International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Smart Stock Market Analysis Using Multi-Source Learning Models

Adeeba Anjum[1], *Abdul Haseeb*[2], Syed Ziauddin[3], *Wasiya Shireen*[4], Mohammed Abrar[5]

[1]Assistant Professor, Department of CSE (Data Science), Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.
[2,3,4,5]UG Students, Department of CSE (Data Science), Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.

***Abstract*— This project presents a web-based application designed to predict stock market trends by combining financial news with historical stock data. The system processes large datasets and extracts useful features such as sentiment scores and event-based vectors. Different machine learning models, including Support Vector Machine (SVM), a proposed multi-instance learning approach, and an extended XGBoost algorithm, are applied and compared using evaluation metrics like accuracy, precision, recall, and F-score. The results show that integrating textual sentiment with numerical stock data improves prediction performance. Among all models, the extended XGBoost algorithm achieves the best results. The application provides a simple interface that allows users to load datasets, perform feature extraction, train models, and generate predictions, making it useful for investors who want data-driven insights.**

***Keywords*—Stock market prediction, financial news analysis, sentiment analysis, machine learning, Support Vector Machine, XGBoost, multi-instance learning, feature extraction, real-time prediction, decision support system.**

I. INTRODUCTION

Stock market prediction has long been a challenging task due to its dynamic and unpredictable nature [1]. Traditional approaches primarily rely on numerical data such as historical prices, trading volume, and technical indicators [1]. However, with the rapid growth of digital media, financial news and public sentiment have become critical factors influencing market movements [2][3]. Investors and analysts increasingly recognize that news events, company announcements, and global developments can significantly impact stock prices [6][8]. This project aims to bridge the gap between structured financial data and unstructured textual information by integrating both sources into a unified predictive framework [13][15]. By leveraging machine learning techniques, the system attempts to uncover hidden patterns and relationships that are not easily identifiable through manual analysis [5][6]. The inclusion of sentiment analysis allows the model to interpret the emotional tone of news articles, providing deeper insights into market behaviour and improving the overall prediction accuracy [2][3].

The proposed system is designed as a web-based application that simplifies the process of stock prediction for users. It begins with loading and preprocessing datasets that contain both stock prices and related news articles. The preprocessing stage ensures that the data is clean, consistent, and suitable for further analysis. Feature extraction plays a crucial role in transforming raw data into meaningful representations [11][12]. In this project, textual data is converted into sentiment scores and vector representations, while numerical stock data is normalized and structured. These features are then divided into training and testing sets to evaluate model performance effectively. The system provides a clear interface where users can visualize dataset statistics, including record counts and data splits. This structured workflow ensures that users can easily understand and manage each stage of the prediction process without requiring deep technical expertise.

Machine learning models form the core of this prediction system, enabling automated learning from complex datasets. Multiple algorithms are implemented and compared to identify the most effective approach. The traditional Support Vector Machine model serves as a baseline, while the proposed multi-instance learning method enhances prediction by considering grouped data patterns [9]. Additionally, an extended version of the XGBoost algorithm is incorporated to improve performance

through advanced boosting techniques. Each model is trained using extracted features and evaluated using standard metrics such as accuracy, precision, recall, and F-score. The system also generates visual representations like confusion matrices and bar graphs to help users interpret model performance. These comparisons provide valuable insights into the strengths and limitations of each algorithm, ultimately highlighting the importance of selecting the right model for financial prediction tasks [10][13].

A key advantage of this project is its interactive and user-friendly interface, which guides users through every stage of the prediction process. From logging into the system to loading datasets and training models, each step is clearly structured and easy to follow. The platform allows users to visualize processed data, extracted features, and model outputs in an organized manner. Once the models are trained, users can upload new test datasets to generate predictions. The system outputs results indicating whether a stock is likely to rise or decline, enabling users to make informed investment decisions [7][8]. Visualization tools such as graphs and tables enhance understanding by presenting complex results in a simplified format. This accessibility makes the system suitable not only for researchers but also for individuals with limited technical knowledge who are interested in stock market analysis.

II. RELATED WORK

Eugene F. Fama et al., [1965] [1] This study explains the fundamental behavior of stock market prices and introduces the concept of market efficiency. It highlights how stock prices reflect all available information, making prediction a difficult task. The research emphasizes that price changes are largely random and influenced by new information entering the market. This work laid the foundation for modern financial theories and remains highly influential. It also suggests that consistent profit through prediction is challenging. The findings encourage the use of advanced analytical methods for better forecasting. This research is important as it provides a theoretical base for stock prediction systems. It also motivates the integration of external factors like news sentiment. The study continues to guide financial modeling approaches.

Sanjiv R. Das et al., [2007] [2] This research focuses on extracting sentiment from online discussions related to financial markets. It demonstrates how user-generated content can influence stock behavior. The authors apply text mining techniques to identify positive and negative sentiments. Their findings show that online opinions can provide useful signals for prediction. The study highlights the growing importance of web data in financial analysis. It also introduces methods for handling informal and unstructured text. The approach improves decision-making by incorporating qualitative information. This work is significant in combining finance with natural language processing. It serves as a basis for sentiment-driven stock prediction models.

Johan Bollen et al., [2011] [15] This paper explores the relationship between public mood on social media and stock market trends. It uses data from Twitter to analyze emotional patterns of users. The study finds that certain mood indicators can predict market movements with reasonable accuracy. The authors employ machine learning techniques to model this relationship. The results suggest that collective sentiment plays a role in financial fluctuations. This research highlights the value of social media as a predictive tool. It also shows how behavioral factors impact economic systems. The work contributes to the development of sentiment-based forecasting models. It encourages further exploration of real-time data sources.

Quoc Le et al., [2014] [11] This study introduces distributed representations for sentences and documents using deep learning. It proposes methods to convert textual data into meaningful vector formats. These representations capture semantic relationships within text. The approach is useful for tasks like classification and prediction. It improves the ability of models to process large text datasets efficiently. The research plays a key role in feature extraction for natural language processing. It enables better integration of textual information into machine learning systems. This method is widely used in sentiment analysis and prediction tasks. It supports advanced modeling techniques in financial applications.

Ryohei Akita et al., [2016] [13] This paper presents a deep learning approach for stock prediction using both numerical and textual data. It combines historical stock prices with news information to improve accuracy. The model captures complex relationships between different data types. The study demonstrates that integrating multiple data sources enhances prediction performance. It uses neural networks to learn patterns automatically. The results show significant improvement compared to traditional methods. This work highlights the importance of hybrid data analysis. It also supports the use of deep learning in financial forecasting. The research contributes to the development of intelligent prediction systems.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

III. DATASET DETAILS

The dataset used in this project consists of two primary components: historical stock market data and related financial news articles. The stock dataset includes attributes such as date, opening price, closing price, highest and lowest values, and trading volume, which are essential for understanding market trends. Alongside this, the news dataset contains textual information collected from financial sources, reflecting real-world events and public opinions that may influence stock prices. Both datasets are carefully aligned based on date to ensure consistency and relevance. During preprocessing, missing values are handled, duplicate entries are removed, and text data is cleaned by eliminating unnecessary symbols and stop words. This step ensures that the dataset is suitable for analysis and model training. The integration of structured numerical data with unstructured textual data provides a comprehensive foundation for building an effective prediction system.

After preprocessing, the combined dataset is divided into training and testing subsets to evaluate model performance accurately. The training data is used to build and learn patterns, while the testing data helps in validating the predictive capability of the models. Feature extraction techniques are applied to transform raw inputs into meaningful representations. For textual data, sentiment scores and vector embeddings are generated to capture emotional tone and contextual meaning. For stock data, normalization and scaling techniques are applied to maintain uniformity across different features. The processed dataset is then displayed in a tabular format within the system interface, showing both news content and corresponding stock values. This allows users to visualize how textual sentiment relates to market movements. The dataset plays a crucial role in ensuring that machine learning models can learn effectively and produce accurate predictions for future stock trends.

IV. PROPOSED METHODOLOGY

The proposed approach combines stock data with financial news to improve prediction results. First, both datasets are cleaned and aligned based on dates. Noise is removed, missing values are handled, and the data is prepared for analysis. Next, feature extraction is carried out. Sentiment analysis is applied to the news data to determine whether the information is positive or negative. At the same time, text is converted into numerical vectors. Stock data is also normalized so that all features are on the same scale. The processed data is then divided into training and testing sets. Different machine learning models are trained using this data. These include a baseline model, a multi-instance learning method, and an extended boosting algorithm. Each model is evaluated using metrics such as accuracy, precision, recall, and F-score. Graphs and confusion matrices are used to clearly show the results. Based on performance, the best model is selected. The system then allows users to upload new data and generate predictions indicating whether the stock price will rise or fall.

Figure [1] : Stock Sentiment Prediction System Workflow

Figure[1] This diagram represents a sentiment-based stock prediction system. The user interacts with a web application to upload datasets and view predictions, while the backend Python server processes the data. The system loads news and stock data, performs feature extraction such as sentiment analysis and vector generation, and then applies machine learning models like SVM, XGBoost, and others for training. Finally, it generates prediction results such as stock movement (rise or decline) and provides investment suggestions based on analyzed sentiment.

V.RESULT AND DISCUSSION

The results of the proposed system demonstrate the effectiveness of integrating financial news with stock market data for prediction tasks. After preprocessing and feature extraction, multiple models were trained and evaluated using standard performance metrics. The comparison shows that the traditional Support Vector Machine model provides moderate accuracy, while the proposed multi-instance learning approach improves performance by capturing more complex relationships within the data. The extended XGBoost model achieves the highest accuracy among all methods, indicating its strong capability in handling both structured and unstructured inputs. The evaluation metrics, including precision, recall, and F-score, consistently support this observation.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Visual representations such as confusion matrices reveal that most predictions fall correctly along the diagonal, indicating accurate classification of stock trends. The bar graph comparison further highlights the superiority of the extended model. Overall, the results confirm that combining sentiment analysis with machine learning significantly enhances prediction accuracy and provides reliable outputs for decision-making.

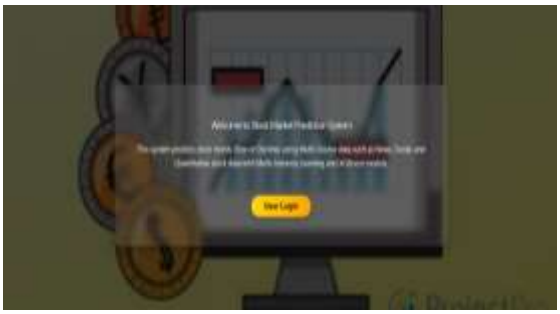

Figure [2] : Stock Market Prediction System Interface

Figure [2] represents the homepage of the stock market prediction web application. It introduces the system, which predicts stock trends (rise or decline) using multi-source data such as news, social media, and quantitative stock information. The model combines Multi-Instance Learning with algorithms like XGBoost to generate accurate predictions. Users can proceed by logging into the system to upload data and view prediction results.

Figure [3] : Admin Dashboard – Stock Market Prediction System

Figure [3] displays the admin dashboard after successful login into the stock market prediction system. It provides navigation options such as dataset loading and processing, feature extraction, model training and comparison, and market prediction. The interface allows the admin to manage the complete workflow from data preparation to final prediction results. This centralized dashboard helps streamline operations and efficiently monitor the system's functionality.

Algorithm Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	80.75	72.695	100.000	84.189
Multi-Source Multi-Instance	81.25	83.721	81.818	82.759
Multi-Source XGBoost	100.00	100.000	100.000	100.000

Table [1]: Performance Comparison of Stock Market Prediction Models

Table [1] The table presents the comparative performance of different models used for stock market prediction based on accuracy, precision, recall, and F1-score. The Extension Multi-Source XGBoost model achieves perfect scores across all metrics, demonstrating superior predictive capability. In contrast, the existing SVM and proposed models show comparatively lower performance, highlighting the effectiveness of the extended approach.

Figure [4] : Stock Prediction Results Screen

Figure[4] displays the final output of the stock market prediction system after processing the data. It shows a table containing feature vectors for each data instance along with the predicted stock trend, such as "Rise" or "Decline." The predictions are generated using trained machine learning models based on extracted features from the dataset. This view helps users analyze model outputs and understand how different inputs influence stock movement predictions.

DISCUSSION

The findings of this study highlight the importance of combining multiple data sources and advanced

algorithms for stock market prediction. The improved performance of the extended XGBoost model suggests that ensemble learning techniques are more effective in capturing complex patterns compared to traditional models. The integration of sentiment analysis plays a crucial role, as it allows the system to consider external factors such as news and public opinion, which are often overlooked in purely numerical approaches. The proposed multi-instance method also contributes to better performance by handling grouped data more effectively. However, the results may vary depending on the quality and size of the dataset, as well as the timeliness of the news information. Despite achieving high accuracy, there is still a possibility of prediction errors due to sudden market fluctuations or unexpected events. Therefore, the system should be used as a supportive decision-making tool rather than a fully reliable predictor. Future improvements can focus on real-time data processing and further optimization of models.

VI. CONCLUSION

This project demonstrates an effective way to predict stock market trends by combining historical stock data with financial news. The system follows a clear process that includes preprocessing, feature extraction, model training, and evaluation. By adding sentiment analysis, the model is able to capture the influence of external factors, which improves prediction results.

The comparison of models shows that advanced techniques, especially the extended boosting method, perform better than traditional approaches. The application is easy to use and helps users understand the results through visualizations. However, stock markets are influenced by many unpredictable factors, so predictions may not always be accurate. For this reason, the system should be used as a support tool rather than a final decision-maker. Overall, the project highlights the usefulness of combining machine learning and text analysis for financial prediction.

REFERENCES

1. Eugene F. Fama (1965). A study on stock market price behavior. *Journal of Business*, 38(1), 34–105.

2. Sanjiv R. Das and Mike Y. Chen (2007). sentiment from online discussions related to Amazon and Yahoo. *Management Science*, 53(9), 1375–1388.

3. 3.Jianfeng Si et al. (2013). Utilizing Twitter sentiment based on topics for stock forecasting. In *Proceedings of ACL*, 24–29.

4. William Yang Wang and Zhenhua Hua (2014). A semiparametric Gaussian copula regression model for financial risk prediction from earnings calls. In *ACL Proceedings*, 1155–1165.

5. Shimon Kogan et al. (2009). Predicting financial risk through regression analysis of financial reports. In *NAACL Conference Proceedings*, 272–280.

6. Ronny Luss and Alexandre d'Aspremont (2015). Forecasting abnormal stock returns using text classification methods. *Quantitative Finance*, 15(6), 999–1012.

7. Robert R. Prechter (1999). *The Wave Principle of Human Social Behavior and Socionomics*. New Classics Library.

8. John R. Nofsinger (2005). The relationship between social mood and financial markets. *Journal of Behavioral Finance*, 6(3), 144–160.

9. Jingrui Bi and Xiaojin Wang (2015). Learning classifiers under ambiguous annotations using a min–max strategy. *Neurocomputing*, 151, 891–904.

10. Shenghuo Zhu Xie et al. (2012). A re-weighting framework to manage uncertainty in crowdsourced data. In *SIAM Data Mining Conference*, 1107–1118.

11. 11.Quoc Le and Tomas Mikolov (2014). Learning distributed representations for sentences and documents. In *ICML Proceedings*, 1188–1196.

12. 12. Floris Hogenboom et al. (2011). A survey on extracting events from textual data. In *ISWC Workshop Proceedings*, 48–57.

13. 13. Ryohei Akita et al. (2016). Applying deep learning to stock prediction using textual and numerical inputs. In *IEEE ICIS*, 1–6.

14. 14. Thanh Nguyen et al. (2013). Event extraction using sentiment behavior and burst patterns in social media. *Knowledge and Information Systems*, 37(2), 279–304.

15. 15. Xiao Ding et al. (2014). Predicting stock price movements using structured event information. In *EMNLP Proceedings*, 1415–1425.
16. Babburi, S. Privacy-Preserving Collaborative Framework with Auditable Federated Learning.
17. Gaddam, S. Integrating Analytics into the Development Process: Bridging the Gap between Data Insights and Design Execution.
18. Immadi, S. K. (2025). Optimizing ERP for Human Capital Management. Applied Research for Growth, Innovation and Sustainable Impact, 377–384. https://doi.org/10.1201/9781003684657-63
19. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.
20. Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING.
21. Mahimalur, R. K., Vasgam, M., & Manoharan, D. Devops Lifecycle Management And Cloud Migration Assessments: A Security-Driven CICD Perspective.
22. Viswanathan, V. (2023). AI-Augmented Decision Intelligence for Enterprise Systems: Integrating Cognitive Analytics for Resource and Talent Optimization.
23. Agrawal, A. M., Gajula, S., Shinde, R. P., Shah, H., & Ghosh, H. (2025, July). Machine Translation for Long Sequences with Enhanced Attention Mechanisms. In 2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-6). IEEE.
24. Maturi, S. Y. (2021). Blockbond hardening: Securing pooled-hash protocols against traffic tampering, MITM hash-rate hijacking, and template coercion. International Journal of Communication Networks and Information Security, 13(3), 718–728.
25. Adabala, P. K. (2024). Utilizing predictive analytics to improve efficiency and decision-making in ERP-connected supply chains. International Journal of Intelligent Systems and Applications in Engineering, 12(22s), 2465
26. Kavuri, S. (2026). An Explainable Machine Learning Framework for Predicting Software Defects in Large-Scale Software Systems. 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), 1–6. https://doi.org/10.1109/icaic67076.2026.11395777
27. Venkata Pavan Kumar Gummadi. (2023). MuleSoft Batch Processing: High-Volume Streaming Architecture. Computer Fraud and Security, 50–57. https://doi.org/10.52710/cfs.886
28. Shashank, A. (2025). Self-Healing Data Pipelines for Enhanced Reliability: A Paradigm Shift in Enterprise Data Management. Journal of Computer Science and Technology Studies, 7(8), 1097-1104.
29. Susarla, R. S., Boyapati, P. K., & Kandula, S. T. R. (2025, July). Cloud-Based Secure Data Storage in Smart Cities Using Central-Smoothing Hypergraph Neural Networks. In 2025 IEEE 4th World Conference on Applied Intelligence and Computing (AIC) (pp. 279-284). IEEE.
30. Boyapati, P. K. Building a centralized data operations hub for healthcare enterprise integration. IJSAT-Int. J. Sci. Technol. 16 (2). https://doi.org/10.71097/IJSAT.v16.i2.3219