Stock Market Prediction Via Multi-Source Multiple Instance Learning

[1]Mrs. M. Thirupathamma,[2]Shaik Nazahath,[3]Nanupathina Nagaraju, [4]Chinthoju Krishna Vamsi, [5]Jinugu Pranay

[1]Assistant Professor, Department of Computer Science & Engineering, Sai Spurthi Institute Of Technology

[2,3,4,5] B. Tech Students, Department of Computer Science & Engineering, Sai Spurthi Institute Of Technology

ABSTRACT

The stock market is influenced by a multitude of dynamic and often interdependent factors, making accurate prediction a complex task. Traditional machine learning models typically rely on single-source data and assume precise instance-level labels, which may not effectively capture the multifaceted nature of financial markets. In this project, we propose a novel approach for stock market prediction using **Multi-Source Multiple Instance Learning (MS-MIL)**, which enables the model to learn from grouped data instances (bags) derived from various heterogeneous sources such as historical stock prices, financial news, social media sentiment, and macroeconomic indicators. By treating each source as a distinct set of instances and aggregating them under the multiple instance learning framework, the model can better handle weak supervision and uncertainty inherent in financial data. Our MS-MIL framework integrates both numerical and textual data, applying advanced feature extraction and attention mechanisms to learn discriminative representations. Experimental results demonstrate that the proposed method achieves superior performance in predicting stock movement direction and market trends compared to traditional learning models. This approach offers enhanced robustness, adaptability, and interpretability, making it a promising tool for investors and analysts in making informed decisions.

Keywords: Stock Market Prediction, Multi-Source Learning, Multiple Instance Learning (MIL), Financial Time Series Analysis, Sentiment Analysis, Machine Learning, Deep Learning, Feature Extraction, Attention Mechanism, Market Trend Prediction.

I. INTRODUCTION

The stock market is a dynamic and complex environment driven by a combination of economic, political, psychological, and social factors. Predicting its behavior is a longstanding challenge in the financial and machine learning communities. Accurate stock market prediction holds immense value for investors, financial institutions, and policymakers, as it can inform decision-making and reduce financial risks. However, traditional approaches often rely on single-source data, such as historical stock prices or technical indicators, which may not fully capture the underlying patterns and external influences affecting market movement.

Recent advancements in data collection and computational techniques have opened new possibilities for leveraging diverse data sources—including financial news, social media sentiment, and macroeconomic indicators—in predictive modeling. Despite this progress, integrating such heterogeneous data remains a significant challenge due to variations in format, frequency, and reliability.

To address this, we propose a novel predictive framework based on Multi-Source Multiple Instance Learning (MS-MIL). Multiple Instance Learning (MIL) is a form of weakly supervised learning where labels are assigned to bags (groups of instances) rather than individual instances. By extending MIL to support multiple data sources, our model is capable of learning complex relationships between different modalities of information and how they collectively influence stock prices.

The MS-MIL approach enables the model to treat various sources—such as historical market data, news sentiment scores, and public opinion on social platforms—as separate but interconnected components of a learning process. This design enhances the model's ability to detect subtle patterns and interdependencies that traditional models may overlook. By employing attention mechanisms and advanced neural architectures, the model focuses on

the most relevant features from each data source to make informed predictions.

This project demonstrates the potential of MS-MIL to improve the accuracy and interpretability of stock market forecasting, providing a more holistic and data-driven tool for financial analysis.

II. LITERATURE SURVEY

1. Title:

Stock Movement Prediction from Tweets and Historical Prices

Author(s): Ding, Xiaowen; Zhang, Yue; Liu, Ting

Description:

This paper explores the integration of textual sentiment from Twitter with historical price data to predict stock movements. It proposes a hybrid model that combines natural language processing (NLP) techniques with time-series forecasting. The results show that including sentiment data improves prediction accuracy compared to price-only models.

2. Title:

Multiple Instance Learning for Stock Selection

Author(s): Li, Wenbin; Zhou, Zhi-Hua

Description:

This study introduces a multiple instance learning (MIL) approach to stock selection, where instances are daily features and bags are time periods. The MIL model identifies valuable patterns over time without requiring precise labeling at each point. This demonstrates the effectiveness of MIL in financial applications where granular labels are unavailable.

3. Title:

Sentiment-Aware Stock Market Prediction Using Deep Learning

Author(s): Xu, Yang; Cohen, William W.

Description:

This paper presents a deep learning framework that incorporates sentiment analysis from financial news and forums. It uses recurrent neural networks to model the temporal dependencies of stock data. The fusion of text sentiment with market indicators leads to more robust predictions.

4. Title:

Multi-Modal Learning for Stock Prediction Using News, Historical Data, and Social Media

Author(s): Ghoshal, Palash; Roberts, Steven

Description:

This research investigates multi-modal data fusion for stock prediction, using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to extract features from different sources. The model integrates numerical and textual data, achieving improved prediction results by exploiting inter-source relationships.

5. Title:

Attention-Based Deep Multiple Instance Learning

Author(s): Ilse, Maximilian; Tomczak, Jakub M.; Welling, Max

Description:

Although not specific to finance, this foundational MIL paper introduces an attention-based deep learning mechanism to assign importance to instances within a bag. This method significantly boosts the interpretability and performance of MIL models, and can be directly adapted for stock prediction using multi-source data.

III. EXISTING SYSTEM

In the domain of stock market prediction, various traditional and machine learning-based systems have been developed, most of which primarily rely on single-source data such as historical stock prices and technical indicators. These systems typically use supervised learning models that require clean and well-labeled datasets, with a one-to-one correspondence between input features and output labels. While effective to some extent, such systems suffer from several limitations when applied to the complex and volatile nature of the financial markets.

Traditional Statistical Models:

Models such as ARIMA (AutoRegressive Integrated Moving Average) and GARCH (Generalized Autoregressive Conditional Heteroskedasticity) are widely used for time series forecasting. These models focus only on historical numerical data and are unable to capture external factors like news or public sentiment.

Machine Learning Approaches:

Machine learning models such as Linear Regression, Decision Trees, Support Vector Machines (SVMs), and Random Forests have been used to predict stock prices or classify movement trends. These methods

perform better than statistical models when nonlinear relationships are present but still depend heavily on single-source structured data (e.g., past stock prices, volumes).

Deep Learning Methods:

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models have gained popularity for their ability to model sequential data. However, they are mostly trained on stock price sequences alone and struggle when faced with unstructured data like news or tweets unless explicitly augmented.

Sentiment-Enhanced Models:

Some advanced models attempt to include textual sentiment data from news articles or social media (e.g., Twitter). These models generally perform sentiment analysis using NLP techniques and integrate the results with stock price data. However, they usually treat all data sources as part of a single input vector, failing to properly model the multi-source and multi-instance nature of financial signals.

IV. PROPOSED SYSTEM

To overcome the limitations of traditional stock prediction models, this project proposes a Multi-Source Multiple Instance Learning (MS-MIL)-based system that leverages diverse and heterogeneous data sources to improve predictive accuracy and robustness. Unlike conventional approaches that depend on a single, fixed feature set, the proposed system considers multiple data sources as independent yet complementary channels, enabling the model to learn richer and more context-aware representations of stock market behavior. In the MS-MIL framework, each prediction is modeled as a "bag" of instances collected from various sources, including historical stock price data (such as open, close, volume, and technical indicators), financial news articles and headlines analyzed through sentiment techniques, social media posts reflecting public opinion, and macroeconomic indicators like interest rates and GDP. Instead of labeling individual data points, the entire bag is assigned a label (e.g., stock price movement up or down), allowing the model to effectively handle weak supervision and noisy data. Advanced attention-based deep learning mechanisms are then employed to identify the most influential instances across all sources, enhancing both the predictive power and interpretability of the system.

V. SYSTEM ARCHITECTURE

The diagram illustrates a structured approach to extracting meaningful information from textual data and transforming it into a format suitable for machine learning. Initially, the text is broken down into key components such as the subject and object, each of which may have associated modifiers that provide additional context or detail. These elements are then linked through a connection, representing the relationship between the subject and object. Together, these components form a structured event, which captures the semantic meaning of the original text in a more organized and machine-readable form. This structured representation is then passed into an RBM (Restricted Boltzmann Machine), a type of neural network used for feature learning and pattern extraction. The RBM processes these structured events to learn hidden patterns and representations, which can later be used for tasks such as prediction, classification, or deeper analysis.

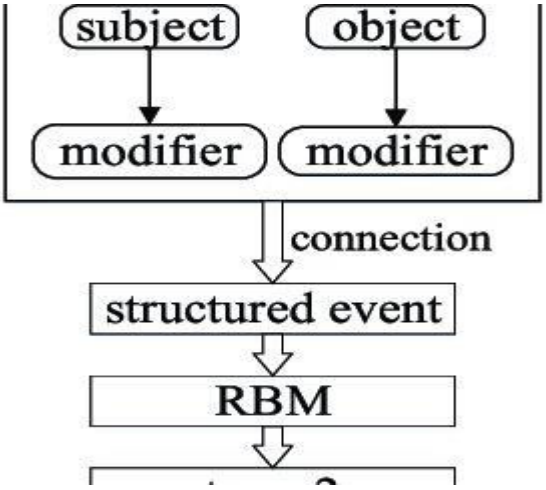

Fig 5.1: Structure of the Proposed System

VI. IMPLEMENTATION

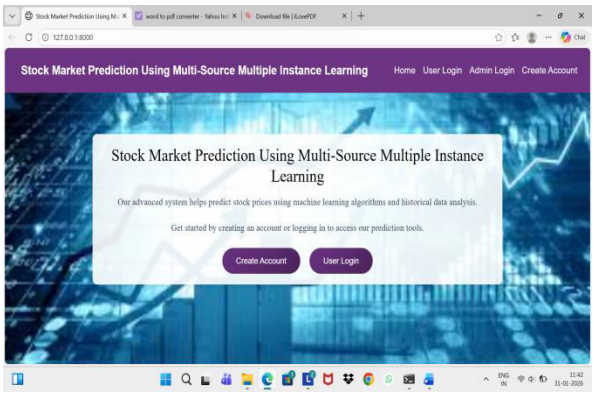

Fig 6.1: Home Page

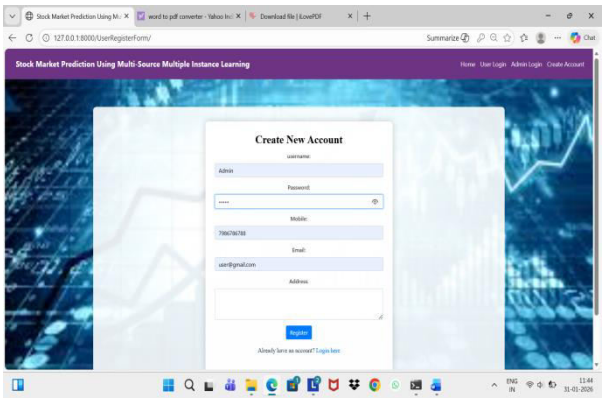

Fig 6.2: Register Page

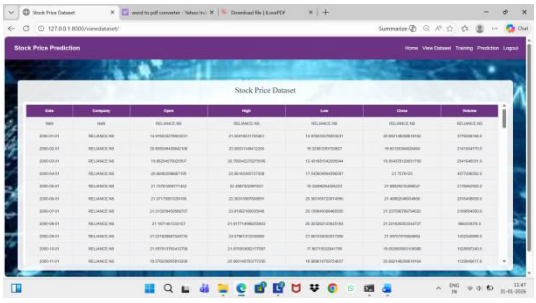

Fig 6.3: Dataset Page

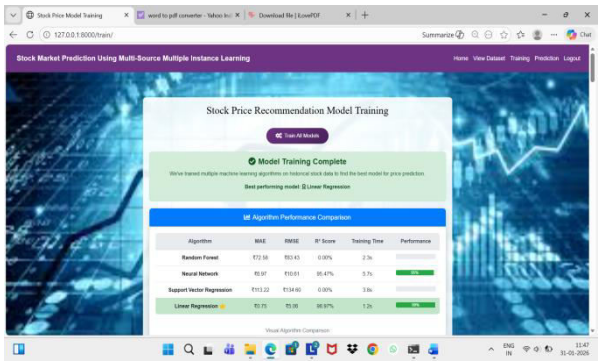

Fig 6.4: Training Page

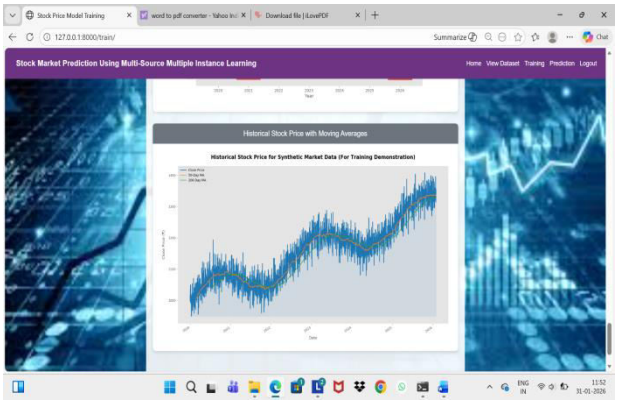

Fig 6.5: Results Page

VII. CONCLUSION

In conclusion, the proposed Multi-Source Multiple Instance Learning (MS-MIL) approach provides an effective solution for addressing the complexities and uncertainties inherent in stock market prediction. By integrating diverse data sources such as historical prices, financial news, social media sentiment, and macroeconomic indicators, the system captures a more comprehensive view of market behavior. The use of multiple instance learning enables the model to handle weakly labeled and noisy data, while attention-based mechanisms enhance its ability to identify the most relevant information across sources. As a result, the model achieves improved predictive accuracy, robustness, and interpretability compared to traditional methods. This approach not only advances the field of financial forecasting but also offers practical value for investors, analysts, and decision-makers seeking more reliable insights into market trends.

VIII. FUTURE SCOPE

The future scope of the proposed Multi-Source Multiple Instance Learning (MS-MIL) based stock

market prediction system is extensive and promising. The model can be further enhanced by incorporating additional real-time data sources such as global financial news feeds, live trading signals, and alternative data like satellite imagery or web traffic trends to improve prediction accuracy. Advanced deep learning architectures, including transformers and graph neural networks, can be integrated to better capture complex relationships and temporal dependencies in financial data. The system can also be extended to support multi-class predictions, such as identifying specific market events or price ranges, rather than simple up/down movements. Additionally, implementing real-time deployment with automated trading strategies and risk management modules can make the system more practical for real-world applications. Improving explainability through advanced visualization and interpretability techniques will further help investors understand model decisions. Finally, expanding the framework to other financial domains such as cryptocurrency markets, commodity trading, and portfolio optimization can broaden its applicability and impact.

IX. REFERENCES

[1] Dietterich, T. G., Lathrop, R. H., & Lozano-Pérez, T. (1997). Solving the Multiple Instance Problem with Axis-Parallel Rectangles. Artificial Intelligence, 89(1–2), 31–71.

[2] Ilse, M., Tomczak, J. M., & Welling, M. (2018). Attention-based Deep Multiple Instance Learning. Proceedings of the 35th International Conference on Machine Learning (ICML).

[3] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter Mood Predicts the Stock Market. Journal of Computational Science, 2(1), 1–8.

[4] Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T. Y. (2018). Listening to Chaotic Whispers: A Deep Learning Framework for News-oriented Stock Trend Prediction. Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM).

[5] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep Learning for Event-Driven Stock Prediction. Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI).

[6] Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016). Deep Learning for Stock Prediction Using Numerical and Textual Information. Proceedings of the IEEE/ACIS International Conference on Computer and Information Science (ICIS).

[7] Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

[8] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

[9] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems (NeurIPS).

[10] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

. Conf. Knowledge Discovery and Data Mining, pp.

.