Section: Research Paper

An Investigation of Social Media Influence on Stock Market Price Prediction Using Sentiment Analysis and Machine Learning

Dr. Binita Verma, Assistant Professor, JECRC University, Jaipur
Dr. Swati Namdev, Assistant Professor, Career College, Bhopal

***Abstract*— In Machine Learning (ML) stock market price prediction is one of the complex problem due to volatile nature and dependency on various real world consequences. However, there are a number of efforts and claims are exist which are offering the accurate price prediction. But most of them are not effectively fit for all the scenarios of stock market price prediction. In this paper, we have proposed a ML based stock market price prediction model for investigating the influence of social media information for accurate stock market prediction models. In this context, first we have carried out a survey on recently contributed data models for stock market price prediction. Next, a simple deep neural network is trained on Yahoo Query Language (YQL) database to predict the stock market price. Third, the deep learning model has been extended to incorporate the social media sentiment data with the historical price data for finding influence on price prediction. The experiments on two popular Indian banks namely State Bank of India (SBI) and ICICI bank has been carried out, additionally for inclusion of social media information we have used the Twitter based bank news. The experiments have been carried out and the comparison among their prediction accuracy performed. Additionally, the improvements are reported in different scenarios of experiments. The finding demonstrate the deep learning model provides the accurate prediction of stock market price, additionally the social media information can help to understand the possible price movements and also can improve the accuracy up to 3-5%. Finally some more facts have been identified based on which we have proposed the future extension of the proposed model.**

***Keywords*—stock market price prediction, deep learning, sentiment analysis, social media data analysis, influence measurement.**

I. INTRODUCTION

Machine Learning (ML) [1] has become popular and acceptable in various real-world applications. The ability to learn with historical records, the discovery of relationships among different attributes and the identification of patterns make it valuable in different domains of applications in engineering, business, banking, medicine, and others [2]. These applications use ML techniques to classify, recognize, categorize and predict a specific kind of pattern. Among different kinds of pattern analysis, prediction is one of the popular applications; predictive techniques are used for approximating a continuous value. In this presented work, predictive data analysis is the key area of investigation. Therefore stock market price prediction problem has been considered for investigation. The stock market price prediction is a classical area of research in machine learning. Additionally, significant research work has been done in this area. But most of the work has directly utilized the historical price data on machine learning algorithms but the stock market price has been influenced by various real-world events. The stock prices can be influenced by politics [3], natural disasters, and terrorist attacks [4]. Additionally also based on company sentiments such as a number of contracts, running projects, client base, and others [5]. All this news can be obtained from social media.

Therefore in this presented work, we have proposed to measure the influence of social media news on stock market price prediction accuracy using machine learning technique. In this paper, we first proposed a deep neural network and trained it with the historical stock market price data. Next, the prediction accuracy has been measured. Secondly, the deep learning model with social media sentiments indicator and stock market price has been used to train the network. Additionally, the performance of the algorithm has been measured. Based on the experiments and the difference between the performance in terms of accuracy has been measured. Additionally, different facts that influence the performance of the prediction algorithms have been concluded and the future extension of the work has been proposed. This section has provides an overview of the proposed work, and the next section will provide the study of different recently available models for accurate stock market price prediction. Next, a model has been proposed for stock market price prediction and influence study. Further, the experimental analysis has been carried out with different experimental scenarios and their results are reported. Finally, the future extension of the work is provided with the experimental observations.

II. LITERATURE SURVEY

S. M. Carta et al [6] propose a binary classification to predict future stock price variations in terms of high or low. Using globally published articles sets of lexicons are generated to identify essential words in a specific time interval and within business sector. Feature engineering is performed on lexicons, and used with a Decision Tree. The predicted classes demonstrate the company's next day stock price variation. The performance evaluation using walk-forward technique has been done, which shows that technique perform better than competitors.

Table 1 review of recent stock market price prediction models based on sentiment analysis and ML algorithms

Ref	Research issue	Features and Dataset	Prediction algorithms	Results
[6]	predict the magnitude (high or low) of stock price variations using binary classification	S&P 500 index, most impactful words in a specific time and in certain business sector, generated lexicons.	Decision Tree classifier	achieving a Balanced Accuracy between 0.5 and 0.6 scores
[7]	Understanding the determinants of satisfaction in P2P hosting	100,000 customer reviews left on the Airbnb, NLP algorithms to find clues that help to segment customers	multiple linear regression and support vector regression	Individual MLR (0.660), SVR (0.657), Couple MLR (0.684), SVR (0.676) and Family MLR (0.712), SVR(0.711)
[8]	Attempts to use the latest technologies to design a framework for financial asset price prediction	Yahoo Finance API, news data from SeekingAlpha, Features using FinBERT, FFT, ARIMA, stacked auto-encoder, PCA and XGBoost	Generative adversarial network, where the generator is Seq2Seq and the discriminator is GRU.	its RMSPE is averaged around 2.95%, i.e., it can predict the next day's financial asset price with a margin of error of 5%
[9]	Use advantage of sentiment analysis in stock market	Apple products related Tweets, by StockTwits. Yahoo Finance.	SVM	Accuracy of 76.65% in stock prediction.
[10]	Investigates if and to what point it is possible to trade on news sentiment and DL	Dow Jones industrial average, 25 daily news headlines between 2008 and 2016	LSTM	53-58% accuracy
[11]	How to prevent risk of investment in specific field of stock market	TF-IDF, Word2Vec, CountVectorizer, Doc2Vec, dimension reduces by PCA.	Adaboost, XGboost and GBDT, decision tree, logistic regression	Training precision of 62.282%. The test accuracy of GBDT Model is 64.557%.
[12]	Survey the different approaches of ML that can be incorporated in applied finance	--	Walkthrough in direction of applicability of ML stock market prediction.	--
[13]	find the correlations between sentiments and trends in stock market	Saudi Arabian stock market, TADAWUL. More than 277K Arabic tweets	Pearson's correlation coefficient, Kendall rank correlation and Spearman's correlation	(1) Largest Arabic tweets dataset in finance. (2) Influence of Twitter on the Saudi stock market
[14]	Financial time-series prediction has been long and the most challenging issues	stock indexes Hang Seng Index's (HSI), Deutscher Aktienindex (DAX), S&P 500	Combined GRU-CNN	Accuracy 56.2% on HIS, 56.1% on DAX and 56.3% on S&P500
[15]	The existing combination (news and numerical data) cannot exploit their complementarity.	News corpus and numerical data from the CSI300 and S&P500	Numerical based attention (NBA) method	Accuracy D-CSI300 (0.528), D-S&P500 (0.596), M-CSI300 (0.644)
[16]	Explore predictive power of historical news sentiments based on financial market performance to forecast financial sentiments.	Wikipedia and Gigaword five corpus articles, DJIA 30 Index	recurrent neural network with long short-term memory	Yields P = 0.76%.
[17]	Show that information extracted from news is better at predicting the direction of asset volatility movement	Latent Dirichlet Allocation to represent information	Simple naïve Bayes classifiers	Average directional prediction accuracy for volatility, is 56%, asset close price random at 49%
[18]	Performance of LSTM is dependent on hyper-parameters and no guidelines for configuring LSTM.	A dataset was created from Indian stock market	Long Short Term Memory (LSTM)	P-values were as follows: ICICI 0.898, TCS 0.216, RELIANCE 0.002 and MARUTI 0.263.
[19]	hypothesizing about markets through the application of behavioral finance	Thomson Reuters and Cable News Network	knowledge graph and deep learning	Accuracy of 61.63%, 59.18%, and 58.48% for Apple, Microsoft, Samsung
[20]	Difficulties of training RNNs have discourage their adoption	Dow Jones Industrial Average (DJIA)	deep RNN	f1 measure is 0.82
[21]	find the sentiment from the Bengali paragraph	social network sites, Bengali blogs	Naive Bayes	accuracy is 86.67%
[22]	Reviews architecture and recent advancement of deep structured learning	-	-	useful for readers and students in the early stage of deep learning studies
[23]	Accurate modelling of stock market trends via news disclosures	sentiment-related features	SVM and PSO-based model	Accuracy 0.5915

[24]	reviews on different available SMP techniques	-	Regression and LSTM	-
[25]	Sentiment dataset was constructed from lingual dataset which is unrelated to financial sector and led to poor performance	financial news dataset and Harvard IV-4, S&P 500 prices	GRU, Stock2Vec	accuracy of 66.32 percent

M. Chiny et al [7] took a ML approach to examine 100K reviews of customer of Airbnb to identify different dimensions which describe customer satisfaction for each category. The data does not have any information of customer category. So, NLP is used on review data to identify features which help to categorize them, and used multiple linear regression and support vector regression. Additionally, calculate 6 scores i.e. cleanliness, precision, check-in, location, communication and value. The category of customers based on accommodation the performance is considered. The results show that customers are not equally satisfaction level. In addition, difference was observed depending on the category of client. Results shows improvements made to the rating system for each category of clients.

Q. Bi et al [8] try to utilize the new AI techniques for designing a system for financial resource price prediction. The technique has three parts (1) Feature Engineering to recognize different features using FFT, FinBERT, ARIMA, PCA, stacked auto-encoder, and XGBoost. (2) Regressor includes a generative adversarial network, which usages Seq2Seq as generator and GRU as discriminator. (3) HyperOptimizer used to tune GAN parameters through Bayesian algorithm. Experiments show that the framework performs better than similar method.

R. Batra et al [9] performed sentiment analysis for Apple products on tweets extracted by StockTwits. Additionally Yahoo finance based market index data have used. The sentiment score is calculated through SVM by categorizing twits as bullish or bearish. Finally the market data and sentiment score is used to predict next day's stock movement. Results demonstrate a positive relation among market data and people opinion with an accuracy of 76.65%.

M. Vicari et al [10] investigates when we can trade on the basis of news sentiment. The Deep Learning models are developed with the large amounts of data and can learn automatically. This may help to investors who want to enhance their trading system. In this context LSTM is good choice. They explained how DL models are trained and use market sentiment for forecasting. The prediction is performed for Dow Jones and by analyzing 25 news headlines. The analysis is performed on two scenarios which provide five step ahead predictions and tested on real-world.

J. S. Yang et al [11] used NLP for text analysis using Word2Vec, TF-IDF, Doc2Vec and CountVectorizer. For dimension reduction they used PCA, and finally ML techniques such as XGboost, Adaboost and GBDT, logistic regression and decision tree has used. The results show GBDT is providing accurate stock prediction. As compared to the previous methods based on sentiments based on news provide more accurate prediction.

H. Vachhani et al [12] survey the different approaches of ML that can be incorporated in applied finance. The motivation behind ML is to draw out the specifics from the available data from different sources and to forecast from it. Different ML algorithms have their abilities for predictions and depended on the number of parameters as input features. This work is trying to offer a walkthrough and objective for applicability of the ML algorithm for stock market prediction.

M. Alshahrani et al [13] has finding the correlations among Saudi Arabian stock market and Arabic sentiments using the ML and NLP. They were crawled 277K tweets among 114K were annotated. For measuring correlations, Kendall rank, Pearson's coefficient, and Spearman's rank is used. They report that the users, who have a significant impact on the stock market, could be predictable. They contribute by collecting a largest Arabic tweets dataset based on finance. Second is that it is the first to study the influence of Twitter on the Saudi stock market.

M. Zulqarnain et al [14] combine CNN and RNN architectures to take the advantages of prediction. The model for financial time series using CNN is presented, and then provides it to a gated recurrent unit (GRU) layer to get long-term dependencies. GRU perform better in sequential learning. The evaluation is performed on three datasets Deutscher Aktienindex (DAX), S&P 500, and Hang Seng Index's. Results show that the GRU-CNN model offers best accuracy on HIS (56.2%), DAX (56.1%) and S&P500 (56.3%).

G. Liu et al [15] provide a technique known as numerical based attention (NBA) method for dual sources stock market forecast. The contributions are: (1) an attention-based method to use connection among news and numerical data. The trend in news is mapped into the numerical data. This method has reducing noise and use news information. To evaluate NBA news and numerical data three dataset namely CSI300 and S&P500 is used. Experiments have done for proving NBA is superior.

W. Souma et al [16] examine the power of news sentiments based financial market prediction to forecast financial news sentiments. The sentiments are defined based on stock price returns over one minute after release of a news article. If the stock price is positive (negative), they classify article released just prior to the practical stock. They use Wikipedia and

Gigaword, and for word representation apply global vectors for deep learning input. They analyze intraday Thompson Reuters News and Dow Jones (DJIA 30) Index. The combination of deep RNN and LSTM is used to train, and test the prediction ability. They find accuracy will improve when we switch news selection from random to positive and negative.

A. Atkins et al [17] demonstrate the news information is good for predicting the direction of volatility movement, or statistics, as compared to direction. The results provided using ML models and Latent Dirichlet Allocation from news, and naïve Bayes is used to predict the movements. Results show that the directional prediction accuracy, on new information, is 56%, while for close price is 49%. They provide results by range of stocks in the US market. They found that volatility movements are more predictable than asset price.

A. Yadav et al [18] was created a model using LSTM and Indian stock market. Then tuning was used to optimize models by comparing stateless and stateful models.

Y. Liu et al [19], establish a technique to predict stock price movement using knowledge graph based on financial news. This technique utilizes the event tuple properties. The feature selection models were used on the Thomson Reuters and Cable News Network. Experiments were conducted to demonstrate effectiveness of knowledge graph. A comparison of the accuracy with the same feature combinations for six stocks shows that method provides better performance. This work demonstrates the utility of knowledge graph in business.

M. Fabbri et al [20] provide a solution using deep RNN for trading with Dow Jones (DJIA). The result shows 50% higher accuracy than similar technique. The trading activities are performed based on predictions. For reliability of results, they used a long period of series. The solution has increased the initial capital. They tested this model by inverting the series of DJIA.

Md. R. H. Khan et al [21] performed sentiment of Bengali text to classify into happy or sad using different ML algorithms. In this context data is collected from different blogs and social media. Among different phases of analysis the preprocessing is one of the most complex parts. They use Count Vectorizer to tokenize the data and used six algorithms to predict. Results demonstrate Multinomial Naive Bayes provide us the maximum accuracy (86.67%).

S. Lee et al [22] provide a review on architecture and advancement in structured deep learning. They also explained applications and advantages. This work is beneficial for the early stage learners of deep learning.

R. Chiong et al [23] sentiment analysis-based technique is proposed for prediction of stock market. In pre-processing Sentiment analysis is done for finding relevant features. Historical stock market data and extracted features are used to train SVM, with parameter optimization using particle swarm optimization (PSO). Results demonstrate SVM-PSO-based technique is performing better than deep learning. The results are best in the literature.

S. D. Killekar et al [24] reviews available SMP techniques utilizing Regression and LSTM based ML for prediction of stock values. Main aspects for prediction are close, open, high, low and volume.

D. L. Minh et al [25] provide a method to predict stock prices directions using financial news and sentiment dictionary. That includes a proposal of a two-stream gated RNN and Stock2Vec based on word embedding based on Harvard IV-4 and financial news dataset. Two experiments are conducted (1) predicts S&P 500 index using historical prices and Reuters and Bloomberg articles, (2) predict price trends of VN-index using stock prices from cophieu68 and VietStock news. Results show: 1) GRU outperforms; 2) Stock2Vec is more efficient for financial datasets; 3) proves that model is effective for the stock prediction.

III. PROPOSED SYSTEM

Among different ML challenges the stock market price prediction is one of the most complicated tasks. Different real-world factors can influence the prices of stocks. A number of studies have been done in this area for making an accurate data model for stock market price prediction. According to the literature, we can categorize them into two categories: a method based on historical prices and trends, and the second type of method that is usages sentiment analysis and historical prices. In the first method, we can use the ability of ML techniques to analyze historical stock price data and recover patterns for prediction. In this task mostly supervised learning is used. But these techniques are not much accurate and fail in various conditions due to different dependencies of the stock market on real-world events.

Therefore, in the recent method of stock market prices prediction authors are trying to incorporate the sentiment analysis for improving the accuracy of prediction. However, this strategy is an effective technique and can enhance accuracy, but the available dataset for social media sentiment analysis is focused on language classification and another similar task. The incorporation of such a dataset with the stock market prices is not much effective and not able to improve the performance of the prediction accuracy. Therefore the proposed work is aimed to study and investigate the influence of social media data and key events which are having a high influence on stock market prices. Additionally, we are trying to measure the impact of involving social media information on the stock market prices.

Therefore the proposed work involves the following task for proposed investigative work.

1. Design a deep learning model which analyze the historical prices and predict next day close price
2. Design an intelligent model which usages historical stock market price data as well as the social media news for preparing a new dataset.
3. Utilize the newly organized dataset for training with the simple CNN model for performing training and price prediction.

A. Simple ML based prediction model

The aim of designing this simple ML based model is know the prediction performance of the ML algorithm by training only with the stock market price data. In this context the proposed model is demonstrated in figure 1.

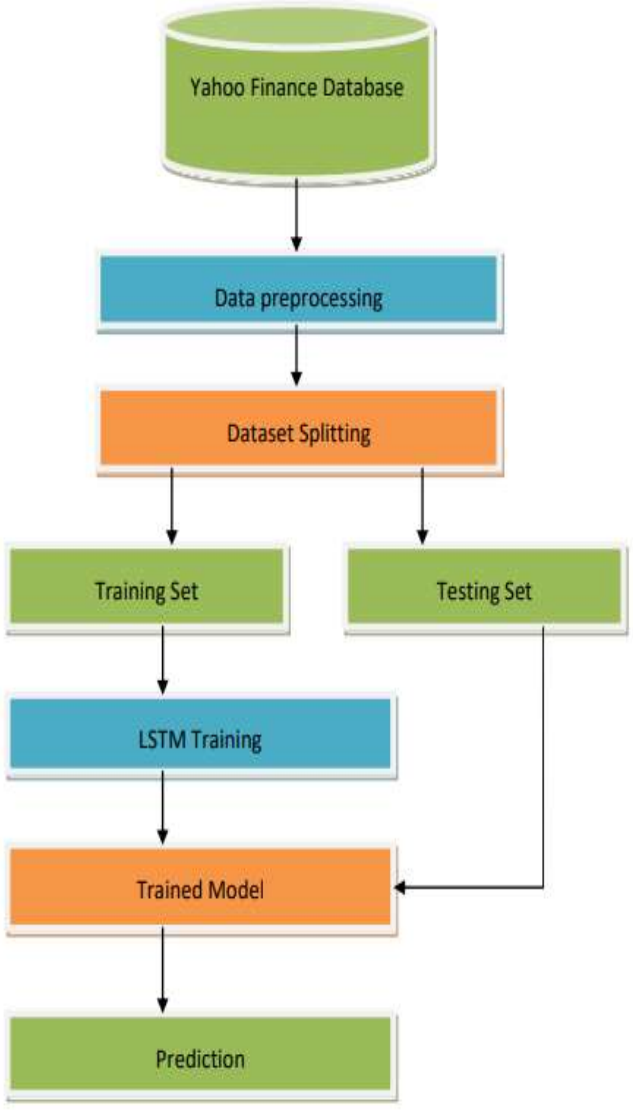


Figure 1 basic ML based stock market price prediction model

The yahoo finance database is a one of the most utilized data source by different researchers for stock market price prediction model deployment. That is a collection of historical price index for different listed companies. In this experiment we utilized historical prices of two popular banks in India namely State bank of India (SBI), and ICICI bank. We have extracted the stock price using yahoo finance API. The data consist of Date, Open, High, Low, Close, and Adj-Close. A raw data format of the SBI price index is demonstrated in figure 2.

Date	Open	High	Low	Close	Adj Close
2012-01-03	9.79	9.87	9.77	9.79	6.420575
2012-01-04	9.86	9.86	9.75	9.81	6.433694
2012-01-05	9.80	9.87	9.78	9.85	6.459926
2012-01-06	9.81	9.83	9.63	9.79	6.420575
2012-01-09	9.82	9.83	9.80	9.83	6.446809

Figure 2 raw data for SBI stock price

The obtained data is in raw format therefore we need to adopt some preprocessing techniques to enhance the quality of data. Therefore first we are trying to remove the data instances which have missing or null values. Additionally, we have used the min-max normalization for scaling the data between 0-1. Now we have divided the entire extracted data into two parts i.e. training and testing samples. The training samples consist of 70% of data and the test sample contains 30% of the entire data samples. Now we have trained the simple deep learning model namely LSTM using the training samples prepared. The configured LSTM model consists of 6 layers which consist of the following attributes:

Table 2 LSTM configuration

S. No.	Layers	Activation	Additional attribute
1	Input layer, LSTM	Relu	Return sequence “True”
2	LSTM	Relu	Return sequence “True”
3	LSTM	Relu	Return sequence “False”
4	Dropout		0.2
5	LSTM	Relu	Return sequence “True”
6	Dense		

The configured LSTM has trained on five different epochs for measuring the performance of the predictive algorithm i.e. 100, 300, 500, 800, and 1000 epoch. After training the model has validated using the 30% test samples to measure the performance of the prediction.

This section provides a base line model for studying the performance of predictive model using only the stock market price data. The next section involves the study of social media data influence on stock market price prediction and their dependency study.

B. Sentiment Based price prediction

The stock market is very sensitive against various real-world changes and events such as political changes, natural disasters, terrorist attacks, war conditions, and many more. Additionally, there are some specific factors that are directly related to the target companies such as new contracts, projects, stock, market, and others. Social media is one of the platforms where we get most of such kinds of NEWS. Therefore, in this experimental study, we proposed to develop an approach for searching and utilizing the news on social media to improve the stock market price prediction. In this section, we have explained two major things in order to develop a new dataset using multi-source information and utilized for stock market price prediction.

1. First we describe how sentiment score is computed
2. Second we explained how we combine it with the financial price index to be used to train the deep learning model.

The proposed task is initialized with a query database which is manually prepared. The queries are relevant to the bank name such as “state bank of India stock price”, “New Contracts of state bank of India” and similar others. These queries are utilized

as input to the Twin API for searching the information from Twitter. Basically, the Twin API is a scraper that is used to extract information from Twitter. The resultant information from the API is stored in a data structure for further processing. The news obtained from Twitter is in form of text (unstructured data). Here we have two different techniques are adopted for analyzing the twits in terms of sentiment scores.

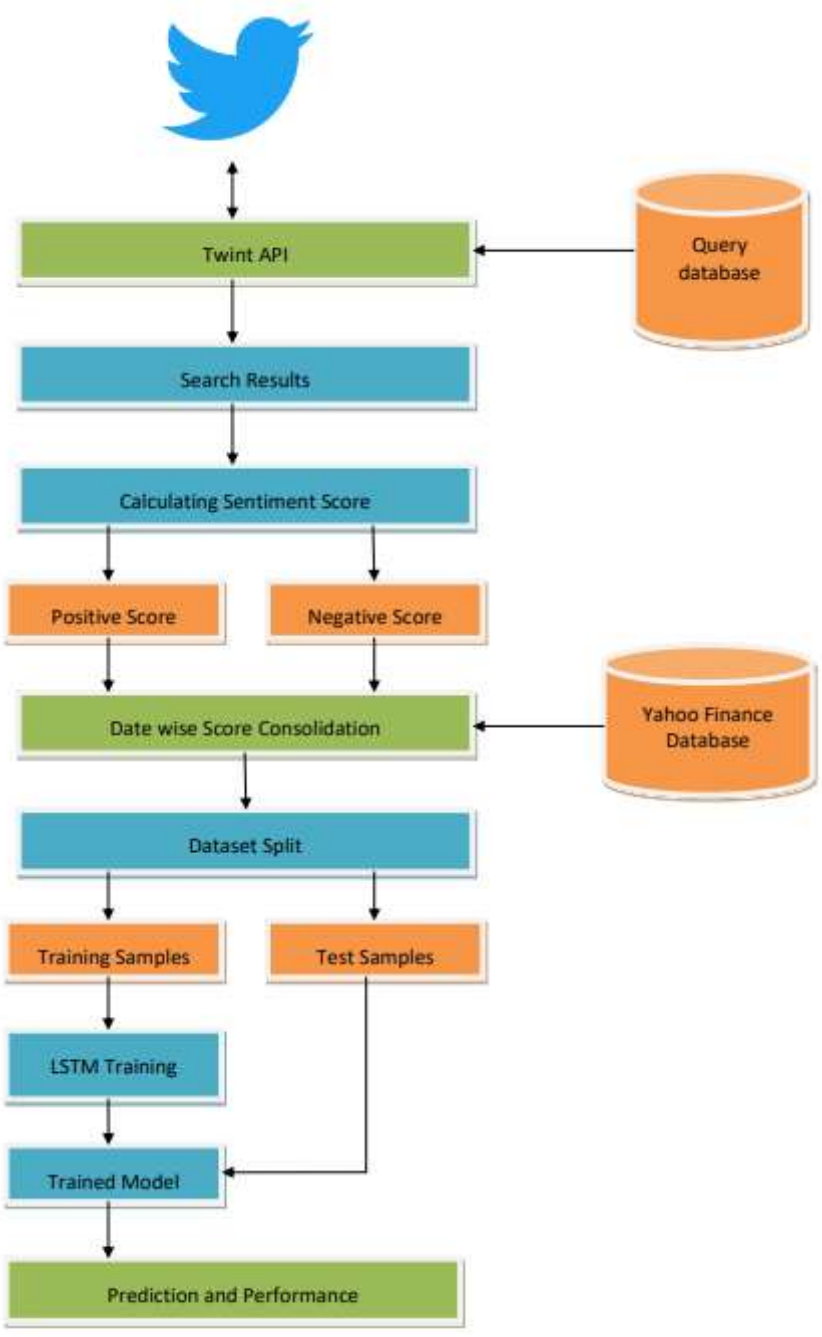


Figure 3 shows the proposed system to predict stock market price prediction using social media sentiments

The first method of evaluating the sentiment of the social media text is Sentiment Intensity Analyzer. That is also known as Valence Aware and sEntiment Reasoner (VADER). It is a rule-based sentiment analysis tool. This calculates sentiment into four classes i.e. positive, negative, neutral, and compound. A compound score is the sum of all words in the lexicon and returns the values between -1 and 1. According to their literature, it is not recommended to use text preprocessing techniques before the calculation of VADER. Therefore the obtained results from Twitter are used with the sentiment intensity analyzer and get the score for each tweet without any text processing. Next, we also utilized the traditional technique of sentiment

analysis therefore we pre-process the text to remove stop words and special characters. After text pre-processing, we used the preprocessed text with the NLP parser. That parser helps to parse the sentences in Part of Speech (POS). Using the POS tags we identify the essential keywords from the entire sentence and then use a third-party API SentiWordNet [2]. That is a lexical database that includes the attributes as given in table 3:

Table 3 SentiWordNet Database format

POS	ID	Positive score	Negative score	SysNetTerms

According to the SentiWordNet attributes, the database defines the cross POS relationships among the similar keywords therefore a same word the SentiWordNet contains the positive score as well as negative score too. Therefore, for each keyword in a twit contains S_p a positive score and S_n negative score. Thus for a same twit we calculate two different scores positive and negative such that.

$$T_p = \frac{1}{N}\sum_{i=1}^{N} S_p$$

And

$$T_n = \frac{1}{N}\sum_{i=1}^{N} S_n$$

Where, T_p and T_n is the positive and negative score of twit respectively, and N is the number of word in the twit.

However, we calculated the three types of sentiment scores for a single twit i.e. positive score, negative score, and compound sentiment intensity. But we need to aggregate the financial index and news sentiments, here the major problem is that for a single date financial index has only one entry but there are more than entries of social media news for a single date. Therefore we need to combine the scores of the different twit's sentiments into a single date. Therefore we utilize the mean of these scores according to the target date. Now for these three scores, we find the mean of the scores according to their dates. Thus the date wise sentiments scores are:

$$DS_p = \frac{1}{S}\sum_{i=1}^{S} {T_p}^i$$

$$DS_n = \frac{1}{S}\sum_{i=1}^{S} {T_n}^i$$

And

$$DC = \frac{1}{S}\sum_{i=1}^{S} C_i$$

Where, DS_p is the date wise consolidated positive sentiment, DS_n is date wise consolidated negative sentiment score, and DC is the compound sentiment intensity.

Now we combine the historical stock market price of companies and the computed sentiment scores. The combined dataset after combining the financial data and sentiment scores are final dataset is demonstrated using table 4.

Table 4 Final dataset for training

Date	Open	High	Low	Close	Adj-Close	DS_p	DS_n	DC

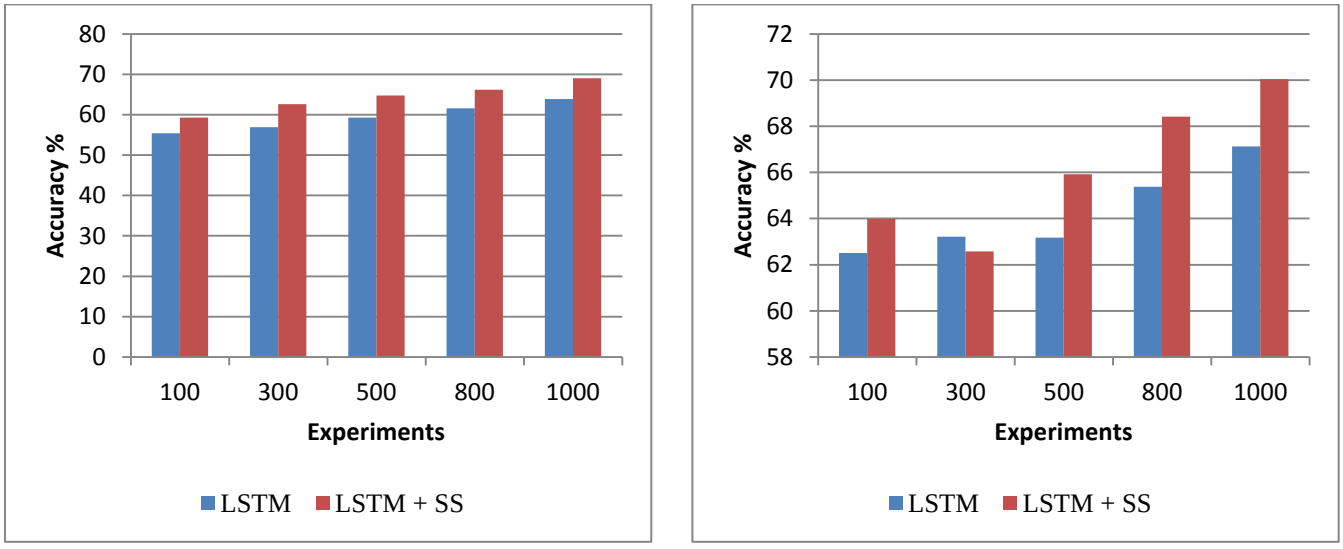
After preparing the combined dataset we are utilizing a similar LSTM model for performing the training. After training, we have tested out a model for their prediction with the 30% of test data with their sentiment scores. The obtained performance of the model and the impact of the extended dataset are described in the next section.

IV. RESULTS ANALYSIS

This section demonstrates how the performance of the machine learning algorithm for predicting the stock market price with the help of social media sentiment analysis. Therefore we first demonstrate the accuracy of models. However, the LSTM does not provide performance in terms of accuracy but it is an essential parameter for analyzing the performance of the ML models. The accuracy demonstrates how accurately a prediction algorithm does. To calculate the accuracy we use the following equation.

$$Accuracy(\%) = 100 - \left(\frac{|actual\ value - predicted|}{|actual\ value|} X100\right)$$

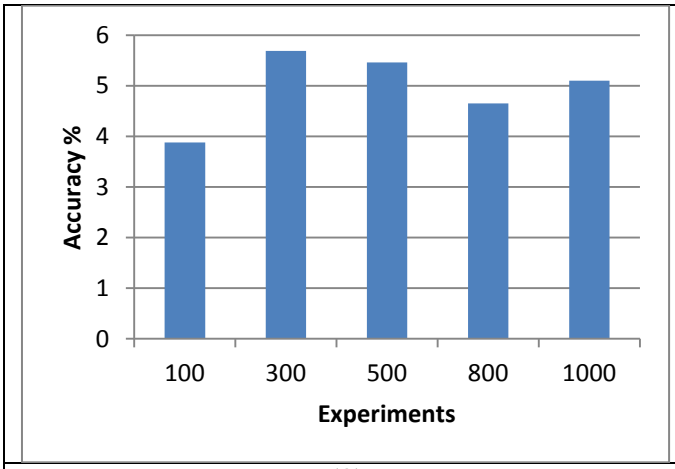
The comparative accuracy of both the techniques i.e. sentiment score-based and without sentiment analysis is demonstrated in Figure 3. The performance of both the models has been compared with the increasing number of epochs for two Indian banking companies namely SBI and ICICI. The accuracy of SBI based price prediction is given in figure 3(A), and 3(B) contains the performance of ICICI bank.



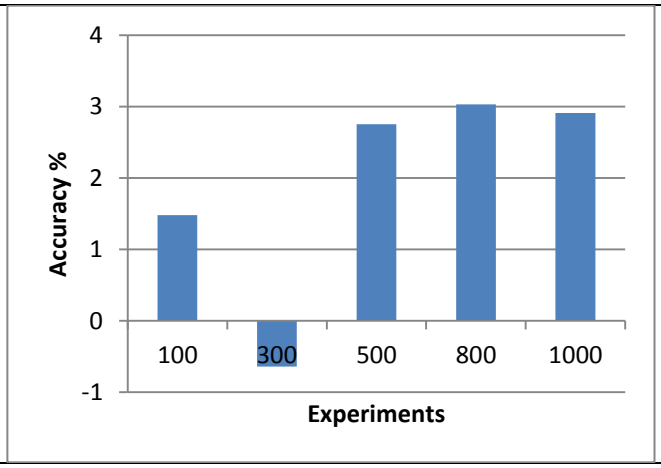
(A) (B)

Figure 3 Comparative Accuracy with increasing number of each using both kinds of models for (A) SBI price prediction and (B) ICICI bank price prediction

According to the obtained performance, the accuracy of the proposed approach with SBI stock price is much more consistent and improving with the increase of epoch cycles. On the other hand, ICICI Bank's stock price is sometimes increasing and decreasing due to the availability of news data over social media. but according to overall performance, the proposed approach improves the performance of prediction accuracy. In order to understand the amount of performance improvement, we have also measured the performance improvement of the algorithm by finding the difference among both models' accuracy we prepared figure 4. Figure 4(A) demonstrates the performance improvement of the model for SBI stock price and figure 4(B) demonstrates the performance influence when we apply the proposed model with the ICICI stocks. According to an experimental performance study we have found the proposed model can improve the prediction accuracy up to 3-5% from the traditionally applied model. However, there is a significant improvement in prediction accuracy is observed but there is some space to improve the performance of the model more effectively.



(A)



(B)

Figure 4 performance influence of the sentiment analysis over the traditional method of stock market price prediction

V. CONCLUSION AND FUTURE WORK

Machine learning provides us the ability to deal with the complex real-world problem. One of such real world problem is the prediction of stock market price index. There are a number of solutions are developed in order to deal with the stock market price prediction using classical ML techniques as well as new deep learning based techniques. Additionally some of them are only considering the historical price data and some of the works are also considering the influencing facts such as social media sentiments and news articles. However, we found a great improvement on prediction with the social media sentiments on stock market price but most of the available sentiment analysis datasets are not much effective for improving the prediction performance.

Therefore in this paper we have proposed to incorporate the deep learning technique and social media data (specifically based on target stock) for producing a new dataset for stock market price prediction. The new dataset has incorporated the historical price index, sentiment intensity analyzer based compound sentiment score, and SentiWordNet based positive and negative sentiment scores. This data is used with the popular deep learning model namely LSTM for training and prediction. In addition, for measuring the influence of the sentiment analysis on the traditional methods of stock market price prediction model we have trained the same model with only the historical price data. The performance of both the models are measured in terms of accuracy and compared for finding the improvement after involving the sentiment analysis.

The experimental analysis indicates there is 3-5% positive improvement in prediction accuracy. But during investigation we have found there is more probability to improve the prediction. We have found some critical research gaps for more improvement in stock market price prediction for our future work:

1. Social media NEWS is not much authentic thus we need more accurate source of NEWS by which the fluctuations on stock market can be measured more precisely.
2. We need to monitor the most recent news for making more accurate decisions
3. The news sentiments are helpful for accurate prediction but there are significant variations during various positive and negative events

REFERENCES

[1] T. H. Nguyen, K. Shirai, J. Velcin, "Sentiment Analysis on Social Media for Stock Movement Prediction", Expert Systems With Applications Volume 42, PP. 9603–9611, 2015.

[2] Binita Verma, Ramjeevan Singh Thakur, Shailesh Jaloree, "Sentiment Analysis using Lexicon and Machine Learning Based Approaches: A Survey", Proceedings of International Conference on Recent Advancement on Computer and Communication. Lecture Notes in Networks and Systems, Springer, Singapore, pp. 441-447, 2017.

[3] S. L. Pandhripande, A. Dixit, "Prediction of 2 Scrip Listed in NSE using Artificial Neural Network", International Journal of Computer Applications (IJCA), Volume 134, No.2, 2016.

[4] P. Sobkowicz, M. Kaschesky, G. Bouchard, "Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web", Government Information Quarterly, Volume 29, PP. 470–479, 2012.

[5] Mrs. K. S. Mahajan, R. V. Kulkarni, "A Review: Application of Data mining Tools for Stock Market", International Journal Computer Technology & Applications, Volume 4,No. 1, PP. 19-27, 2013.

[6] S. M. Cart, S. Consoli, L. Piras, A. S. Podda, D. R. Recupero, "explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting", IEEE Access, VOLUME 9, 2021

[7] M. Chiny, O. Bencharef, M. Y. Hadi, Y. Chihab, "A Client-Centric Evaluation System to Evaluate Guest's Satisfaction on Airbnb Using Machine Learning and NLP", Hindawi Applied Computational Intelligence and Soft Computing Volume 2021, Article ID 6675790, 14 pages

[8] Q. Bi, H. Yan, C. Chen, Q. Su, "An Integrated Machine Learning Framework for Stock Price Prediction", c Springer Nature Switzerland AG 2020, CCIR 2020, LNCS 12285, pp. 99–110, 2020.

[9] R. Batra, S. M. Daudpota, "Integrating StockTwits with Sentiment Analysis for better Prediction of Stock Price Movement", 2018 International Conference on Computing, Mathematics and Engineering Technologies – iCoMET 2018, 978-1-5386-1370-2/18/$31.00 ©2018 IEEE

[10] M. Vicari, M. Gaspari, "Analysis of news sentiments using natural language processing and deep learning", AI & SOCIETY (2021) 36:931–937

[11] J. S. Yang, C. Y. Zhao, H. T. Yu, H. Y. Chen, "Use GBDT to Predict the Stock Market", Procedia Computer Science 174 (2020) 161–171

[12] H. Vachhani, M. S. Obiadat, A. Thakkar, V. Shah, R. Sojitra, J. Bhatia, S. Tanwar, "Machine Learning Based Stock Market Analysis: A Short Survey", Springer Nature Switzerland AG 2020, ICIDCA 2019, LNDECT 46, pp. 12–26, 2020

[13] M. Alshahrani, F. Zhu, A. Sameh, L. Zheng, S. Mumtaz, "Evaluating the Influence of Twitter on the Saudi Arabian Stock Market Indicators", © Springer International Publishing AG, part of Springer Nature 2018, 5th International Symposium on Data Mining Applications, pp. 113–132, 2018.

[14] M. Zulqarnain, R. Ghazali, M. G. Ghouse, Y. M. M. Hassim, I. Javid, "Predicting Financial Prices of Stock Market using Recurrent Convolutional Neural Networks", I.J. Intelligent Systems and Applications, 2020, 6, 21-32

[15] G. Liu, X. Wang, "A Numerical-Based Attention Method for Stock Market Prediction With Dual Information", IEEE Access, VOLUME 7, 2019

[16] W. Souma, I. Vodenska, H. Aoyama, "Enhanced news sentiment analysis using deep learning methods", Journal of Computational Social Science (2019) 2:33–46

[17] A. Atkins, M. Niranjan, E. Gerding, "Financial news predicts stock market volatility better than close price", The Journal of Finance and Data Science 4 (2018) 120e137

[18] A. Yadav, C. K. Jha, A. Sharan, "Optimizing LSTM for time series prediction in Indian stock market", Procedia Computer Science 167 (2020) 2091–2100

[19] Y. Liu, Q. Zeng, J. O. Meré, H. Yang, "Anticipating Stock Market of the Renowned Companies: A Knowledge Graph Approach", Hindawi Complexity Volume 2019, Article ID 9202457, 15 pages

[20] M. Fabbri, G. Moro, "Dow Jones Trading with Deep Learning: The Unreasonable Effectiveness of Recurrent Neural Networks", In Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018), pages 142-153

[21] Md. R. H. Khan, U. S. Afroz, A. K. M. Masum, S. Abujar, S. A. Hossain, "Sentiment Analysis from Bengali Depression Dataset using Machine Learning", IEEE – 49239

[22] S. Lee, "Deep Structured Learning: Architectures and Applications", International Journal of Advanced Culture Technology Vol.6 No.4 262-265 (2018)

[23] R. Chiong, Z. Fan, Z. Hu, M. T. P. Adam, B. Lutz, D. Neumann, "A sentiment analysis-based machine learning approach for financial market prediction via news disclosures", GECCO'18 Companion, July 15–19, 2018, Kyoto, Japan, © 2018 Copyright held by the owner/author(s). ACM ISBN 978-1-4503-5764-7/18/07

[24] S. D. Killekar, Dr. S. S. Sannakki, P. Y. Niranjan, G. R. Deshpande, "An Analytical Approach for Stock Market Forecasting Based on Machine Learning", International Journal of Scientific Research in Science, Engineering and Technology, 2020, Volume 7, Issue 2

[25] D. L. Minh, A. S. Niaraki, H. D. Huy, K. Min, H. Moon, "Deep Learning Approach for Short-Term Stock Trends Prediction Based on Two-Stream Gated Recurrent Unit Network", IEEE access, VOLUME 6, 2018