2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) | 979-8-3503-9737-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACCS57279.2023.10112768

Stock Market Prediction Using Sentiment Analysis and Incremental Clustering Approaches

[1] Leela Satish Parvatha, [2] Devu Naga Veera Tarun, [3] Mentem Yeswanth, [4]Jonnalagadda. Surya Kiran

[1,2,3,4] Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation,
Vaddeswaram, A.P., India-522502.

[1]satishparvatha@gmail.com, [2]tarunnagveer24@gmail.com, [3]mentemyeswanth@gmail.com, [4]kiransurya93@kluniversity.in

***Abstract*—Sentimental analysis is one of the techniques of Natural Language Processing. It helps us to determine the polarity of the given data: Negative, Positive or Neutral. We know that predicting the prices of Stocks are hard as they are Unstable. Sentimental Analysis helps us to predict the prices of Stocks. The stock prices can be predicted by taking raw data from various social media platforms and these are converted into valuable data by using Sentimental analysis. We observed that we can get more accuracy when we supply more data. DB Scan is an algorithm which is used to form clusters dynamically with the help of eps value. The number of clusters formed will be dependent on the eps value given to the algorithm. And comparing the results of different algorithms tells us which algorithm is best for stock price prediction. Developing the DBSCAN from scratch will gives the flexibility to change the algorithms accordingly and also, we can use the values like centroids in the model. and these results are compared with the algorithms like KMEANS, LSTM, CNN, with the parameters like MAE, MSE, RMSE.**

eps = epsilon, hyperparameter that controls cluster size

***Keywords*—Sentiment analysis, k-means clustering, LSTM, GRU, CNN, RMSE, MSE, MAE.**

I. Introduction

The main aim of stock price prediction is to determine the future movement of the trends and the future value of stocks of a financial exchange. A good accurate prediction of stocks will give financial investors the flexibility to make decisions accordingly. The prediction of stock market will give huge profit and will be and gives a motivation for research in that area. Stock prices predictions can also be done using sentimental analysis where we will gather data from the tweets and comments of Facebook and using that we will calculate sentiments of the data. We'll gather data from different kinds of sources and apply various types of ml and AI algorithms to predict the stock price.

As we know that Stock market is known for being nonlinear and Dynamic. So, it's extremely challenging to predict the stocks accurately because of multiple factors such as economic changes which will occur globally and also company's financial performance. By using Machine learning techniques, we can find several strategies which can be implemented to predict the stock accurately. People generally predict the stocks by previous stock results and use social media sites like twitter and Facebook. As people use these networking sites to share their opinions and their views on a topic it's easy to use their information available in those platforms and it can be used as user feedback and opinion about the specific topic or any product. These opinions can be used by Business analysts to predict further.

II. Literature Survey

A. *Title: Applying LSTM for Stock Price Prediction with Sentiment Analysis*

Authors: Alexandre Heiden, BrazilRafael Stubs Parpinell

Objective: For prediction the stock price using sentimental analysis he used Vander sentiment analysis model to feed the sentiments as a feature in to LSTM model. We collected the data from New York Times and used Vander model on that and predicted sentimental analysis and supplied LSTM as a feature.

Limitations: It has shown good results. But still there is chance of improving the accuracy if we supply more data to the model [1].

B. *Title: Feasibility Study of Stock Market Prediction for Sentiment Analysis using Artificial Intelligence*

Author: Surbhi Soni, Ashok Kumar Shirvastava, Deepak Motwani

Objective: They have given a brief explanation of applications of ML. AI and sentimental analysis and stock price prediction using LSTM Arima and RNN. A financial system should understand ai in order to form a perfect model. It is mandatory to follow rules and principles which will follow the procedure of AI. Using AI, ANN we can get good accuracy [2].

C. *Title: Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks*

Authors: Sai Vikram Kolasani, Rida Assaf.

Objective: We use stock data which contains aapl keywords and used two models multilayer perceptron and boosted regression. First, we collected the data from New York Times and used vander model on that and predicted sentimental analysis and supplied LSTM as a feature and along with that we have supplied historical data of assets.

Limitations: We want to use this data on different markets. If we supply more data, we will get more results [3].

979-8-3503-9737-6/23/$31.00 ©2023 IEEE

D. *Title: The Stock Prediction using Machine Learning Techniques: A Comprehensive and Systematic Literature Review*

Authors: Wiranata, R. B, Djunaidy.

Objective: There are around 81 studies which were investigated. It also provides information methods and those are combined and performed to get better results. The 9 methods which are best among 48 methods used for performing stock prediction model as they give good accuracy and very less error rate such as LSTM, RNN, CNN, RF, XGBoost.

Limitations: We want to perform with combination of input dataset types. And also to improve performance various machine learning techniques need to be combined [4].

E. *Title: Sentiment analysis and prediction of Indian stock market amid Covid-19 pandemic*

Authors: Gondaliya, C, Patel, A, Shah, T

Objective: As they compared with six different ml algorithms, they observed that logistic regression and svm produced more better compared to the other models as they compared with six different ml algorithms, they observed that logistic regression and svm produced more better compared to the other models.

Limitations: We can expand this to longer period of data and other type of techniques [5].

F. *Title: Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19*

Authors: Mujahid, M., Lee, E, Rustam, F, Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf.

Objective: In this paper we have taken tweets about e learning and identified the. The sentiments of people using Text blob, vader, sentiword, and word polarity. Two feature techniques like used for classifying the data into neutral, positive and negative and DT, SVM, and RF achieves the accuracy of 0.95using Bow and SMOTHE. Vader and text blob are also used for performance comparison [6].

G. *Title: Sentiment Analysis as A Factor Included In The Forecasts Of Price Changes In The Stock Exchange*

Authors: Wojarnik, G.

Objective: We will create variables with the help of classifiers which are used to improve the model accuracy and these clusters are created using technical analysis indicators. This research showed us the data from the discussion forums can feed a mechanism to predict the performance for the companies.

Limitations: It tried to correlate the data from different sources of texts with the different sources of texts [7].

H. *Title: Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation*

Authors: Kyo-Joong Oh, Dongkun Lee, Byungsoo KO

Objective: The prediction was performed based on high demand stocks. These are represented in Candlestick chart images for better classification. Here the proposed joint network achieved good results in predicting the stock. It gave 75.38% accuracy for data set of stock with data of ten day's time period.

Limitations: We would also want to use enhanced techniques to get better performance while predicting stock price over shorter periods of time [8].

I. *Title: Prediction of stock values changes using sentiment analysis of stock news headlines*

Authors: László Nemes & Attila Kiss.

Objective: In this paper they have used economic news headlines and uses different tools like Bert, Vander, Text blob, RNN to compare the results with the stock in that same period they compared the various sentimental analysis with Bert as a bench mark.in that comparison the RNN model outshined various other tools. The other tools gave results quite close to the Bert model.

Limitations: Include new tools and features to this model [9].

J. *Title: Sentiment Analysis in social media and Its Application: Systematic Literature Review*

Authors: Zulfadzli Drus, Haliyana Khalid.

Objective: In this paper the social media data which is the raw data of the social media platforms is transformed into useful information by using sentiment analysis using naive bayes, lexicon method. In this paper the study demonstrated that the sentiment analysis in social media. Many papers used twitter tweets as their social media context.

Limitations: More investigation is needed to develop the universal model of sentiment analysis that is which helps in the exploring data [10].

K. *Title: Sentiment analysis of financial news using unsupervised approach*

Authors: Anita Yadava, C K Jhaa, Aditi Sharanb, Vikrant Vaishbad.

Objective: The sentimental analysis of financial news is done with unsupervised learning approach. The framework is proposed and has been used to make sentiments in financial text more strong. both hybrid approach and also noun-verb approach gave more promising accuracies when compared to Turney's approach.

Limitations: The unsupervised approaches which gave promising results can be used for building a real time model [11].

L. *Title: Stock Market Data Prediction Using Machine Learning Techniques*

Authors: Torres P, E. P., Hernández-Álvarez, M., Torres Hernández, E. A., & Yoo, S. G.

Objective: Random forests and multilayer perceptron algorithms are used in order to predict closing prices using Apple Inc. An accuracy analysis was also performed to see how accurate supervised machine learning algorithms in the financial area. The current study has demonstrated how artificial intelligence approaches, notably machine learning algorithms, used to foresee and predict future of stock market data. Stock market data more open. This would be a significant advancement for the stock market [12].

M. Title: Stock Price Prediction Using News sentiment Analysis

Authors: Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia1and David C. Anastasiu

Objective: The accuracy was improved as large amount of time series data Cloud computing was used for Training Prediction models. Prediction models are build using ARIMA, RNN, FACEBOOK PROPHET. Promising results were found in prediction model which is based on RNN.

Limitations: We want to implement a domain specific model it can be made by grouping companies based on their sector [13].

N. Title: Stock market prediction analysis by incorporating social and news opinion and sentiment news opinion and sentiment

Authors: Zhaoxia ,Seng-Beng HO Zhiping LIN

Objective: The stock price is predicted by observing the relation between the stock market trends and the sentiments of the news. Mean square error (MSE) is reduced hence performance is improved. To optimize performance various algorithms were applied. Among these 4 algorithms Trainlm gave the best performance for prediction model.

Limitations: We want work on the sentiments of the effects the prices of stock. Also Sentiment Effect Parameters are to be selected in future [14].

O. Title: Stock trend prediction using setimental analysis.

Authors: Kalyani Joshi, Bharathi H. N, Prof. Jyothi Rao

Objective: News articles have impact on stockRF, SVM, naïve bayes are performed in testing. Own dictionary was created for removing stop words related to stocks and finance. RF, SVM, Naïve bayes are performed while testing where RF, SVM gave better results than Naïve bayes. This prediction model gave more than 80% accuracy.

Limitations: To implement this by adding more data of company. We would like to use Twitter data for analysis and also include algorithmic trading strategies [15].

P. Title: Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty

Authors: Aditya Bhardwaja, Yogendra Narayanb, Vanrajc, Pawana, Maitreyee Duttaaa.

Objective: The indicators of stock market like nifty and Sensex are used to predict the trends of stocks and sentimental analysis is used to fetch the opinion from the users.Sensex and niftylive values are taken from the servers with different time intervals and that data is used to predict the stock market values using the sentimental analysis.

Limitations: Used live servers data from Sensex and nifty without advanced options in the python script [16].

III. METHODOLOGY

Our system consists of three phases.

a. Clustering and Classification (incremental)

b. K means Clustering and Classification

c. LSTM vs GRU vs CNN comparison

A. Clustering and Classification (incremental)

We've used a DBSCAN algorithm for forming clusters. It is an algorithm which will group the points that are together or the points which are close to each other which uses the Euclid distance or a minimal distance as parameter.

(DBSCAN) is an algorithm which is based on unsupervised learning. As a first step it identifies the clusters which are distinct in the data. Based on the parameter high density and lower density. We need to give eps value and min clusters for forming the clusters. After that these clusters we be supplied as target variable for LSTM algorithm. Using LSTM we've done classification.

Parameters of DBSCAN algorithm

1)*MinPts:* Minimum number of points that are required to consider them as a region.
2)*Eps (ε):* The distance from one point to the other in that neighbourhood.
3)*Core:* A point from a distance from it tone has at least m points.
4)*Noise:* If a point neither belong to core nor a border and if it has lower than m points then it is called noise.

B. K means Clustering and Classification

It is part of unsupervised learning, it groups the data which is unlabeled and forms the clusters for the data in this algorithm the k indicates the number of clusters it should form if k=1 it forms one cluster. If it is 2 it forms two clusters like that it arranges the data into different clusters. It allows the data to arrange into different groups and without the. Training of the data we can identify the orders of group. The algorithm is based on its centre, in this approach each cluster is linked with the centroid.

The main motive of the algorithm is to decrease the sum of distance between its corresponding points. An iterative step by step process is used to determine the best value for k centroids. Each data will be assigned to its nearest c. And these points which are nearer to the centre will form a cluster. After that these clusters we be supplied as target variable for LSTM algorithm. Using LSTM we have done classification. After that these clusters we be supplied as

target variable for LSTM algorithm. Using LSTM we have done classification.

C. LSTM, CNN , GRU

1)*LSTM:* Long short-term memory will come under artificial neural networks which is also a application in machine learning. Like all other networks it also contains a feed forward networks. It also process sequence of data as well as sequence of data.

2)*CNN:* CNN A Convolution Neural Network can also be called as convent. It is used in processing the data which contains grid structure which is similar like an image. A digital image can also be called as double representation of visualized data. It also contains a series of pixels placed in a grid structure such like fashion that contain pixels value to note what color each pixel should be and how bright the pixel.

3)*GRU:* A gated recurrent unit comes under recurrent neural network. Which uses the connections from a node which are in a sequence for performing the tasks which are associated with it and also clustering. Block diagram for model is shown in Fig. 1.

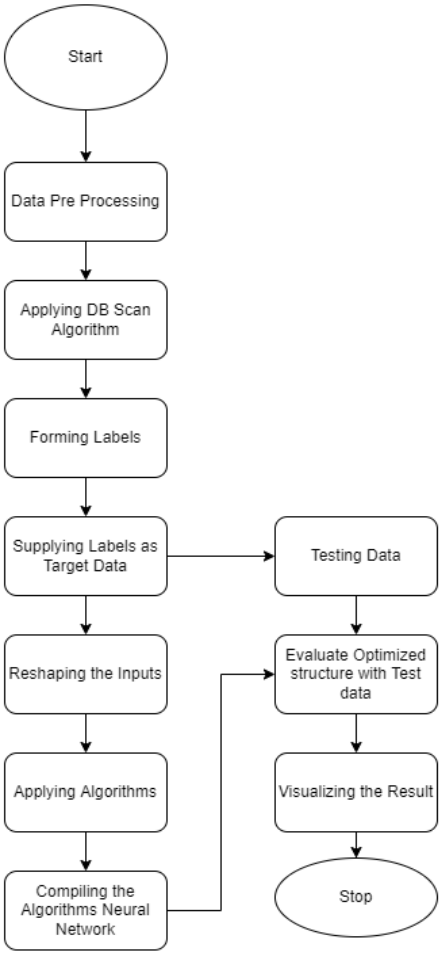

Fig. 1. Block diagram for model

D. Workflow Of KMEANS Algorithm

Step 1: Importing the necessary libraries like numpy, pandas etc… to execute the program

Step 2: importing and reading the AAPL dataset which contains the daily stocks details

Step 3: pre-processing techniques to be applied to delete unwanted and null values

Step 4: Apply kmeans algorithm by taking closing stocks as a target value by giving max clusters as 3

Step 5: After applying the algorithm now supply the formed clusters as a target variable

Step 6: Now use LSTM algorithm for classifying the clusters

Step 7: Result analysis using mse, mea, mse as parameters.

E. Workflow Between Algorithms

Step 1: Importing the necessary libraries like to execute the program

Step 2: importing and reading the dataset which contains the daily stocks details

Step 3: pre-processing techniques to be applied to delete unwanted and null values

Step 4: Now use LSTM algorithm for classifying the data

Step 5: Result analysis using rmse, mea, mse as parameters

Step 6: Now use cnn algorithm for classifying the data

Step 7: Result analysis using rmse, mea, mse as parameters

Step 8: Now use gru algorithm for classifying the data

Step 9: Result analysis using rmse, mea, mse as parameters

IV. IMPLEMENTATION

Python: Libraries, Tools

Spyder is an open source Development Environment for Python, it is a free (IDE) which also contains Anaconda

Python with the version of 3.8 is used for implementing the code and have used the kmeans library which is predefined library. Python and generally used for clustering data. And also we have used DBSCAN for clustering which is used to form the clusters dynamically. Imported the necessary libraries for the model isshown in Fig. 2. Imported aapl dataset and called DBSCAN function with eps value and dataset as arguments is shown in Fig. 3. Used LSTM model for prediction is shown in Fig. 4. Used Gru model for prediction is shown in Fig. 5.

```
import numpy as np
import pandas as pd
import random
```

Fig. 2. Imported the necessary libraries for the model

```
d=pd.read_csv('/content/AAPL.csv')

[ ] df=d['Open']

[ ] k=cluster_with_stack(1.5,4,df)
```

Fig. 3. Imported aapl dataset and called DBSCAN function with eps value and dataset as arguments

```
[ ] model=Sequential()
    model.add(LSTM(50,return_sequences=True,input_shape=(100,1)))
    model.add(LSTM(50,return_sequences=True))
    model.add(LSTM(50))
    model.add(Dense(1))
    model.compile(loss='mean_squared_error',optimizer='adam')
```

Fig. 4. Used LSTM model for prediction

```
[ ] import tensorflow as tf
    from tensorflow.keras.layers import LSTM, GRU ,Dropout

    tf.keras.backend.clear_session()
    model=Sequential()
    model.add(GRU(32,return_sequences=True,input_shape=(time_step,1)))
    model.add(GRU(32,return_sequences=True))
    model.add(GRU(32))
    model.add(Dropout(0.20))
    model.add(Dense(1))
    model.compile(loss='mean_squared_error',optimizer='adam')
```

Fig. 5. Used Gru model for prediction.

V. Experimental Set Up And Results

A. Dataset

The dataset (aapl) was taken from Kaggle website which contains the information of daily stocks.

B. Experiment Results

Below are the screen shots of our experiment results

The results of LSTM algorithm using DBSCAN clustering is shown in Fig. 6. The results of LSTM algorithm using kmeans clustering is shown in Fig. 7. The results of LSTM without classification is shown in Fig. 8. The results of Gru without classification is shown in Fig. 9. The results of cnn without classification is shown in Fig. 10. The comparison between various algorithms results performed earlier is shown in Fig. 11. Graph of different methods is shown in Fig. 12. Graph between eps value and clusters formed is shown in Fig. 13.

```
Train data RMSE:  0.23901811254249553
Train data MSE:  0.05712965812337706
Train data MAE:  0.0837692795695129
```

Fig. 6. Indicates the results of LSTM algorithm using DBSCAN clustering. In this we have given eps as 1.5 which forms four clusters

```
Train data RMSE:  0.3368307426987388
Train data MSE:  0.11345494922698399
Train data MAE:  0.3345916628146129
```

Fig. 7. Indicates the results of LSTM algorithm using kmeans clustering .In this we have given max clusters as 5

```
Train data RMSE:  8.58319269407566
Train data MSE:  73.67119682363378
Train data MAE:  3.6485560872708134
```

Fig. 8. Indicates the results of LSTM without classification

```
Train data RMSE:  8.618053502828117
Train data MSE:  74.27084617760798
Train data MAE:  3.8059277465431585
```

Fig. 9. Indicates the results of Gru without classification

```
Train data RMSE:  10.474963607387632
Train data MSE:  109.72486257609532
Train data MAE:  4.313637202694393
```

Fig. 10. Indicates the results of cnn without classification

	DBSCAN	Kmeans	LSTM	Gru	Cnn
RMSE	0.23901	0.3368	8.58	8.61	10.474
MSE	0.0571	0.11	73.67	74.27	109.724
MEA	0.083	0.33459	3.6485	3.805	4.313

Fig. 11. Indicates the comparison between various algorithms results performed earlier

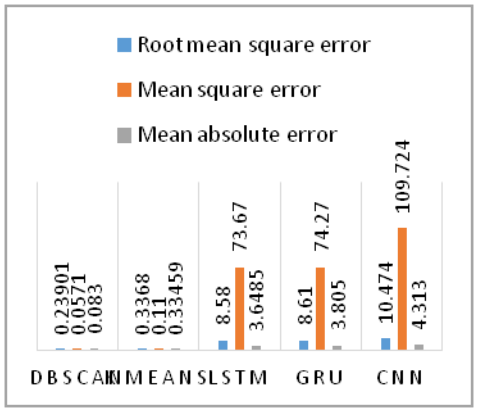

Fig. 12. Graph of different methods

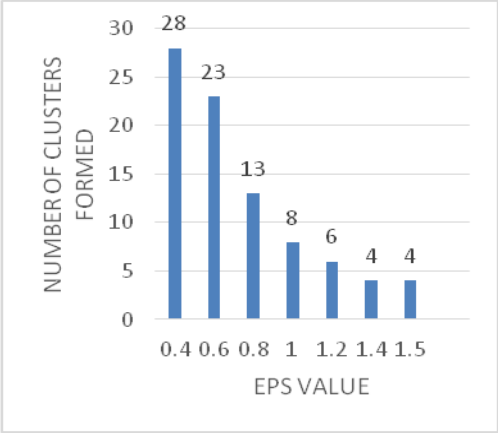

Fig. 13. Graph between eps value and clusters formed

VI. Conclusion

Stock price prediction will be useful to predict the stock price of the company and other financial aspects. Here the concept of using sentimental analysis helps to predict the stock .if the value of sentiment is negative it indicates the fall of stock price or else if the value of sentiment is positive it indicates the rise of stock price, and then stock price may will down. DB Scan is used for creating the clusters and classifying those clusters and also using Kmeans clustering the clusters are classified and also they are classified without clustering by using LSTM,GRU,CNN. All the results are compared and DBSCAN gives better results.

Including market data as well as textual data from will gives chance to a better prediction accuracies, Further work includes a hybrid model which can be made of combining two networks such as(ex:- ARIMA,AR).

References

[1] Heiden, A., & Parpinelli, R. S. (2021). Applying LSTM for stock price prediction with sentiment analysis. Anais do 15. Congresso Brasileiro de Inteligência Computacional. https://doi.org/10.21528/cbic2021-45

[2] Surbhi Soni, Ashok Kumar Shirvastava, & Deepak Motwani. (n.d.). Feasibility Study of Stock Market Prediction for Sentiment Analysis using Artificial Intelligence . Retrieved November24, 2022, from https://www.amity.edu/gwalior/journal/papers/vol4-issue1- 3.pdf.

[3] Kolasani, S. V., & Assaf, R. (2020). Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks. In Journal of Data Analysis and Information Processing (Vol. 08, Issue 04, pp. 309–319). Scientific Research Publishing, Inc. https://doi.org/10.4236/jdaip.2020.84018.

[4] Wiranata, R. B., & Djunaidy, A. (2021). The stock exchange prediction using machinelearning techniques: A comprehensive and systematic literature review. Jurnal IlmuKomputer dan Informasi, 14(2), 91-112. https://doi.org/10.21609/jiki.v14i2.935

[5] Gondaliya, C., Patel, A., & Shah, T. (2021). Sentiment analysis and prediction of Indianstock market amid COVID-19 pandemic. IOP Conference Series: Materials Science and Engineering, 1020(1), 012023. https://doi.org/10.1088/1757-899x/1020/1/012023

[6] Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf, (2021). Sentiment analysis and topic modeling on tweets about online education during COVID-19. Applied Sciences, 11(18), 8438. https://doi.org/10.3390/app11188438

[7] Wojarnik, G. (2021). Sentiment analysis as a factor included in the forecasts of price changes in the stock exchange. Procedia Computer Science, 192, 3176-3183. https://doi.org/10.1016/j.procs.2021.09.090

[8] Ho, T., & Huang, Y. (2021). Stock price movement prediction using sentiment analysis and candlestick chart representation. Sensors, 21(23), 7957. https://doi.org/10.3390/s21237957

[9] László Nemes & Attila Kiss (2021) Prediction of stock values changes usingsentiment analysis of stock news headlines, Journal of Information and Telecommunication, 5:3,375-394, DOI: 10.1080/24751839.2021.1874252

[10] Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application:Systematic literature review. Procedia Computer Science, 161, 707-714. https://doi.org/10.1016/j.procs.2019.11.174

[11] Yadav, A., Jha, C. K., Sharan, A., & Vaish, V. (2020). Sentiment analysis of financialnews using unsupervised approach. Procedia Computer Science, 167, 589-598. https://doi.org/10.1016/j.procs.2020.03.325

[12] Torres P, E. P., Hernández-Álvarez, M., Torres Hernández, E. A., & Yoo, S. G. (2019,February). Stock market data prediction using machine learning techniques. In International conference on information technology & systems (pp. 539-547). Springer, Cham. DOI: 10.1007/978-3-030-11890-7_52

[13] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 205-208, doi: 10.1109/BigDataService.2019.00035.

[14] Z. Wang, S. -B. Ho and Z. Lin, "Stock Market Prediction Analysis by Incorporating Social and News Opinion and Sentiment," 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018, pp. 1375-1380, doi: 10.1109/ICDMW.2018.00195.

[15] kalyani joshi, et al. STOCK TREND PREDICTION USING NEWS SENTIMENT ANALSIS. arxiv.org/ftp/arxiv/papers/1607/1607.01958.

[16] Aditya Bhardwaj, Yogendra Narayan, Vanraj, Pawan, Maitreyee Dutta,Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty,Procedia Computer Science,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2015.10.043.