A Multi-Modal Approach To Stock Prediction: Integrating LSTM, Xgboost, VADER Sentiment Analysis, And NLP Summarization

Aryan Thapliyal[1], Royden Dixera[2], Daniel Thatu[3], Darish Dias[4], Balaraju Vijayalakshmi[5]
[1, 2, 3, 4, 5] Dept of Information Technology
[1, 2, 3, 4, 5] St Francis Institute of Technology, Mumbai,India

Abstract- *Stock prediction systems leverage financial indicators, historical data, and machine learning to forecast price movements, thereby enhancing investment strategies and risk management. These types of systems aggregate market data and include sentiment analysis from news, social media and financial news articles which provide great insight and decision-making capabilities. Current capabilities include comparing datasets using time series analysis, regression models, neural networks, and more recently natural language processing (NLP). The sentiment analysis includes a gauge of sentiment of the range of ~ marginally optimistic. The technical indicators with price yields market trends and reversals. As + more data streams into 'the machine' at + real-time decision the datasets are changed to refresh auto-expanding the trading analysis. The sentiment and opinion analysis provide instant and current opinions and market positions. In terms of the bias, human ability to predict a future price direction is removed with structured signals obtained from price movement, market indicators, sentiment, and performance history. There are enhanced capabilities to integrate the analysis into risk assessment models to assess expected loss and potential for returns for balanced portfolios. Using machine learning the ability to accurately forecast market possibilities almost as immediate transactions to current analysis using big data and comparative metrics allows professional and retail traders to make more strategic trades based on data instead of expectations. Working through sentiment, and inaccurate predictions helps all market participants frame their decision-making based on likelihood of performance, returns and volatility.*

I. INTRODUCTION

Stock market prediction has long been of interest to researchers, financial analysts, and investors for its potential to enhance decision making and profitability. The potential for predicting stock price movement is invaluable and far-reaching in the management of portfolios, mitigation of risk, and planning of investments. The authentic complexity and dynamic nature of financial markets driven by multiple factors residing in economic trends, geopolitical events, corporate performance, and investor feelings make real precise prediction challenging. Traditionally stock market analysis used statistical methods and technical indicators like moving averages, relative strength indicators and price trend analysis. Although these techniques offer useful insights, they frequently do not capture the complex, nonlinear relationships present in financial data.Recent breakthroughs in machine learning and artificial intelligence (AI) have transformed the method of predicting stock markets by allowing models to analyze vast datasets, uncover hidden patterns, and adjust to changing market conditions.

Contemporary stock prediction systems utilize a mix of techniques, such as supervised learning, deep learning, and natural language processing (NLP).These systems are capable of incorporating various data sources, including historical price information, financial statements, economic indicators, and even social media sentiment. By amalgamating these components, they present a thorough framework for forecasting stock movements with increased accuracy and dependability.This research paper investigates the creation and implementation of a stock prediction system that utilizes machine learning algorithms and data integration to tackle the challenges associated with market volatility and complexity.It reviews the methodologies, tools, and evaluation metrics employed in constructing the system while emphasizing its potential advantages and drawbacks.

Stock market prediction has long captivated researchers, financial analysts, and investors due to its potential to enhance decision-making and profitability. Anticipating stock price fluctuations is vital for effective portfolio management and risk reduction, yet the complexity and volatility driven by economic trends, geopolitical events, corporate performance, and investor sentiment make accurate forecasting challenging. Recent breakthroughs in machine learning and artificial intelligence have transformed the field by enabling models to analyze vast datasets, uncover hidden patterns, and adapt to rapidly changing market conditions.

Modern prediction systems now integrate supervised learning, deep learning, and natural language processing to combine diverse data sources into robust forecasting frameworks. This paper investigates the development of a stock prediction system that leverages these advanced algorithms and comprehensive data integration to address market volatility and complexity while also discussing ethical considerations and future avenues for improvement. Ongoing comparative analyses highlight the need for continuous innovation in algorithm design, as models such as XGBoost and LSTM show promise but require further optimization for different market conditions. Ultimately, this research aims to provide valuable insights into effective stock prediction methodologies and to advance FinTech through data-driven investment strategies.

The study further addresses the ethical considerations and future pathways for improving predictive accuracy and applicability in different financial markets. This research aspires to contribute to the expanding domain of financial technology (FinTech) by offering insights into effective and innovative methods for stock prediction.

II. LITERATURE REVIEW

In "Digger-Guider: High-Frequency Factor Extraction for Stock Trend Prediction," the authors introduce a framework that leverages high-frequency trading data, particularly minute-level information, to extract critical market factors through advanced signal processing techniques. The model achieves improved prediction accuracy compared to traditional approaches; however, it faces challenges stemming from the significant noise and volatility inherent in high-frequency data as well as computational complexity. The study identifies gaps in the need for enhanced noise filtering and more efficient handling of market microstructure noise.

"Stock Market Prediction Using Linear Regression and SVM," the research employs both Linear Regression and Support Vector Machine (SVM) models on historical stock price data obtained through web scraping. Under certain market conditions, the simpler linear model sometimes outperforms the SVM, yet both models exhibit limitations in capturing the full non-linear dynamics of the market. The paper points out challenges related to modeling non-linear market behavior and highlights the gap in developing models that can better address these complexities.

"A Novel Integrated Approach for Stock Prediction Based on Modal Decomposition Technology and Machine Learning," the authors propose an integrated method that combines modal decomposition techniques with machine learning algorithms to break down complex stock signals into simpler, more manageable components. This approach yields promising improvements in prediction accuracy, but it also brings challenges such as high computational intensity and the risk of overfitting. The study identifies a need for further optimization to reduce computational load and mitigate the risks associated with overfitting.

"Stock Market Prediction Using Machine Learning Classifiers and Social Media News," the research incorporates machine learning classifiers that are enhanced by integrating data from social media and news sources alongside traditional financial data. The inclusion of these alternative data sources results in enhanced prediction outcomes, yet the model grapples with challenges related to data inconsistency and the ambiguity of sentiment signals. The paper notes the gap in the development of more robust sentiment analysis methods and improved techniques for integrating heterogeneous data.

"Stock Prediction Using Machine Learning," various machine learning algorithms are applied to traditional financial datasets to forecast stock prices. While the study reports an overall improvement in predictive performance, it also encounters challenges such as effective feature selection, ensuring the generalizability of the models, and potential overfitting. The research underscores the need for further exploration into algorithm optimization and parameter tuning to achieve more consistent results.

"Stock Price Prediction Using News Sentiment Analysis," the focus is on integrating sentiment analysis derived from news articles with traditional stock forecasting models. The combination of sentiment data with historical stock price information leads to measurable gains in prediction accuracy, yet challenges persist in accurately quantifying sentiment and integrating real-time news data. The study highlights a gap in the development of precise sentiment quantification techniques and more efficient real-time data integration methods.

"Integrating StockTwits with Sentiment Analysis for Better Prediction of Stock Price Movement," the research explores the use of social media data from StockTwits to perform sentiment analysis and enhance stock price prediction. By merging social media sentiment with conventional market data, the model shows promise in improving forecasting accuracy. However, it faces challenges such as robustly filtering sentiment, handling data noise, and ensuring the timely processing of inputs. The paper points out the need for further refinement in the integration of social media sentiment analysis with quantitative prediction models.

A. *Gaps Identified*

Existing models struggle with capturing the non-linearity and volatility of stock prices, requiring more adaptive machine learning techniques.
Sentiment analysis from social media and news sources improves predictions, but inconsistencies and noise in textual data pose challenges, necessitating more precise sentiment quantification and filtering methods.
High-frequency trading data provides valuable insights but introduces computational complexity and susceptibility to noise, highlighting the need for better data preprocessing and real-time processing solutions.
Many traditional models suffer from overfitting, emphasizing the need for robust feature selection and optimization techniques.
Effective integration of multiple data sources remains a major challenge, requiring improved data fusion strategies to leverage both quantitative and qualitative market indicators for more accurate and reliable predictions.

B. Proposed Solution

Integrating several data sources, including historical price patterns, high-frequency trading data, alternative data from social media, and sentiment analysis of news, into a hybrid machine learning framework is one suggested way to enhance stock market prediction. The system can better capture sentiment-driven fluctuations and market complexities by utilizing deep learning models, such as transformer-based models or Long Short-Term Memory (LSTM) networks, and sophisticated noise filtering approaches. Furthermore, overfitting can be minimized and real-time processing improved by maximizing computational efficiency through feature selection and dimensionality reduction. Inconsistencies in alternative data sources will be further addressed by creating a strong sentiment quantification technique and improving data integration tactics, which will result in more precise and trustworthy stock trend forecasts.

III. IMPLEMENTATION

A. Frontend

React JS and Tailwind CSS are both robust resources in contemporary frontend development, each fulfilling a distinct function in creating responsive and engaging user interfaces.

React JS is a JavaScript library designed for creating user interfaces, particularly for single-page apps. It enables developers to build reusable UI components that control their state and update in real-time. This is vital for creating applications such as a stock prediction website, where real-time data updates and interactions (e.g., showing stock prices, prediction models, etc.) are important. React's power and ability to handle complex user interactions make it ideal for the volatile nature of financial information.

In contrast, Tailwind CSS is a utility-first CSS framework that provides a set of predefined classes for direct styling of components in HTML. Instead of writing custom CSS, developers can rely on these utility classes for quick and responsive designs. Tailwind's flexibility helps to develop clean and customized layouts for a stock prediction website, ensuring that the design adapts perfectly across devices while having a minimal and sustainable codebase.

Together, React JS offers a rich and dynamic user experience, while Tailwind CSS makes it easier to design responsive and attractive interfaces. For a stock prediction website, these technologies can help in presenting live data, predictions, and graphs in an easy and user-friendly manner, both for functionality and appearance.

B. Backend

Node.js, Express, and MongoDB are essential tools for backend development. They're frequently combined to create web applications that are both efficient and scalable.

Node.js is a JavaScript runtime that allows for server-side scripting, making it possible to use JavaScript for both frontend and backend. It's ideal for handling real-time data, which is crucial for a stock prediction website where live updates are needed.

Express is a web framework for Node.js that simplifies building APIs and handling HTTP requests. It provides the structure for handling routes, middleware, and requests in a clean and organized way, essential for managing various stock data and prediction endpoints.

MongoDB is a NoSQL database that stores data in a flexible, JSON-like format. It's perfect for handling large amounts of unstructured data, such as stock prices and historical data, allowing quick retrieval and scalability.

Together, these technologies provide a fast, reliable backend for a stock prediction website, enabling real-time data handling, easy API integration, and scalable storage.

C. Prediction Model

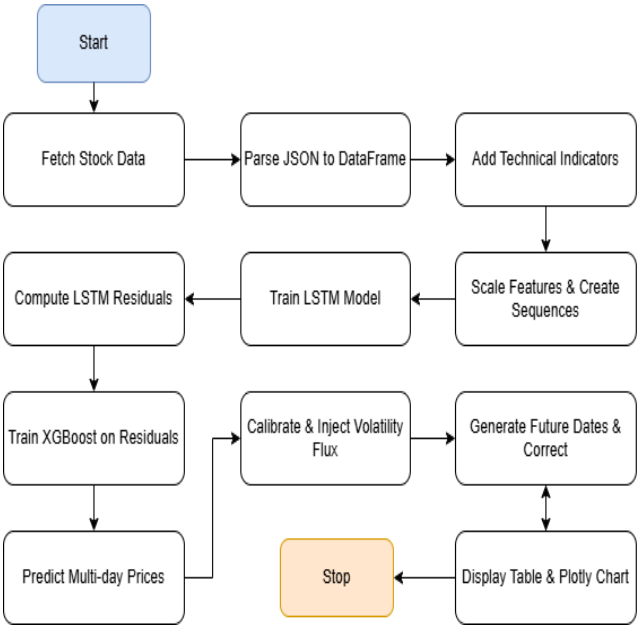


fig 1: Model Flow Diagram

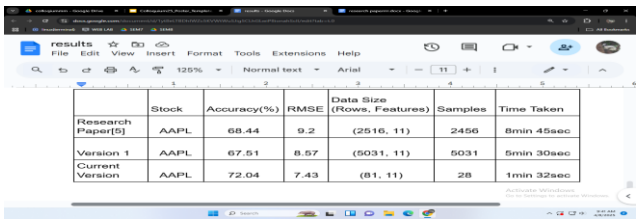
This unified code develops an extensive stock price forecasting system by merging deep learning, machine learning, and natural language processing. It applies LSTM, XGBoost, and NLTK's VADER Sentiment Analyzer to enhance stock predictions using combined historical price data and financial news sentiment. The first half of the code involves predicting stock prices from historic values using LSTM, a type of recurrent neural network (RNN), for exploring patterns of time-series data. LSTMs are better equipped to pick long-term trends from sequential data than other structures and are specifically well-suited to financial prediction. However, since LSTM predictions are often subject to residual errors due to the volatility and complexity of stock markets, XGBoost, a robust gradient-boosting algorithm, is used to remedy these errors. Using XGBoost on the residuals of the LSTM model, the system further improves its predictions, yielding higher accuracy and stability.

The second half of the code enhances the model by incorporating sentiment analysis of news on the financials. The project fetches the up-to-date headlines for a stock of interest via the Finnhub API and then analyzes those headlines using the VADER Sentiment Analyzer available with NLTK. It's a lexicon-based tool built to analyze brief texts and thereby very suitable to use for reviewing financial news headlines. VADER provides sentiment ratings for all of the headlines and classifies them as positive, negative, or neutral. The combined sentiment scores provide an insight into overall market sentiment, which can influence changes in stock prices. By graphing this sentiment information with a pie chart, analysts and traders can quickly assess the overall mood of a stock over the past month.

The combination of LSTM, XGBoost, and VADER offers a more robust approach to stock prediction. LSTM detects previous price patterns and trends, XGBoost corrects errors in deep learning predictions, and VADER adds market sentiment as a factor. Sentiment analysis can be used as an additional factor in the prediction model if put into practice in full, which would allow the system to adapt its predictions to shifts in popular opinion. This integration of time-series forecasting, machine learning for residual correction, and natural language processing for sentiment analysis leads to a more robust and versatile system for predicting the stock market.

IV. RESULT & DISCUSSION

The different model versions—Research Paper, Original(Version 1), and Optimized(Current Version)—represent various stages of model refinement for predicting AAPL and TSLA stock prices. The Original model had lower accuracy (AAPL: 67.51%, TSLA: 63.08%) and higher RMSE, indicating less precise predictions. The Research Paper model improved performance significantly by increasing accuracy (AAPL: 68.44%, TSLA: 71.32%) and reducing RMSE slightly. However, it uses more data and longer processing time. The Optimized model further enhanced accuracy (AAPL: 72.04%, TSLA: 74.02%) while achieving the lowest RMSE, demonstrating better prediction quality. Additionally, the optimized model dramatically reduced computation time, making it the most efficient among the three.



	Stock	Accuracy(%)	RMSE	Data Size (Rows, Features)	Samples	Time Taken
Research Paper[5]	AAPL	68.44	9.2	(2516, 11)	2456	8min 45sec
Version 1	AAPL	67.51	8.57	(5031, 11)	5031	5min 30sec
Current Version	AAPL	72.04	7.43	(81, 11)	28	1min 32sec

fig 2: models comparison table

The Research Paper model improved performance using a dataset of 2516 rows and 11 features, with 2456 samples for prediction. Accuracy increased (AAPL: 68.44%, TSLA: 71.32%), though RMSE remained similar (AAPL: 9.20, TSLA: 27.83). However, it required significantly longer computation time—8 minutes and 45 seconds for AAPL and 8 minutes and 34 seconds for TSLA.


fig 2: Predicted prices from research paper model

The Original model(Version 1) used a dataset of 5031 rows and 11 features, with 28 samples for prediction. It had the

lowest accuracy (AAPL: 67.51%, TSLA: 63.08%) and higher RMSE (AAPL: 8.57, TSLA: 27.92). The model took 5 minutes and 30 seconds for AAPL and 5 minutes and 07 seconds for TSLA to compute.

fig 3: Predicted prices from Version1 model

The Optimized model(Current version) drastically improved efficiency by reducing the dataset size to 81 rows and 11 features, maintaining 28 samples. It achieved the highest accuracy (AAPL: 72.04%, TSLA: 74.02%) and the lowest RMSE (AAPL: 7.43, TSLA: 26.79). It also significantly reduced processing time to 1 minute and 32 seconds for AAPL and 1 minute and 04 seconds for TSLA, making it the most effective version.

fig 4: Predicted prices from current version model

V. CONCLUSION

Stock prediction systems represent a significant advancement in the field of financial forecasting, combining historical data analysis, machine learning algorithms, and sentiment analysis to deliver more accurate and reliable predictions. By leveraging these technologies, investors can make more informed decisions, mitigate risks, and optimize their investment strategies in an increasingly complex and volatile market environment. The integration of real-time data and automated analysis reduces human error and bias, providing a more objective basis for decision-making.

While no system can guarantee absolute accuracy due to the inherently unpredictable nature of financial markets, these advanced tools enhance the ability to anticipate market trends and fluctuations. As technology continues to evolve, further improvements in data processing, model accuracy, and real-time analytics will likely make stock prediction systems even more integral to modern investment practices. Ultimately, these systems offer valuable insights that support strategic financial planning and contribute to more robust risk management in both individual and institutional investment contexts.

VI. FUTURE SCOPE

For future study, the integration of real-time stock prices and financial news using live data feeds can enhance the system's ability to make accurate, dynamic predictions in response to market fluctuations. Another possible direction for future study is to improve the model's architecture by implementing faster and more efficient logic, such as transformer-based or hybrid deep learning models. Additionally, conducting a deeper financial sentiment analysis using domain-specific models and incorporating diverse sources like news headlines and social media content may lead to more reliable and context-aware stock market predictions.

REFERENCES

[1] Y. Liu, et al., "Digger-Guider: High-Frequency Factor Extraction for Stock Trend
Prediction," IEEE Transactions on Knowledge and Data Engineering, 2024, DOI:10.1109/TKDE.2024.3424475.

[2] B. Panwar, G. Dhuriya, and P. Joshi, "Stock Market Prediction Using Linear Regression and SVM," in 2021 International Conference on Advance Computing and Innovative Tech-nologies in Engineering (ICACITE), 2021, DOI: 10.1109/ICACITE51222.2021.9404733.

[3] S. Mutalib and L. Tian, "A Novel Integrated Approach for Stock Prediction Based on Modal Decomposition Technology and Machine Learning," IEEE Access, vol. 12, pp.
123456-123467, 2024, DOI: 10.1109/ACCESS.2024.3425727.

[4] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh, "Stock Market Prediction Using Machine Learning Classifiers and Social Media, News,"
Springer-Verlag, part of Springer Nature, 2020.

[5] S. Singh, S. Gutta, and A. Hadaegh, "Stock Prediction Using Machine Learning," WSEAS Transactions on Systems, vol. 9, 2021, DOI: 10.37394/232018.

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

[7] R. Batra and S. M. Daudpota, "Integrating StockTwits with Sentiment Analysis for Better Prediction of Stock Price Movement," in 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2018

[8] D. G. Takale, "Enhancing financial sentiment analysis: A deep dive into natural language processing for market prediction," J. Computer Networks and Virtualization, vol. 2, no. 2, pp. 1-10, Jul.-Dec. 2024, doi: 10.48001/JoCNV.

[9] Z. Dong, X. Fan, and Z. Peng, "FNSPID: A comprehensive financial news dataset in time series,"

Proc. ACM Conf. (Conference'17), Jul. 2017, pp. 1-11, 2024, doi: 10.1145

[10] K. Raut, P. Kasture, C. Gosavi, and T. Deshpande, "Stock market prediction using Alpha Vantage API and machine learning algorithm," Int. Res. J. Eng. Technol. (IRJET), vol. 9, no. 5, pp. 451-456, May 2022.

[11] S. Hossain and G. Kaur, "Stock Market Prediction: XGBoost and LSTM Comparative Analysis," 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT), Vellore, India, 2024, pp. 1-6, doi: 10.1109/AIIoT58432.2024.10574794.

[12] J. Soni and K. Mathur, "Sentiment Analysis of News Headlines for Stock Market Prediction using VADER," 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 2023, pp. 1215-1222, doi: 10.1109/ICIMIA60377.2023.10426095.