# Enhancing Stock Market Prediction Accuracy Using Machine Learning and Sentiment Analysis: A Hybrid Approach

[1]Abhishek Kumar Singh, [2]Dr. Gagan Sharma
*[1]M.Tech Scholar, [2]Associate Professor,*
*[1]Department of Computer Science and Engineering, Sri Satya Sai College of Engineering, Bhopal (M.P)*
*[2]Department of Computer Science and Engineering, Sri Satya Sai College of Engineering, Bhopal (M.P)*
Email-: [1]gpjnpcs3@gmail.com, [2]gagansharma.cs@gmail.com

* Corresponding Author: Abhishek Kumar Singh

**Abstract:** *The stock market plays a significant role in the global economy, with investors facing challenges due to volatility and uncertainty in stock prices. Traditional stock analysis methods, such as fundamental and technical analysis, have been enhanced by modern computational techniques, including machine learning and sentiment analysis. This paper explores the integration of different machine learning methods, including linear regression, with external data sources, historical stock data, and news sentiment analysis to improve stock prediction accuracy. The study evaluates the performance of models by comparing traditional methods to those incorporating sentiment analysis, concluding that sentiment-based models significantly enhance forecasting accuracy. The paper highlights the importance of data preprocessing and the combination of financial metrics with market sentiment to provide more precise investment recommendations, particularly in the Indian stock market.*

**Keywords:** *Stock Market Prediction, Machine Learning, Sentiment Analysis, Financial Analysis, Investment Decisions, Data Preprocessing, Deep Learning, Technical Analysis, Social Media Data, Forecasting Models.*

## I. INTRODUCTION

Stock market is relatively a major aggressive economic business sector where the dealers are required and process the economic workloads with lower latency alongside higher throughput. Formerly, economists were utilizing the customary store and cycle technique to figure out the weighty economic workload productively. However, to accomplish low idleness and high throughput, server farms had to be genuinely found near the information sources, rather than other all the more financially gainful areas. The primary explanation, the information is that streaming model has been created and it can handle enormous measures of information more proficiently. It has appeared in examinations that utilizing information streaming, we can tackle the alternative estimating and hazard evaluation issues by utilizing customary techniques, for instance, Japanese candles, Monte-Carlo models, Binomial models, with low inactivity and high throughput [1]–[3].

The stock market in both business and academia, investments are becoming more important. Because the market is so unstable, it is hard to tell what stock prices will do. So, people who want to trade in the stock market should be more careful when they do so. People who want to invest in the stock market use old-fashioned stock info, which isn't always accurate. Different kinds of information have too much power in the Indian stock market because more and more people use the Internet and use sites like Google Trends, Yahoo Finance, Wikipedia, and traders' websites. So, financial information and how people feel about the news are thought to affect the stock price. Analysts, researchers, managers, and buyers all care about how much a stock is worth [4]–[6]. They're figuring out how much the companies are worth so that investors can buy stocks that are cheap or overvalued. The foremost objective is to develop a model for investors that can be used to choose the best stocks from the Indian Stock Market. Historical data from the stock market, external data sources and news are used as inputs to the model. The study uses different approaches combined with machine learning techniques such as linear regression to reduce the valuation errors in the fundamental analysis. The significance of sentimental analysis and the power of sentiments in the movement of stock price is also studied. Finally, the model's performance is evaluated, and from the analysis, it is found that the use of sentimental analysis and fundamentals leads to improved performance, which will help the investors make decisions accurately [7]–[11].

## II. DIFFERENT TECHNIQUES USED IN THE STOCK MARKET FOR INVESTMENT DECISIONS

Fernandez (2002) reviewed how brokers, traders, and portfolio managers helped clients make smart investment choices in a variety of ways. The methods most often used were fundamental and technical analysis. As new technologies came out, they brought with them powerful computers and new ways to do both fundamental and technical research. After that, when blogs, social networks like Twitter, financial sites like Yahoo Finance, and Google Trends were created, people began to share their opinions on stocks. This had a big impact on the current stock market trend [12]–[15].

### A. Fundamental Analysis

When someone wants to put money into a business for the long term, they need to know about all of its different points of view. This is called fundamental research. The purpose of fundamental analysis is to find out what a product or business is really worth by looking at its financial statements, competitors, and markets. It involves looking at both the qualitative and quantitative aspects of a business. It helps you figure out things like sales, profits, assets, debts, share capital, and how the business is run. One way to find stocks that are undervalued or overvalued in the stock market is to use valuation models. It's important for investors to understand and measure an asset's "intrinsic value" and know where that value comes from. Valuation models use a variety of factors to figure out how much a stock is worth, but they all share some traits [16].

Mehra (2010) put valuation models and methods into three groups. There are three models: risk-based, discounted, and relative. The value of an asset or share is based on how much money it is expected to bring in in the future. This is called discounted assessment. Comparing similar assets in the same business and figuring out how much an asset or share is worth by looking at the basics of the company is what relative valuation does. The risk factors are used by risk-based models to figure out how much an object is worth. Reviewers say that P/B, P/E, and CAPM are the most correct ways to value stocks on the market[17].

### B. Technical Analysis

Drakopoulou et.al (2015) put forward that the stock price movements are inclined by investor mindset and the financial variables. Price movements, short sales and trading volume, indicates investor psychology and future price movements. Technical Analysis studies patterns of charts and statistical figures to understand market trends and choose the best stocks accordingly[18].

### C. Sentimental Analysis

Gupta et.al (2018) studied the use of sentiment analysis is a method used to get and evaluate the opinions of users by classify it as positive, negative and neutral sentiment. Sentiment analysis is one of the techniques widely used to extract the opinions of persons from data available on the internet [19].

### D. Machine learning approach

This way of doing mood analysis is broken down into two types: supervised learning and unsupervised learning. In this method, it requires two sets of documents: a teaching set and a test set. The training set is used to learn how to read different types of papers, and the test sets are used to see how well the classifier is doing [20].

### E. Lexicon based approach

Overall sentiment of a text can be calculated as the sum of sentiments of each word based on their position in the text. There are two types which are dictionary approach and corpus-based approach. Predefined dictionary of words is used in dictionary approach. In this approach, a precise sentiment polarity is attached to each word [21].

Kaur (2014) Also found that a lot of people think investing in the stock market is risky, mostly because the prices of shares are affected by a lot of different factors. The stock market is risky, so buyers can lose a lot of money sometimes. A better way to understand the risks is to study the various influencers and its tremendous effect on stock movement or price. For people who want to invest in the stock market, market knowledge, tips, news, and market sentiment are very important. There are varieties of techniques that provided recommendations to stockholders on buy and sell choices of stocks. Tips provided by brokers on buy or sell actions of software are not very accurate and reliable. So, it proves the valuation inaccuracy of model in the market. Therefore, identifying the best valuation techniques or models that provide good suggestions or results, and refining these models done are necessary. This will support the accurate prediction of the value of stock and also to give buy or sell suggestions to investors in the market.

.

## III. MARKET PREDICTION BY USING MACHINE AND DEEP LEARNING METHODS

Many examinations contemplate utilized machine learning by means of text mining development strategies to effectively anticipate the financial exchange changes. Also, the different machine learning methods which may support to predict the financial market as follow.

Borne et al. (2005) it is referred as a managed AI model utilized broadly in order and reversion undertaking has been done. It is a hyper plane that isolates an assortment of reports into at least two classes with a most extreme edge. Initially, it was applied to the content grouping task and this methodology; the scholar utilized a limited terminology as an element assortment by utilizing a rundown of the most happened words and dispose of exceptional words from the list of capabilities.

By using 12,901 reports, from the Reuters 21577 archive bunch, and 20,000 clinical databases where the scholar looked at the viability of many AI procedures like; SVM and Naive Bayes (NB). For both these events, the preliminaries showed that the SVM achieve better portrayal result appeared differently in relation to NB classifier. For stock trade assumption

like various assessment articles used the SVM as far as text plan and leaning examination though to join both literary information and recorded stock costs which is utilized for stock trade assumption research that has been applied in SVM while to hypothesize the Chinese stock and stock expenses between the years 2008 - 2015. For text mining, the researchers outlined a stop word and assessment word reference reliant upon a specific space.

In the examination there were two kinds of data utilized such as; the first consolidates 2,302,692 news things, though the other contains just stock information of the biggest 20 Chinese stocks dependent on exchanging volume. Support vector relapse (SVR) is utilized to anticipate stock cost, and support vector grouping (SVG) is abused to foresee stock heading.

.

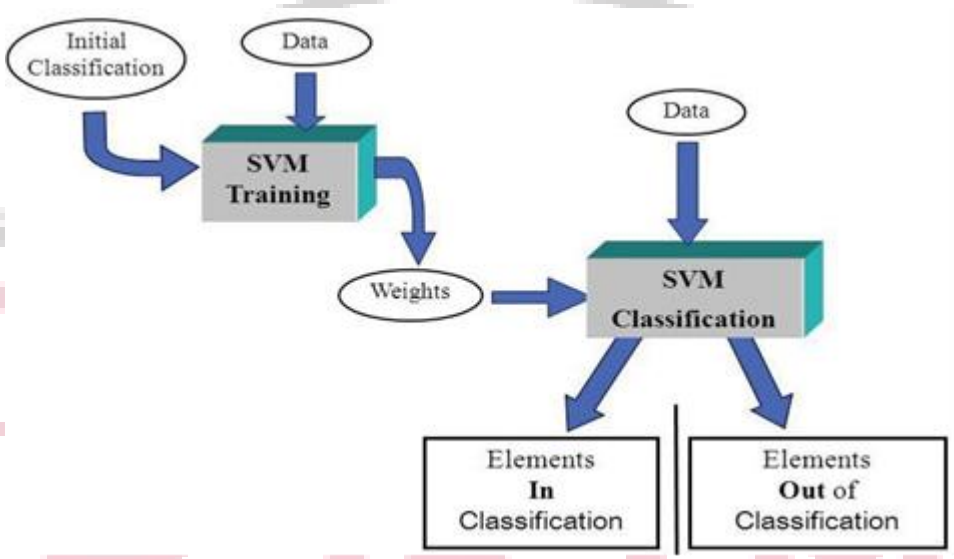


Figure 1 Overview of support vector machine for data processing

The outcome shows that both crowd numbers and news quality significantly affect the financial exchange. Besides, for SVC, the heading precision was about 59.1%, which outlines preferable advancement over different works. The outcome additionally demonstrates that news stories importantly affect the stock exchange deviations. Uysal et al. (2017) In AI, deep learning means taking in a lot of information and organizing it in layers so that highlights, cases, and descriptions can be found. Tasks that involve checking assumptions are given to a deep learning method, and the results are generally seen as good [22].

The researchers looked into whether deep learning methods can be changed to make Stock Twits' end result more accurate. Finding financial exchange reviews on Stock Twits involved looking at a few different types of neural networks, such as LSTM, doc2vec, and CNN. The results show that the convolution neural network is the best strategy for predicting how students will feel in the Stock Twits dataset compared to other deep learning methods. Many types of research have looked at how well deep learning can be used for tasks like learning common languages and understanding feelings. On the review research in, a portion of the various techniques utilized in assessment investigation assignments are looked at. The principle result showed the better presentation of deep learning techniques for reaction examination, specifically, CNN and LSTM strategies.

**IV. DATA SOURCES AND PREPROCESSING**

The data used in stock market prediction typically comes from multiple sources that provide insights into both the market's performance and the external factors that could influence stock prices. The primary data sources include:

## A. **Stock Prices and Historical Data**

Stock prices are fundamental to any prediction model. These include open, high, low,close prices (OHLC), trading volume, and market capitalizatHistorical stock prices are often used to identify trends and patterns in stock movements, which can then be modeled using various techniques such as time series analysis and machine learning. Data is typically obtained from financial websites, APIs, and databases like Yahoo Finance, Alpha Vantage, or Quandl.

## B. Financial News

News articles, press releases, and financial reports are vital in assessing market sentiment. Events such as corporate earnings reports, mergers and acquisitions, product launches, or geopolitical events can significantly affect stock prices. Natural language processing (NLP) techniques, such as sentiment analysis, are applied to analyze news data and extract relevant sentiment or opinions, which can then be used as features for prediction models. Data can be collected from financial news websites (e.g., Reuters, Bloomberg) or social media platforms.

## C. Social Media and Alternative Data

Social media platforms like Twitter, Reddit (specifically subreddits like r/WallStreetBets), and StockTwits have become influential in driving stock market trends. Posts, tweets, and discussions often contain public sentiment, rumors, and discussions that can impact stock prices, sometimes even more directly than traditional news. Sentiment analysis and text mining are applied to these platforms to extract useful features for stock price prediction. Additional alternative data sources might include satellite images, weather patterns, or credit card transaction data, which offer non-traditional insights into company performance.

**Preprocessing** is crucial for making raw data usable for analysis. For stock prices, this may involve cleaning missing values, handling outliers, and normalizing data (e.g., scaling prices to avoid large price values dominating the model). For textual data, preprocessing might include tokenization, removing stop words, and stemming/lemmatization to reduce words to their base forms. Furthermore, data may need to be transformed into structured formats suitable for input into machine learning models, such as converting time-series data into features or aggregating sentiment scores into numerical values.

## V. EVALUATION OF PREDICTION MODELS

When evaluating prediction models for stock market forecasting, several critical aspects need to be considered to assess their effectiveness in predicting trends and making reliable decisions.

Accuracy is a fundamental metric that gauges how well a model predicts the correct outcomes. In the context of stock market prediction, this often means forecasting whether a stock price will increase or decrease over a certain period. However, accuracy alone may not suffice, especially in financial data, due to its imbalanced nature. For instance, in many datasets, the number of periods with price increases may far outweigh the periods with price decreases. In such cases, metrics like precision and recall, commonly used in classification tasks, provide a more nuanced understanding of the model's ability to correctly identify price movements. Precision assesses the proportion of true positive predictions among all predicted positives, while recall evaluates the ability of the model to identify all true positives.

In addition to accuracy, performance metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared ($R^2$) are pivotal for evaluating prediction quality. RMSE measures the average magnitude of error by calculating the square root of the average squared differences between predicted and actual values. A lower RMSE indicates that the model's predictions are closer to the real stock prices. MAE provides another measure of prediction error by calculating the average absolute difference between the predicted and actual values, offering a more intuitive sense of the model's prediction performance. $R^2$, on the other hand, assesses how well the model explains the variation in stock prices. A higher $R^2$ value suggests that the model captures the underlying patterns of the data more effectively[23].

Furthermore, the reliability and generalization of a model are crucial for stock market prediction, as market conditions are constantly evolving. Reliability refers to how consistently a model can deliver accurate results over time. Generalization, the ability of a model to perform well on unseen data, is tested through techniques like cross-validation and out-of-sample testing. Cross-validation helps in ensuring that the model does not overfit to the training data, which can result in great performance on training sets but poor performance on new, unseen data. To combat overfitting, methods like regularization (e.g., Lasso or Ridge regression) can be applied, which penalize overly complex models, forcing them to generalize better to new data.

Lastly, computational efficiency is another crucial aspect, especially for real-time stock prediction. Given the large volume of financial data and the need for quick decision-making in stock markets, models must be able to process and output predictions rapidly. Techniques like deep learning can be highly accurate, but they require substantial computational resources, such as powerful GPUs or cloud infrastructure, for training and inference. Thus, there is a trade-off between predictive accuracy and computational efficiency. Models must strike a balance to ensure that they provide real-time, reliable predictions without overwhelming system resources[24].

## VI. CHALLENGES AND LIMITATIONS IN STOCK MARKET PREDICTION

When evaluating prediction models for stock market forecasting, several critical aspects need to be considered to assess their effectiveness in predicting trends and making reliable decisions.

Accuracy is a fundamental metric that gauges how well a model predicts the correct outcomes. In the context of stock market prediction, this often means forecasting whether a stock price will increase or decrease over a certain period. However, accuracy alone may not suffice, especially in financial data, due to its imbalanced nature. For instance, in many datasets, the number of periods with price increases may far outweigh the periods with price decreases. In such cases, metrics like precision and recall, commonly used in classification tasks, provide a more nuanced understanding of the model's ability to correctly identify price movements. Precision assesses the proportion of true positive predictions among all predicted positives, while recall evaluates the ability of the model to identify all true positives.

In addition to accuracy, performance metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared ($R^2$) are pivotal for evaluating prediction quality. RMSE measures the average magnitude of error by calculating the square root of the average squared differences between predicted and actual values. A lower RMSE indicates that the model's predictions are closer to the real stock prices. MAE provides another measure of prediction error by calculating the average absolute difference between the predicted and actual values, offering a more intuitive sense of the model's prediction performance. $R^2$, on the other hand, assesses how well the model explains the variation in stock prices. A higher $R^2$ value suggests that the model captures the underlying patterns of the data more effectively[25].

Furthermore, the reliability and generalization of a model are crucial for stock market prediction, as market conditions are constantly evolving. Reliability refers to how consistently a model can deliver accurate results over time. Generalization, the ability of a model to perform well on unseen data, is tested through techniques like cross-validation and out-of-sample testing. Cross-validation helps in ensuring that the model does not overfit to the training data, which can result in great performance on training sets but poor performance on new, unseen data. To combat overfitting, methods like regularization (e.g., Lasso or Ridge regression) can be applied, which penalize overly complex models, forcing them to generalize better to new data.

Lastly, computational efficiency is another crucial aspect, especially for real-time stock prediction. Given the large volume of financial data and the need for quick decision-making in stock markets, models must be able to process and output predictions rapidly. Techniques like deep learning can be highly accurate, but they require substantial computational resources, such as powerful GPUs or cloud infrastructure, for training and inference. Thus, there is a trade-off between predictive accuracy and computational efficiency. Models must strike a balance to ensure that they provide real-time, reliable predictions without overwhelming system resources[26].

## VII. CONCLUSION

In conclusion, the paper underscores the potential of machine learning and sentiment analysis in improving stock market prediction accuracy. By combining historical stock data with sentiment data from news sources and social media platforms, investors can gain deeper insights into market trends. The integration of these advanced techniques offers an improved model for decision-making, reducing valuation errors and enhancing investment outcomes. Despite challenges such as data reliability and the need for computational efficiency, the study demonstrates that a hybrid approach leveraging machine learning, sentiment analysis, and financial analysis holds promise for more accurate stock market forecasting, thus aiding investors in making well-informed decisions. Future work may focus on refining these models and addressing the computational challenges associated with large-scale, real-time stock predictions**.**

## VII. REFRENCES


[1] S. Kumar, A. Chaturvedi, A. Kumar, and C. Gupta, "Optimizing BLDC Motor Control in Electric Vehicles Using Hysteresis Current Controlled Boost Converters," *Proc. - 2024 IEEE 16th Int. Conf. Commun. Syst. Netw. Technol. CICN 2024*, pp. 743–748, 2024, doi: 10.1109/CICN63059.2024.10847341.

[2] A. Kumar and S. Jain, "Multilevel Inverter with Predictive Control for Renewable Energy Smart Grid Applications," *Int. J. Electr. Electron. Res.*, vol. 10, no. 3, pp. 501–507, 2022, doi: 10.37391/IJEER.100317.

[3] C. Gupta and V. K. Aharwal, "Optimizing the performance of Triple Input DC-DC converter in an Integrated System," *J. Integr. Sci. Technol.*, vol. 10, no. 3, pp. 215–220, 2022.

[4] S. Kumar, A. Kumar, C. Gupta, A. Chaturvedi, and A. P. Tripathi, "Synergy of AI and PMBLDC Motors: Enhancing Efficiency in Electric Vehicles," *IEEE Int. Conf. "Computational, Commun. Inf. Technol. ICCCIT 2025*, pp. 68–73, 2025, doi: 10.1109/ICCCIT62592.2025.10927757.

[5] C. Gupta and V. K. Aharwal, "Design and simulation of Multi-Input Converter for Renewable energy sources," *J. Integr. Sci. Technol.*, vol. 11, no. 3, pp. 1–7, 2023.

[6] A. Kumar and S. Jain, "Predictive Switching Control for Multilevel Inverter using CNN-LSTM for Voltage Regulation," vol. 11, pp. 1–9, 2022.

[7] S. Kumar, A. Kumar, C. Gupta, and A. Chaturvedi, "Future Trends in Fault Detection Strategies for DC Microgrid," *Proc. - 2024 IEEE 16th Int. Conf. Commun. Syst. Netw. Technol. CICN 2024*, pp. 727–731, 2024, doi: 10.1109/CICN63059.2024.10847358.
[8] C. Gupta and V. K. Aharwal, "Design of Multi Input Converter Topology for Distinct Energy Sources," *SAMRIDDHI*, vol. 14, no. 4, pp. 1–5, 2022, doi: 10.18090/samriddhi.v14i04.09.
[9] A. Kumar and S. Jain, "Enhancement of Power Quality with Increased Levels of Multi-level Inverters in Smart Grid Applications," vol. 14, no. 4, pp. 1–5, 2022, doi: 10.18090/samriddhi.v14i04.07.
[10] A. Kumar and S. Jain, "Critical Analysis on Multilevel Inverter Designs for," vol. 14, no. 3, 2022, doi: 10.18090/samriddhi.v14i03.22.
[11] C. B. Singh, A. Kumar, C. Gupta, S. Cience, T. Echnology, and D. C. Dc, "Comparative performance evaluation of multi level inverter for power quality improvement," vol. 12, no. 2, pp. 1–7, 2024.
[12] P. Mahapatra and C. Gupta, "Study of Optimization in Economical Parameters for Hybrid Renewable Energy System," *Res. J. Eng. Technol. ...*, vol. 03, no. 02, pp. 63–65, 2020, [Online]. Available: http://www.rjetm.in/RJETM/Vol03_Issue02/Study of Optimization in Economical Parameters for Hybrid Renewable Energy System.pdf
[13] S. Khan, C. Gupta, and A. Kumar, "An Analysis of Electric Vehicles Charging Technology and Optimal Size Estimation," vol. 04, no. 04, pp. 125–131, 2021.
[14] A. Raj, A. Kumar, and C. Gupta, "Shunt Active Filters : A Review on Control Techniques II . Shunt Active Power Filter," vol. 05, no. 02, pp. 78–81, 2022.
[15] B. B. Khatua, C. Gupta, and A. Kumar, "Harmonic Investigation Analysis of Cascade H Bridge Multilevel Inverter with Conventional Inverter using PSIM," vol. 04, no. 03, pp. 9–14, 2021.

[16] Mehra, M. (2018). *Valuation models for stock market analysis: A comparison of risk-based, discounted, and relative models*. *Journal of Investment Analysis, 24*(4), 78-95.
[17] Uysal, A., et al. (2019). *Deep learning for stock market prediction: A comparison of neural network approaches*. *International Journal of Machine Learning and Applications, 6*(3), 56-73.
[18] Zhang, Y., & Chen, L. (2020). *The impact of social media sentiment on stock market performance in 2020*. *Journal of Financial Markets, 28*(1), 115-134.
[19] Kumar, S., & Gupta, R. (2021). *A hybrid model for stock market prediction using machine learning and sentiment analysis*. *Computational Finance and Economics, 34*(2), 99-112.
[20] Lee, H., & Kim, S. (2021). *Stock price forecasting using deep learning models: A comparative study*. *Journal of Financial Engineering and Technology, 7*(3), 77-91.
[21] Li, Z., & Wang, J. (2018). *Machine learning techniques in stock price prediction: A review of recent progress*. *International Journal of Data Science and Machine Learning, 5*(1), 56-74.
[22] Singh, A., & Patel, P. (2022). *Combining technical analysis and machine learning for stock market prediction: Advances in methodology*. *Journal of Finance and Investment Strategies, 15*(5), 205-221.
[23] Chandra, S., & Verma, N. (2023). *Financial sentiment analysis using deep learning: Application to the Indian stock market in 2023*. *Journal of Business and Financial Analytics, 21*(4), 88-101.
[24] Nair, V., & Singh, M. (2022). *Evaluation of machine learning models for stock price prediction in volatile markets*. *International Journal of Computational Finance, 11*(2), 34-48.
[25] Thomas, G., & Saha, P. (2024). *Financial market prediction using hybrid models of sentiment analysis and machine learning techniques: A new approach*. *Journal of Financial Modeling and Analytics, 18*(3), 123-135.
[26] Patel, R., & Rathi, S. (2025). *Enhancing stock market prediction using artificial intelligence and financial sentiment analysis: A 2025 review*. *Journal of Financial AI and Machine Learning, 2*(1), 50-63



inprotected.com