Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 4, April 2024

Stock Market Prediction Based on Stock Data and News Sentiment

Mr. Sourav Dutta
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Dr. Chandra Das
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Mr. Sohom Saha
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Ms. Disha Biswas
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Ms. Ankana Das
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Ms. Shrestha Banerjee
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Dr. Shilpi Bose
Dept. of Computer science and Engineering
Netaji Subhash Engineering College
Kolkata, India

***Abstract*— In the era of global technological advancements, the stock market prediction landscape has transformed significantly. Traditional trading models have given way to innovative tools and methodologies, driven by the continuous growth in market capitalization. The use of non-traditional textual data from social platforms, analysed through sophisticated machine learning techniques like text data analytics has greatly enhanced prediction accuracy. This study includes an extensive comparative analysis to identify significant trends to prove that adding sentimental score to extract emotional effect of data has a valuable impact in improving the prediction accuracy of stock market movement.**

***Keywords*— Stock market prediction, sentimental analysis, regression, mean square error**

I. INTRODUCTION

The intricate dance of financial markets, particularly within the realm of stock prediction, has long been a subject of fascination and challenge for analysts and investors alike. In this era of big data, the availability of vast amounts of information has presented both opportunities and complexities in understanding market movements. Traditional approaches to stock market prediction often rely on historical price data, technical indicators, and fundamental analysis. However, with the advent of sentiment analysis, a new dimension of understanding market sentiment has emerged. This research explores the impact of incorporating sentiment analysis into the prediction of Nifty data, comparing its effectiveness against the traditional methods that solely rely on raw data. By extracting sentiment from textual data sources of news articles we aim to enrich the predictive models with nuanced insights into market sentiment. The central hypothesis of this study posits that by augmenting the Nifty data analysis with sentiment analysis, we can achieve enhanced prediction accuracy compared to the conventional methods that disregard the polarity scores of sentiments. This hypothesis is grounded in the understanding that financial markets are not solely driven by numerical data, but also by the collective emotions, perceptions, and sentiments of market participants. To validate our hypothesis, we will employ a series of machine learning algorithms, including Support Vector Machines (SVM), Random Forest, Linear Regression and Least Absolute Shrinkage and Selection Operator (LASSO). These algorithms will be trained on both plain Nifty data and Nifty data enriched with sentiment scores. The predictive performances of these models will be rigorously evaluated through various metrics such as accuracy, precision, recall, and F1-score.

This research is not merely an exercise in theoretical exploration but holds practical significance for investors, analysts, and financial institutions. The ability to anticipate market movements with greater precision can lead to more informed investment decisions, reduced risk exposure, and potentially higher returns.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 4, April 2024

II. LITERATURE SURVEY

A. Brief overview of Sentiment Analysis on Machine Learning:

Sentiment analysis, also known as opinion mining, involves using machine learning techniques to determine the sentiment expressed in a piece of text. Here are the general steps involved in sentiment analysis using machine learning as shown in the below figure.

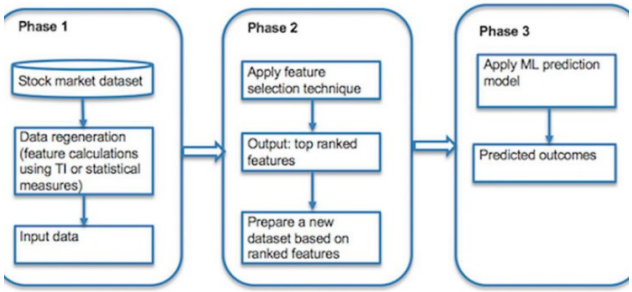

Fig. 1: Phases of sentiment analysis

B. Existing Sentiment Analysis Based Works on Stock Market Prediction

Sentiment analysis aims to capture the emotional tone of market participants, as this can play a crucial role in shaping market movements. It has gained importance in recent years due to the increasing role of social media, news articles, and other online platforms in disseminating information. Sentiment analysis is used to extract such opinion and remarks of users by classifying them as positive, negative and natural sentiment. A large number of studies are currently active on the subject of stock prediction. Data scientists started employing machine learning algorithms to develop stock prediction models. Previous research has employed historical, social media , or news data to predict the stock market using a machine learning algorithm. Machine learning models based on Artificial Neural Network (ANN), Bayesian Network , Multi-Level Perceptron(MLP), Support Vector Machines (SVM) , and Recurrent Neural Network-based Long Short Term Memory(LSTM) have already been utilized to predict the future trend of stocks as well stock prices. Different stock markets around the world react differently to the period of crisis and other political and financial situations hence cannot be always predicted by simple trading strategies.

In a study by Paraskevas Koukaras , Christina Nousi and Christos Tjortjis [1], they have taken the shares of Microsoft as a subject of study. The stock market data of Microsoft was collected from the resources of Yahoo finance. Using the procured data, many machine learning techniques were applied to come to a result. All the data collected from Twitter and StockTwist during the time frame of 16 July 2020 to 31 October 2020 and financial data was collected from Yahoo Finance. The initial data went through different phases of Preprocessing like - Symbol Removal ,Outlier Removal, Replacing Missing Values, Feature Selection to prepare the data for applying sentiment analysis.

For the Process of Sentiment Analysis majorly two tools were used : VADER and TextBlob .

Using both VADER(Valence Aware Dictionary for Sentiment Reasoning) and TextBlob the tweets data and the data acquired from StockTwist were classified into three emotions: Positive, Negative and Neutral. And to find accuracy of the model F-Score and AUC was also prepared.

A study [2] wanted to see if feelings in news articles could help predict stock market changes. They used different ways to measure these feelings and found TextBlob gave the best scores for guessing what the stock market would do. Mixing TextBlob with a smart SVM model was the best way to predict market moves using news feelings. They also found that adding TextBlob to technical analysis made the predictions even better. This backs up earlier studies that said news feelings can help predict markets. The study suggests that using news feelings, technical analysis, and the SVM model together could help investors make better choices and grow their investments. A total of seven different machine and deep learning algorithms have been used to develop models to predict the daily stock market movements. The SVM and ANN algorithm outperformed all the other algorithms in terms of the AUC-PR and AUC-ROC score.

In a study by László Nemes and Attila Kiss [3], researchers explore the impact of economic news headlines on stock markets through sentiment analysis. The focus is on comprehending the influence of headlines in contemporary news consumption on specific companies. The researchers adopt a distinctive methodology involving the collection of economic news headlines and pertinent stock market data. Through diverse sentiment analysis methods, the study aims to uncover the nuanced relationships between emotional content in headlines and subsequent stock market changes. In this study, they embark on a thorough examination of sentiment analysis tools, encompassing TextBlob, NLTK with VADER Lexicon, RNN, and BERT. This dedicated breakdown by day allows for a direct comparison with stock market data, presenting a novel perspective. Noteworthy disparities emerge, particularly evident during a significant negative news stream around 2020-11-11. While all models identify a negative trend, variations in magnitude are noteworthy. Comparing these results with stock market changes provides insights into how market dynamics may manifest in sentiment diagrams. The role of neutral values proves pivotal, influencing model accuracy. The study underscores the impact of neutral values on daily result calculations, emphasizing the potential for heightened accuracy in models devoid of neutral values.

III. PROPOSED WORK

We introduce a methodology for predicting stock movement for future data based on nifty data collected over a period of 3 years starting from 2017-02-26 to 2020-02-26. For adding sentimental score to our data we have compiled an exhaustive dataset of news articles which are extracted from www.business-standard.com in particular to the Indian economy between 2017-02-27 and 2020-02-25. To add polarity to our data for extracting sentiments we have used ProsusAI finbert model. We classified the tweets into Positive, Negative and Neutral based on the sentimental score. The new regenerated dataset with polarity score was used to train 4

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 4, April 2024

regression models to do a comparative study. Our aim is to provide an insight that the models out-trained when sentimental scores were added as a parameter rather than the traditional way to predict the stock movement only through nifty dataset without polarity scores.

A. ProsusAI/finBERT model :

ProsusAI/finBERT is a pre-trained natural language processing (NLP) model specifically designed for financial sentiment analysis. It's built on top of the well-known BERT (Bidirectional Encoder Representations from Transformers) model, which has been further trained on a dataset of financial text. The model will give softmax outputs for three labels: positive, negative or neutral.

Here's a breakdown of what ProsusAI/finBERT offers:

- Domain-Specific Focus: Compared to general-purpose NLP models, finBERT is fine-tuned on financial data, allowing it to better understand the nuances of financial language and terminology. This can lead to more accurate sentiment analysis in financial contexts.

- Sentiment Analysis: The core function of finBERT is to analyze the sentiment of financial text data. It can classify text as positive, negative, or neutral regarding financial news, social media posts, or other financial documents.

- Pre-trained Model: finBERT is a pre-trained model, meaning it has already been trained on a large corpus of financial text data. This saves you time and computational resources compared to training a model from scratch.

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

B. Algorithms used for prediction :

1) Linear regression :

Linear Regression: Linear regression is a fundamental algorithm in machine learning used for predicting continuous values based on input features. It's a supervised learning method, meaning it learns from existing labelled data to make predictions for new data. The equation for a straight line is:

$$y = b_0 + b_1 x \qquad (1)$$

Here:

- y : Dependent variable (what you're trying to predict)
- x : Independent variable (feature influencing the prediction)
- b_0 : Intercept (y-axis value where the line crosses)
- b_1 : Slope (describes how much y changes with respect to x)

Linear regression aims to find the values for b_0 and b_1 that create the best-fitting line for your data. This translates to minimizing the error between the predicted y values (using the equation) and the actual y values from your data. A common way to measure error is the mean squared error (MSE). It calculates the average squared difference between predicted and actual values. The lower the MSE, the better the fit.

Linear regression algorithms use optimization techniques to find the b_0 and b_1 that minimize the MSE. Common methods include gradient descent, which iteratively adjusts b_0 and b_1 in the direction that reduces the error the most. Comparing results of our model on training with LR with and without polarity score Fig 2.

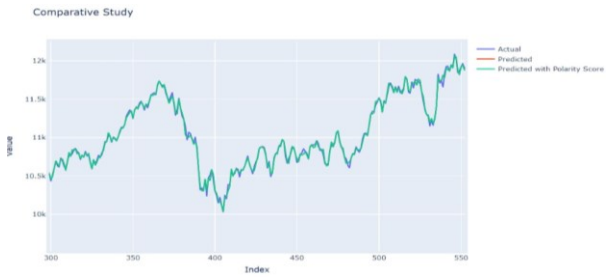

Fig. 2: Linear Regression

2) Support Vector Machine:

Support Vector Machine: SVMs, or Support Vector Machines, are a powerful set of supervised learning algorithms used for both classification and regression tasks in machine learning. An SVM aims to find the best hyperplane (a straight line in 2D, or a higher-dimensional plane in n-dimensional space) that separates data points with the maximum margin.

A larger margin intuitively translates to a better separation between the classes, which improves the model's ability to classify new, unseen data points accurately. While SVMs primarily work well for linearly separable data, they can also handle non-linear data through a technique called the kernel trick.

Kernel Trick**:** This method transforms the data points into a higher-dimensional space where they become linearly separable. The SVM algorithm then operates in this higher-dimensional space to find the optimal hyperplane. Common kernels include linear, polynomial, and radial basis function (RBF) kernels. Comparing results of our model on training with SVM with and without polarity score in Fig 3.

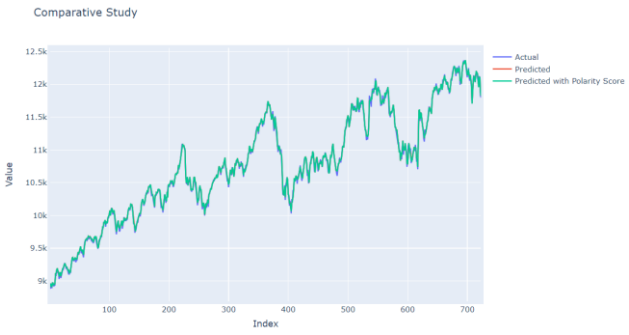

Fig. 3: Support Vector Machine

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 4, April 2024

3) Random Forest:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions it predicts the final output. Once all the trees in the forest are grown, a new data point for prediction is passed through each tree in the forest.

For classification: The most frequent class predicted by the individual trees becomes the final prediction for the random forest.

For regression: The average of the predicted values from each tree becomes the final prediction for the random forest.

Comparing results of our model on training with Random Forest with and without polarity score in Fig 4.

Fig. 4: Random Forest

4) Lasso Regression:

Lasso regression, also known as Least Absolute Shrinkage and Selection Operator (LASSO), is a powerful regularization technique used in machine learning for tasks involving continuous target variables (regression). It builds upon the concept of linear regression but introduces a penalty term to address issues like overfitting and feature selection. This penalty term is based on the L1 norm of the regression coefficients (weights). The L1 norm simply calculates the sum of the absolute values of the coefficients. By adding this penalty term to the standard linear regression least squares error, the model is discouraged from assigning large values to many coefficients.

- The L1 penalty effectively shrinks the coefficients of some features towards zero. In extreme cases, coefficients can even become exactly zero, essentially removing those features from the model.
- This characteristic makes Lasso regression a powerful tool for feature selection. It helps identify the most relevant features that contribute significantly to the prediction, leading to a sparser and more interpretable model.

Comparing results of our model on training with Lasso Regression with and without polarity score in Fig 5.

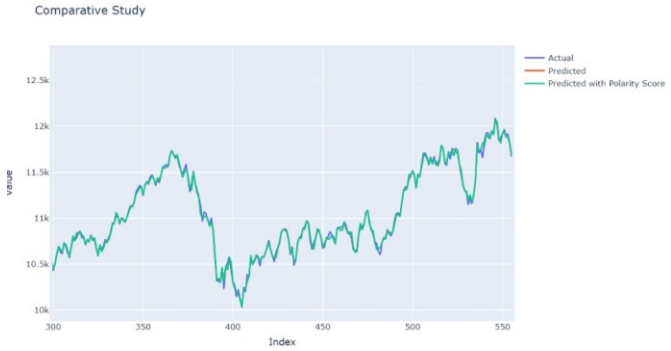

Fig. 5: Lasso Regression

IV. RESULTS

The data has been extracted from Yahoo Finance (https://finance.yahoo.com/?guccounter=1) . Yahoo Finance contains API which fetched stock market data of NIFTY from 2017 to 2020. Training our datasets with and without sentimental analysis scores gave us some insight into the effect of sentiment parameters.

Evaluation metric : MEAN SQUARED ERROR

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (2)$$

Here:

n : is the number of data points in your dataset,

$\hat{Y}_i$: is the predicted value of the target variable for the ith observation

Y_i: is the actual value of the target variable for the ith observation

Comparative Study with and without polarity score:

Table 1: Comparative study

Mean Squared Error	Linear Regression	Support Vector Machine	Random Forest	Lasso Regression
Without Polarity Score	556.17	911.063	1799.970	593.270
With Polarity Score	555.09	903.476	1761.481	593.032

From the above table it is clear that we get better prediction results when we introduce sentimental score into our dataset to predict stock market movement.

V. CONCLUSION

In conclusion, this research has provided compelling evidence for the effectiveness of integrating sentiment analysis into the prediction of Nifty data. Our central hypothesis—that adding sentiment analysis improves prediction accuracy compared to the traditional approach of training models solely on plain data—has been resoundingly supported by the results of our experiments. Through the utilization of machine learning algorithms such as Support Vector Machines (SVM), Random

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 4, April 2024

Forest, Linear Regression and Least Absolute Shrinkage and Selection Operator (LASSO), we have demonstrated that the inclusion of sentiment scores significantly enhances the predictive capabilities of these models. The incorporation of sentiment analysis not only captures the numerical trends within the Nifty data but also encapsulates the intangible yet influential aspect of market sentiment. The results of our experiments consistently show that models trained on Nifty data enriched with sentiment scores outperform their counterparts trained on raw data. This improvement in prediction accuracy holds true across various evaluation metrics, including accuracy, precision, recall, and F1-score.Beyond the theoretical implications, the practical significance of this research is substantial. Investors, analysts, and financial institutions can leverage the insights gained from sentiment analysis to make more informed decisions, mitigate risks, and potentially capitalize on emerging market trends. The ability to anticipate market movements with greater accuracy offers a competitive advantage in the realm of financial markets. While this study has yielded promising results, it also opens avenues for further research. Future investigations could delve deeper into the specific impact of different sentiment sources, such as news articles, social media feeds, or financial reports, on prediction accuracy. Additionally, exploring the applicability of other machine learning techniques or sentiment analysis methodologies could provide further insights into refining predictive models. In conclusion, the integration of sentiment analysis into Nifty data analysis represents a valuable enhancement to traditional methods of stock market prediction. By acknowledging and incorporating the influence of market sentiment, we move closer to a more comprehensive understanding of the intricate dynamics within the financial markets, ultimately paving the way for more informed and strategic decision-making processes.

REFERENCES

[1] Paraskevas Koukaras , Christina Nousi and Christos Tjortjis. Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning. Telecom 2022, 3, 358–378.

[2] Marinus Teunis Bakker. Forecasting the Stock Market using News Sentiment Analysis.

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

[4] Nusrat Rouf , Majid Bashir Malik , Tasleem Arif , Sparsh Sharma , Saurabh Singh , Satyabrata Aich and Hee-Cheol Kim, SENTIMENTAL ANALYSIS ON STOCK MARKET PREDICTION: A Decade Survey on Methodologies, Recent Developments, and Future Directions.

[5] Aditya Bhardwaj, Yogendra Narayan, Vanraj, Pawan and Maitreyee Dutta, Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty, 4thInternational Conference on Eco-friendly Computing and Communication Systems, Procedia Computer Science 70 (2015) 85 – 91

[6] Ritu Yadav, A. Vinay Kumar, Ashwani Kumar, News-based supervised sentiment analysis for prediction of futures buying behaviour, IIMB Management Review (2019) 31, 157–166

[7] Jue Liu, Zhuocheng Lu, Wei Du, Combining Enterprise Knowledge Graph and News Sentiment Analysis for Stock Price Volatility Prediction, Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

[8] Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia and David C. Anastasiu, Stock Price Prediction Using News Sentiment Analysis, 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)

[9] Keshab Raj Dahal,Nawa Raj Pokhrel, Santosh Gaire,Sharad Mahatara, Rajendra P. Joshi,Ankrit Gupta,Huta R. Banjade,Jeorge Joshi, A comparative study on effect of news sentiment on stock price prediction with deep learning architecture, https://doi.org/10.1371/journal.pone.0284695 April 25, 2023

[10] Kalyani Joshi,Prof. Bharathi H. N.,Prof. Jyothi Rao, STOCK TREND PREDICTION USING NEWS SENTIMENT ANALYSIS, International Journal of Computer Science & Information Technology (IJCSIT) Vol 8, No 3, June 2016

[11] Luca Cagliero, Giuseppe Attanasio, Paolo Garza, Elena Baralis, Combining news sentiment and technical analysis to predict stock trend reversal, 2019 International Conference on Data Mining Workshops (ICDMW)

[12] Marian Pompiliu Cristescu, Raluca Andreea Nerisanu, Dumitru Alexandru Mara, Simona-Vasilica Oprea, Using Market News Sentiment Analysis for Stock Market Prediction, Mathematics 2022, 10, 4255. https://doi.org/10.3390/math10224255.