time.hour
- Price: The last tick received at the selected hour
- Sentiment: News sentiment analysis calculated based on the selected scenario outlined in Section 3.3.

So, our input data has 8 features, the formula for the number of features is as follows:

Features=8n where n is window size

The output of our model is the price at the end of the day

3.5 Data Normalization

Since the features extracted from the input data are of different units and scales, normalization is needed to scale the data between 0 and 1, which will also help in faster convergence. To normalize our data, we use the minmaxscaler function provided by the scikit-learn framework. This function gets the maximum and minimum values of each column and performs the following formula:

$$\frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Next, we experiment with various models, namely: Recurrent neural network, Deep neural Network, Support vector Machine and Support vector Regression.

3.6 Models

In this section, we trained different models and compared the effectiveness of recurrent neural network (RNN), feed forward neural network (FFNN), support vector machines (SVM) and support vector regression (SVR) in predicting the direction of today close price with respect to yesterday close price based on the features presented in section 3.4. We tested with the following stocks: AAPL, GOOGL, AMZN and FB for the data collected over 10 years.

For each model, we tried different combinations of window sizes and sentiment scenarios. Window size is a variable, which decides the different number of trading hours during the day; to train our model, we generate data for day d based on first

{4,5,6} trading hours of the day. The data was normalized and split into two sets: training data of 90% and testing data of 10% for RNN, SVM and

SVR models. However, for FFNN we applied the same structure presented in (Arévalo, A. et al. 2016)

FFNN is widely used nowadays for different problems such as classification, regression and pattern recognition in various industries such as financial operations, trading business, analytics and product maintenance. In (Arévalo, A. et al. 2016), the network was formed of 5 layers each with I, 4I/5, 3I/5, 2I/5, I/5 and 1 neuron where I represent the number of inputs. Tanh was the activation function used for all hidden layers and linear function for output layer. This network was applied on H2O platform (Arora, A., et al. 2015); a leading open source data science platform. This platform includes the implementation of deep learning algorithms. After splitting the data into 85% training and 15% testing, we trained the model for 50 epochs and applied ADADELTA (Zeiler, M.D 2012) optimization algorithm to improve learning rate learning process. ADADELTA is a per-dimension adaptive learning rate method for gradient descent, where it is not necessary to search parameters for gradient descent manually and is robust to large gradients and noise. RNN is used for sequence data and differs from DNN by its ability to keep data from previous steps. The memory of RNN could be represented by different cell types: Vanilla RNN (for short term memory), LSTM and GRU (enhance short-term memory of Vanilla RNN using gates mechanism). In our RNN model, we have tried different network structures with different number of neurons at each layer. We tried different network structure through varying the number of layers between 3 and 7 while varying the number of neurons at each layer between 250 and 5 neurons.We tested the implemented networks to get the best results for -layers and 4-layers networks.

We have trained and tested this model on training and testing datasets generated after normalization. The output is the actual price at end of day. Moreover, we have tried different RNN cells provided by TensorFlow. We trained our model on Basic RNN cell, LSTM cell and GRU cell. We trained the model for 100 epochs and applied ADAMOptimizer as our optimization algorithm to get the best learning rate for our model.

SVM, a supervised machine learning algorithm, can be used for both regression and classification problems. This algorithm uses a kernel trick technique that transforms the data and then finds the optimal boundary between outputs. Moreover, SVM shows that it can perform well on non-linear dataset problems, based on the kernel we choose in training SVM model. SVM have been widely used for stock market prediction. In our SVM model, we have tried different kernel algorithms tuning parameters for each model: Linear, Polynomial and RBF. We have trained and tested this model on our training and testing datasets generated. The output is the binary value, 0 when yesterday close price goes down with respect to today close price and 1 when the price goes up. We used scikit-learn library to build this model and we have trained the model and applied GridSearchCV to choose the best parameters to fit our model.

SVR is the same as SVM, however it is used for regression instead of classification. It uses same terms and functionalities as SVM to predict continuous value. In this model, we follow the same process of SVM except for the output, which is not a class, rather end-of-day price.

4. RESULTS AND DISCUSSION

In this section, we show the results obtained for the models defined in Section 3.6 at different stocks. The evaluation metrics are (1) directional accuracy, which analyzes the direction of the predicted value with respect to yesterday's closing price, (2) precision, which measures the consistency of the result, (3) recall, which measures how many correct relevant results are returned, and (4)) F-measure, which measures the weighted average of precision and recall. Based on the directional accuracy metric (Table 2), SVM outperforms RNN, SVR and DNN for different tested stocks. In Table 1. We describe the input data.

Table 1. Stock Data Details

Stock Name	Total Data points	Total Articles	output direction
AAPL	19,243	78,036	1,478 positives 1,271 negatives
FB	11,515	30,198	886 positives 759 negatives
GOOGL	8,225	19,829	625 positives 550 negatives
AMZN	19,243	37,265	1,450 positives 1,299 negatives

According to Table 2, it is very clear that our SVM model is able to achieve more than 50% accuracy. Looking at Table 3, it is also clear that SVM outperforms SVR, DNN, and RNN. All achieved accuracies are above 75% and in case

Table 2. SVM Directional Accuracy Results

Sentiment-Window	Directional Accuracy			
	AAPL	GOOGL	AMZN	FB
S1-4	78.18%	70.94%	75.27 %	68.9%
S1-5	83.36%	80.34%	74.91 %	73.17%
S1-6	81.73%	79.62%	65.82 %	74.66%
S2-4	79.27%	70.94%	74.18 %	73.17%
S2-5	82.64%	77.78%	74.18 %	74.01%
S2-6	81.09%	79.76%	68.36 %	73.27%
S3-4	79.27%	70.09%	75.64 %	75%
S3-5	82.91%	76.92%	70.18 %	73.78%
S3-6	81.64%	76.62%	68.73 %	60.74%

Of APPL, the achieved accuracy is about 83%. All our models achieved better results than the results reported in the literature shown in Table 4.

Based on the reported results, we summarize our contributions as follows:

• We highlighted the impact of news sentiments on stock price movements

• We identified the best time interval for stock price forecasting.

• We have identified the best news scenario and every stock is affected by news differently.

• Our model analysis suggests that the close price or trend with respect to yesterday's closing price can be predicted using various AI models.

• Our proposed model can be used in various ways. First, our model can be used by traders without programming knowledge. These traders can use our model either just to predict price variations and to help traders in their analysis. Also they can use our automated trading system without any supervision, where the system opens and closes trades based on predictions. Finally, our code can be easily configured to trade short-term.

Table 3. All Models Directional Accuracy

	SVM	SVR	DNN	RNN
APPL	**82.91%**	79.2%	81.32%	81.3%
AMZN	**75.27%**	72.26%	74.03%	74.56 %
GOOGL	**80.34%**	66.38%	80.1%	68.38 %
FB	**75%**	68.71%	72.68%	72.39 %

Table 4. Related Work Accuracies

Paper	Metric	Value
Arévalo, A. et al. (2016)	Directional Accuracy	66%
Schumaker, R. P. et al. (2009)	Directional Accuracy	71.18%
Ding, X. et al. (2014)	Accuracy	60%

5. CONCLUSION AND FUTURE WORK

In this paper, we have developed a stock price trend prediction system. To build this model we have collected data from two sources (i) historical stock market data from Reuters and (ii) news sentiment published on certain stock; This data was collected for 4 different stocks over 10 years. Technical features are calculated and used as input data for our model in addition to the 3 scenarios considered while adding sentiment to the calculated features. Our AI framework mainly includes DNN, RNN, SVR and SVM for prediction. We tested our proposed prediction model on APPL, AMZN, GOOGL and FB stock shares for the collected data from (January 1, 2008 to December 31, 2017), resulting in 82.91% accuracy. To our knowledge this is the best accuracy reported in the literature so far. After developing our model, and to show its performance we will implement a risk strategy to test the profit we will get based on our predictions and a few enhancements to our prediction model can be made and studied. One direction is to add additional technical indicators used in the stock market. Another direction is trying different time-frames to group our data. Finally, we can try to increase the exact price prediction.

REFERENCES

[1] James Allan, Jaime Carbonell, George Doddington,Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study: Final report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.

[2] Chinatsu Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen. 1997. A scalable summarization system using robust NLP. In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 66-73, Madrid, Spain.

[3] Breck Baldwin and Thomas S. Morton. 1998. Dynamic coreference-based summarization. In Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3), Granada, Spain, June.

[4] Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In Proceedings of the CL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 10-17, Madrid, Spain.

[5] A. Siddharthan, A. Nenkova, and K. McKeown. Syntactic simplification for improving content selection in multi- document summarization. In Proc.of COLING, 2004.

[6] kumar Mall, Pawan, et al. "Self-Attentive CNN+ BERT: An Approach for Analysis of Sentiment on Movie Reviews Using Word Embedding." International Journal of Intelligent Systems and Applications in Engineering 12.12s (2024): 612-623.

[7] Narayan, Vipul, et al. "7 Extracting business methodology: using artificial intelligence-based method." Semantic Intelligent Computing and Applications 16 (2023): 123.

[8] Narayan, Vipul, et al. "A Comprehensive Review of Various Approach for Medical Image Segmentation and Disease Prediction." Wireless Personal Communications 132.3 (2023): 1819-1848.

[9] Mall, Pawan Kumar, et al. "Rank Based Two Stage Semi-Supervised Deep Learning Model for X-Ray Images Classification: AN APPROACH TOWARD TAGGING UNLABELED MEDICAL DATASET." Journal of Scientific & Industrial Research (JSIR) 82.08 (2023): 818-830.

[10] Chaturvedi, Pooja, Ajai Kumar Daniel, and Vipul Narayan. "Coverage Prediction for Target Coverage in WSN Using Machine Learning Approaches." (2021).