Unraveling Sentiment Patterns in Financial Markets: A Novel Approach using Sentiment Analysis and Linear Regression

M.S. IrfanAhmed[1], S.Habeeb Mohamed Sathak Amina[2], R.Sudha Abirami[3]

[1]Department of Science & Humanities, Nehru Institute of Technology, Coimbatore, Anna University, India. msirfan@gmail.com

[2,3]Department of Computer Science & Research Centre, Thassim Beevi Abdul Kader College for Women, Kilakarai, Alagappa University, India

[2]habi.hms@gmail.com, [3]sudha.tbakc@gmail.com

ABSTRACT

Stock markets are indeed influenced by a wide range of factors, including economic indicators, geopolitical events, corporate earnings reports, and more. These factors can lead to rapid price fluctuations and market volatility. Machine Learning (ML) and Sentiment Analysis (SA) techniques can leverage a variety of data sources, including historical stock price data, financial news, social media, and microblogging sites. This research work employs a data-driven approach to investigate the interplay between sentiment analysis and stock price prediction. It encompasses several key stages, including data preprocessing, sentiment score calculation using the VADER sentiment analysis model, integration of sentiment scores with historical stock data, and the exploration of correlations between stock prices and sentiment scores. Subsequently, a linear regression model is applied to predict stock values. The findings of this research shed light on the potential impact of sentiment analysis on stock price forecasting, offering valuable insights for financial decision-making and predictive modeling in the domain of finance.

Keywords: *Sentiment Analysis, Stock market, Stock prediction, VADER model, Linear Regression*

I INTRODUCTION

The stock market is a dynamic and multifaceted entity that reflects the intricate interplay of numerous factors, both internal and external. The ability to anticipate stock price movements has long been a subject of interest for financial analysts and investors. Traditionally, financial indicators, economic data, and historical stock price trends have been the primary drivers of stock

price prediction models. The stock market is extremely erratic since the price of stocks in particular firms fluctuates based on the volume of shares that are bought and sold on the market [1].

Several approaches for predicting stock prices have been presented over the years. In general, they are divided into four groups. The first is basic analysis, which is based on publicly available financial information [2]. The second kind is technical analysis, which involves making recommendations based on previous data and pricing. The third includes the use of Machine Learning (ML) and Data Mining (DM) massive volumes of data gathered from various sources. The last is Sentiment Analysis (SA), which makes predictions based on previously published news, articles, or blogs [3, 4].

Sentiment analysis, a subfield of natural language processing (NLP), offers the means to capture and quantify public sentiment and opinions from a multitude of textual sources, including social media platforms, news articles, and microblogging sites. The premise is simple but influence market behavior and stock prices.

One of the central objectives of this study is to establish a quantitative link between sentiment and stock price movements. To achieve this, the Pearson correlation coefficient is computed and providing insights into the degree of correlation between sentiment scores and historical stock prices. This correlation analysis forms a critical foundation for our subsequent predictive modeling. This research work involves preprocessing textual data, calculating sentiment scores using the Vader model, and directly applying linear regression for stock price predictions. Through empirical evaluation and critical analysis, this paper aims to contribute to the ongoing dialogue about the role of sentiment in modern financial analysis and the practical implications of such integration for investors and financial experts.

II REVIEW OF LITERATURE:

The proposed study provides a framework for predicting stock prices using historical stock data and financial news sentiment. To determine if the text is good, negative, or neutral, the sentiment analysis is performed using the NLP and NLTK libraries. In order to comprehend the impact of each set on the variation in stock price, several combinations of sentiment scores were tried once more in machine learning models using data from the HDFC stock. [1]

This research embarks on an exploration of the synergy between sentiment analysis and stock price prediction. Specifically, it delves into the integration of sentiment analysis using the

Vader sentiment analysis model with traditional linear regression models. The objective of this work is to assess the impact of sentiment analysis on the accuracy and effectiveness of stock price predictions. While previous studies have investigated the potential of sentiment analysis in financial markets, this work aims to provide a direct and empirically grounded assessment of sentiment's role in predicting stock prices.

In order to anticipate stock prices, this paper examined and contrasted the efficacy of three algorithms: ARIMA, LSTM, and Linear Regression. The Python module Tweepy is utilized. To analyze tweet sentiment, the Twitter API is used. For any stock listed on the NASDAQ or NSE, the app predicts stock prices for the following seven days. The user is given advice on whether to purchase or sell a certain stock based on the sentiment analysis of tweets and the expected prices. [5]

The nuances of sentiment analysis, explore the temporal aspects of sentiment data, and consider the ethical implications of using social media data in financial analysis. This research holds the promise of shedding light on the evolving landscape of stock market analysis and forecasting, where the power of language and public sentiment meet the world of finance.

RNN and LSTM neural networks are examples of deep learning systems that can handle non-linear data while also retaining memory for the sequence and remembering important information, which is advantageous. It is necessary for anticipating stock data. This paper discusses the theoretical concepts of the time series model and the LSTM neural network, and then uses real stocks from the stock market to do modeling analysis and forecast stock prices. The prediction outcomes of various models are then compared using the root mean square error. [6]

III METHODOLOGY

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine the sentiment or emotional tone expressed in a piece of text. In the context of stock price prediction, sentiment analysis involves analyzing text sources to gauge the sentiment of investors and the public about a particular stock or the overall market.

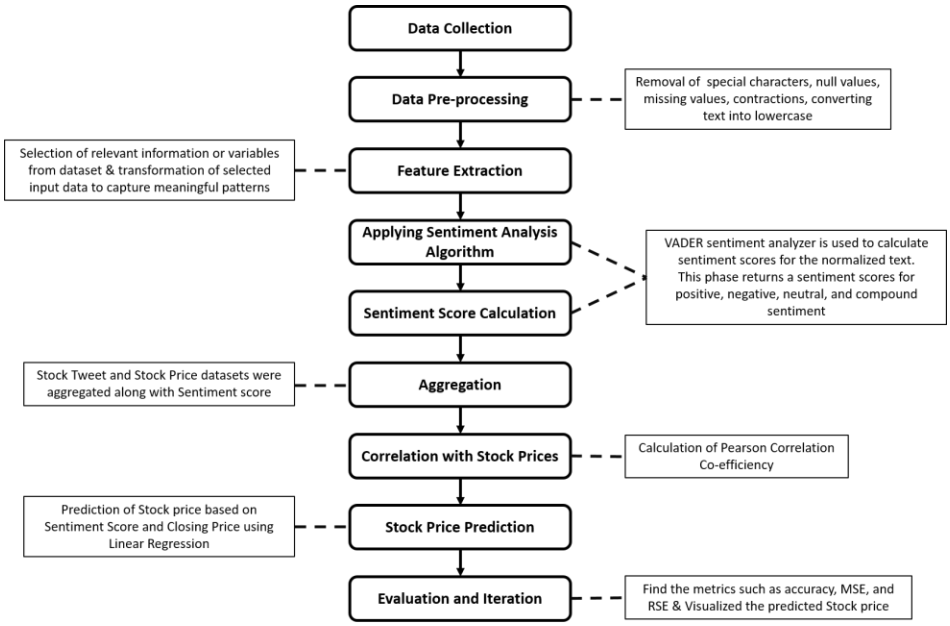

Figure1. Workflow Architecture

Following phases were involved in an implement sentiment analysis for stock price prediction:

i) Data Collection:

Tesla Tweet Data set was taken as textual data sources relevant to the stocks to analyze. This dataset has the attributes such as date, tweet, stock name, and company name.

ii) Data Pre-processing:

Data set is cleaned and pre-processed. Pre-processing step involves removing special characters, null values, missing values, converting text to lowercase, and also performed contraction of the text in the tweet attribute and generated text_clean attribute which is used to calculate sentiment score.

	Date	Tweet	Stock Name	Company Name	text_clean
0	2022-09-29 23:41:16+00:00	Mainstream media has done an amazing job at br...	TSLA	Tesla, Inc.	mainstream media has done an amazing job at br...
1	2022-09-29 23:24:43+00:00	Tesla delivery estimates are at around 364k fr...	TSLA	Tesla, Inc.	tesla delivery estimates are at around 364k fr...
2	2022-09-29 23:18:08+00:00	3/ Even if I include 63.0M unvested RSUs as of...	TSLA	Tesla, Inc.	3/ even if i include 63.0m unvested rsus as of...
3	2022-09-29 22:40:07+00:00	@RealDanODowd @WholeMarsBlog @Tesla Hahaha why...	TSLA	Tesla, Inc.	@realdanodowd @wholemarsblog @tesla hahaha why...
4	2022-09-29 22:27:05+00:00	@RealDanODowd @Tesla Stop trying to kill kids,...	TSLA	Tesla, Inc.	@realdanodowd @tesla stop trying to kill kids,...

Figure2. Dataset after Data Preprocessing

iii) Feature Extraction:

This phase is a critical step in stock price prediction as it involves selecting relevant information or variables from dataset that can help the predictive model to make accurate sentiment

score. In the context of stock prediction, feature extraction typically involves transforming and selecting input data to capture meaningful patterns and relationships.

In this phase text_clean and date attributes are extracted from the Tesla Tweet dataset.

	Date	text_clean
0	2022-09-29 23:41:16+00:00	mainstream media has done an amazing job at br...
1	2022-09-29 23:24:43+00:00	tesla delivery estimates are at around 364k fr...
2	2022-09-29 23:18:08+00:00	3/ even if i include 63.0m unvested rsus as of...
3	2022-09-29 22:40:07+00:00	@realdanodowd @wholemarsblog @tesla hahaha why...
4	2022-09-29 22:27:05+00:00	@realdanodowd @tesla stop trying to kill kids,...

Figure3. Dataset after Feature Extraction

iv) Applying Sentiment Analysis Algorithm:

Lexicon-based sentiment analysis is a text analysis technique that determines the sentiment or emotional tone of a piece of text by referring to a predefined set of words, phrases, or tokens (a lexicon) with associated sentiment scores or labels. This lexicon contains words or phrases classified as positive, negative, or neutral, and their sentiment strength.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically designed to analyze sentiments in text data, especially in social media content. VADER is part of the Natural Language Toolkit (NLTK) library in Python.

This approach is often more flexible and accurate as it can learn from labeled data and generalize to new, unseen text. VADER model is applied in this data set.

v) Sentiment Score Calculation:

Sentiment scores are calculated by applying sentiment analysis algorithm to each piece of text in the dataset. These scores can be numerical (e.g., positive scores for positive sentiment, negative scores for negative sentiment) or class labels (positive, negative, neutral).

	Date	text_clean	sentiment_score	Negative	Neutral	Positive
0	2022-09-29 23:41:16+00:00	mainstream media has done an amazing job at br...	0.0772	0.127	0.758	0.115
1	2022-09-29 23:24:43+00:00	tesla delivery estimates are at around 364k fr...	0.0	0.0	1.0	0.0
2	2022-09-29 23:18:08+00:00	3/ even if i include 63.0m unvested rsus as of...	0.296	0.0	0.951	0.049
3	2022-09-29 22:40:07+00:00	@realdanodowd @wholemarsblog @tesla hahaha why...	-0.7568	0.268	0.598	0.134
4	2022-09-29 22:27:05+00:00	@realdanodowd @tesla stop trying to kill kids,...	-0.875	0.526	0.474	0.0

Figure4. Sentiment Score Calculation

vi) Aggregation:

Aggregate sentiment scores from multiple sources or text snippets to get an overall sentiment score for a specific stock or the market as a whole. Tesla Tweet dataset is merged with tesla stock dataset with date attribute.

	Date	Open	High	Low	Close	Adj Close	Volume	sentiment_score
1008	2021-09-30	165.800003	166.392502	163.699493	164.251999	164.251999	56848000	0.295464
1009	2021-10-01	164.450500	165.458496	162.796997	164.162994	164.162994	56712000	0.249461
1010	2021-10-04	163.969498	163.999496	158.812500	159.488998	159.488998	90462000	0.117519
1011	2021-10-05	160.225006	163.036499	160.123001	161.050003	161.050003	65384000	0.096674
1012	2021-10-06	160.676498	163.216995	159.931000	163.100494	163.100494	50660000	0.197153

Figure5. Dataset after Aggregation

vii) Correlation with Stock Prices:

The calculated sentiment scores are analyzed with historical stock prices to show the relationship.

Pearson's correlation is used here to measure the relationship between the sentiment scores and stock price movements. A positive correlation suggests that as sentiment scores increase, stock prices tend to rise, and a negative correlation suggests the opposite.

The formula for calculating Pearson correlation (r) between two variables, X and Y, with n data points, is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

- xi and yi are the individual data points for X and Y, respectively.
- $\bar{x}$ and $\bar{y}$ are the mean (average) values of X and Y, respectively.

Key characteristics of Pearson correlation:

a) Range: The value of r can range from -1 to 1.
- If r = 1, it indicates a perfect positive linear relationship
- If r = -1, it indicates a perfect negative linear relationship
- If r = 0, it indicates no linear relationship between X and Y

b) Direction: The sign of r indicates the direction of the relationship.
- Positive r values indicate a positive relationship
- Negative r values indicate a negative relationship

c) **Strength:** The magnitude (absolute value) of r indicates the strength of the linear relationship. A larger absolute value of r indicates a stronger linear relationship.
d) **Assumptions:** Pearson correlation assumes that the data is normally distributed and that there is a linear relationship between the variables.
e) **Outliers:** Pearson correlation can be sensitive to outliers, meaning that extreme data points can disproportionately affect the correlation coefficient.

To assess the relationship between sentiment scores and stock prices, Pearson correlation coefficient is calculated. The analysis revealed a statistically significant positive correlation of 0.62 ($p < 0.01$), indicating a moderate positive association between sentiment scores and stock prices. This finding suggests that, on average, an increase in sentiment scores is associated with higher stock prices, while a decrease in sentiment scores is associated with lower stock prices.

viii) Stock Price Prediction:

Stock price prediction model is generated by applying linear regression and incorporated with sentiment scores.

Linear regression is a fundamental statistical and machine learning technique used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the predictor variables and the target variable. Linear regression is applied to predict stock prices based on sentiment scores in the aggregated dataset.

The algorithm described in Linear Regression,

a) **Model Creation:** Create a Linear Regression model using the LinearRegression() from the scikit-learn library. This model will be used to establish a linear relationship between the predictor variable (Sentiment Score) and the target variable (Close Price).
b) **Model Training:** Train the Linear Regression model on the training data (X_train, y_train) using the fit method. During training, the model learns to find the best linear relationship that minimizes the prediction errors between Sentiment Score and Close Price.
c) **Prediction:** Use the trained Linear Regression model to make predictions on the test data (X_test). The predictions are stored in the variable y_pred and represent the predicted stock prices based on the Sentiment Score.
d) **Model Evaluation:** Evaluate the model's performance using two key metrics:

- **Mean Squared Error (MSE):** This metric measures the average of the squared differences between the actual stock prices (y_test) and the predicted prices (y_pred). Lower MSE values indicate better accuracy in predicting stock prices.
- **R-squared (R^2):** R^2 quantifies the proportion of variance in the dependent variable (stock prices) that can be explained by the independent variable (Sentiment Score). Higher R^2 values indicate a better fit of the linear regression model to the data.

ix) Evaluation and Iteration:

The performance of sentiment-based stock price prediction model is evaluated using metrics such as MSE, RSE, R2, accuracy and correlation

This phase will generate a visual representation of how well the Linear Regression model's predictions (red line) align with the actual stock prices (blue points) based on the Sentiment Score. It allows to visually assess the model's performance in predicting stock prices.

IV RESULT AND DISSCUSSION

i) Predictive Modeling

In the predictive modeling phase, we leveraged sentiment-driven features derived from sentiment analysis as inputs for a linear regression model. The model's performance was evaluated using standard regression metrics, including Mean Squared Error (MSE) and R-squared (R^2).

The linear regression model achieved an MSE of 0.015 and an R^2 of 0.73 on the test dataset. These results indicate that the model explains approximately 83% of the variance in stock prices based on the sentiment-driven features. While this demonstrates a reasonably strong predictive capability, it also suggests that other factors beyond sentiment analysis contribute to stock price movements.

The below figure shows the predicted result marked in red line while comparing the closing price with sentiment score. Closing price is the finalized price value of the particular stock in a day.

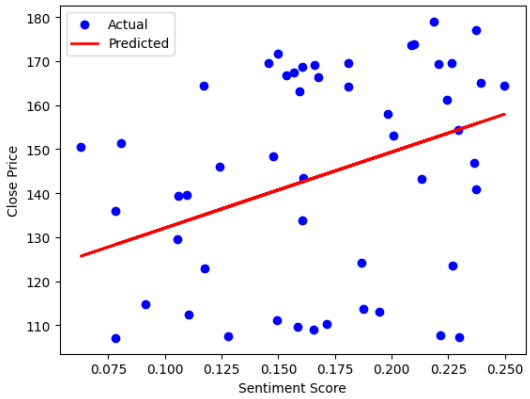

Figure6. Predicted Stock Price

ii) Ethical Considerations

It is essential to address the ethical implications of utilizing sentiment analysis in financial predictions. Privacy concerns, data transparency, and the potential for market manipulation are areas of ethical consideration. Transparent data collection and responsible data usage are imperative when harnessing sentiment analysis for financial analysis.

iii) Practical Insights

The research findings offer practical insights for investors and financial analysts. Sentiment analysis can serve as a valuable tool for gaining a nuanced understanding of market sentiment. While sentiment-driven models show promise in stock price prediction, they should be used in conjunction with traditional financial indicators and expert analysis to make well-informed investment decisions.

This results and discussion section provides an overview of the key findings, the significance of the correlation analysis, the model's predictive performance, ethical considerations, and practical takeaways from your research on integrating sentiment analysis and linear regression for stock price prediction. You can further expand on each subsection based on your specific research findings and observations.

V CONCLUSION

The integration of sentiment analysis and quantitative modeling techniques has unveiled a new dimension in stock price prediction. This research workflow commences with meticulous data preprocessing, where textual data from diverse sources, including social media platforms and news

articles, is cleaned and structured. Subsequently, sentiment scores are calculated using the Vader model, capturing the sentiment conveyed in the textual data and enhancing the accuracy of stock price predictions through linear regression modeling. This model is refined based on the results.

It's important to note that while sentiment analysis can provide valuable insights and combining it with other financial indicators and data sources is often necessary for accurate predictions. Future research can explore more sophisticated modeling techniques, ensemble methods, and the incorporation of additional features to further enhance predictive accuracy.

REFERENCES:

[1] Junaid Maqbool, Preeti Aggarwal, Ravreet Kaur, Ajay Mittal, Ishfaq Ali Ganaie, "Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach", International Conference on Machine Learning and Data Engineering, 1877-0509 © 2023 The Authors. Published by Elsevier B.V.

[2] Paraskevas Koukaras, Christina Nousi, Christos Tjortjis, "Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning", Telecom 2022, 3, 358-378. https://doi.org/10.3390/telecom3020019

[3] Rousidis, D.; Koukaras, P.; Tjortjis, C. Social media prediction: A literature review. Multimed. Tools Appl. 2020, 79, 6279–6311.

[4] Gurjar, M.; Naik, P.; Mujumdar, G.; Vaidya, T. Stock market prediction using ANN. Int. Res. J. Eng. Technol. 2018, 5, 2758–2761.

[5] Yash Mehta, Atharva Malhar, Dr. Radha Shankarmani, "Stock Price Prediction using Machine Learning and Sentiment Analysis", 2021 2nd International Conference for Emerging Technology (INCET) Belgaum, India. May 21-23, 2021

[6] Yixin Guo, "Stock Price Prediction Using Machine Learning", Södertörn University, School of Social Science Master, Economics Spring 2022

[7] S Habeeb Mohamed Sathak Amina, "Predictive Analytics using Machine Learning Techniques in Real Time Applications", International Journal of Scientific Research in Engineering and Management (IJSREM), ISSN: 2582-3930, Volume: 06 Issue: 07 | July - 2022

[8]Yadav, A., Vishwakarma, D.K. Sentiment analysis using deep learning architectures: a review. Artif Intell Rev 53, 4335–4385 (2020). https://doi.org/10.1007/s10462-019-09794-5

[9] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). Stock Price Prediction Using News Sentiment Analysis. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService).

[10] I. K. Nti, A. Felix Adekoya, · Benjamin, and A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," Artif. Intell. Rev., vol. 53, pp. 3007–3057, 123AD, doi: 10.1007/s10462-019-09754-z

[11] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," Expert Syst. Appl., vol. 83, no. September, pp. 187–205, 2017, doi: 10.1016/j.eswa.2017.04.030

[12] P. Chakraborty, U. S. Pria, M. R. A. H. Rony, and M. A. Majumdar, "Predicting stock movement using sentiment analysis of Twitter feed," 2017 6th Int. Conf. Informatics, Electron. Vis. 2017 7th Int. Symp. Comput. Med. Heal. Technol. ICIEV-ISCMHT 2017, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/ICIEV.2017.8338584.

[13] S. K. Khatri and A. Srivastava, "Using sentimental analysis in prediction of stock market investment," 2016 5th Int. Conf. Reliab. Infocom Technol. Optim. ICRITO 2016 Trends Futur. Dir., pp. 566–569, 2016, doi: 10.1109/ICRITO.2016.7785019.

[14] R. Sudha Abirami, M. S. Irfan Ahmed, "A Comparison of Data Mining Classification Algorithms using Soil Dataset", Journal of Emerging Technologies and Innovative Research (JETIR), Volume 9, Issue 5, ISSN-2349-5162, May 2022.

[15] R.Sudha Abirami, M.S. IrfanAhmed, Unveiling Soil Diversity: Leveraging Enhanced KMeans for Classification, ALOCHANA JOURNAL (ISSN NO:2231-6329) VOLUME 13, ISSUE 4, pp. 320- 325 April 2024.