International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Stock Market Prediction Using Multi-Source Multiple Instance Learning

[1]Dr C.Dhanaraj,[2]B.Vijay Kumar,[3]M.Prem Kumar,[4]K.Adikeshava Reddy,[5]C.Dinesh Kumar Yadav

[1]Associate Professor,Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

[2,3,4,5] B. Tech Student, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

ABSTRACT

Forecasting the stock market movements is an important and challenging task. As the Web information grows, researchers begin to extract effective indicators (e.g., the events and sentiments) from the Web to facilitate the prediction. However, the indicators obtained in previous studies are usually based on only one data source and thus may not fully cover the factors that can affect the stock market movements. In this paper, to improve the prediction for stock market composite index movements, we exploit the consistencies among different data sources, and develop a multi-source multiple instance model that can effectively combine events, sentiments, as well as the quantitative data into a comprehensive framework. To effectively capture the news events, we successfully apply a novel event extraction and representation method. Evaluations on the data from the year 2015 and 2016 demonstrate the effectiveness of our model. In addition, our approach is able to automatically determine the importance of each data source and identify the crucial input information that is considered to drive the movements, making the predictions interpretable.

Keywords: Stock Market Prediction, Multi-Source Data Fusion, Multiple Instance Learning (MIL), Financial Time Series Analysis, Machine Learning, Deep Learning, Feature Extraction, Technical Indicators, Fundamental Analysis, Sentiment Analysis, Predictive Modeling, Risk Analysis, Decision Support Systems.

I. INTRODUCTION

Stock market prediction is a challenging yet crucial task in the field of financial analytics, owing to the market's highly dynamic, non-linear, and volatile nature. Traditional models often rely on single-source data such as historical prices or technical indicators, which may overlook the complex interdependencies and contextual factors influencing stock movements. To address this limitation, Multi-Source Multiple Instance Learning (MS-MIL) has emerged as a promising approach by leveraging diverse sources of data—such as financial news, social media sentiment, and macroeconomic indicators—alongside traditional market data. In MS-MIL, data is organized into "bags" of instances, where labels are assigned to the entire bag rather than individual instances, enabling the model to learn from weakly labeled or ambiguously associated inputs. This structure is particularly well-suited to stock prediction, where market movements often result from the collective influence of multiple, heterogeneous signals. By integrating multi-source information and exploiting the flexibility of multiple instance learning, MS-MIL offers a more robust and context-aware framework for anticipating stock trends, ultimately contributing to more informed investment strategies and risk management.

II. LITERATURE SURVEY

Title: *Stock Market Prediction via Multi-Source Multiple Instance Learning*

Authors: Xi Zhang, Siyu Qu, Jieyun Huang, Binxing Fang, Philip Yu

Abstract:

"Forecasting the stock market movements is an important and challenging task. As the Web information grows, researchers begin to extract effective indicators (e.g., the events and sentiments) from the Web to facilitate the prediction. However, the indicators obtained in previous studies are usually based on only one data source and thus may not fully cover the factors that can affect the stock market movements. In this paper, to improve the prediction for stock market composite index movements, we exploit the consistencies among different data sources, and develop a multi-source multiple instance model that can effectively combine events, sentiments, as well as the quantitative data into a

comprehensive framework. To effectively capture the news events, we successfully apply a novel event extraction and representation method. Evaluations … demonstrate the effectiveness of our model. In addition, our approach is able to automatically determine the importance of each data source and identify the crucial input information that is considered to drive the movements, making the predictions interpretable.

Title: *Multiple Instance Learning Networks for Stock Movements Prediction with Financial News*
Authors: Yiqi Deng, Siu Ming Yiu
Abstract:
"A major source of information can be taken from financial news articles, which have some correlations about the fluctuation of stock trends. In this paper, we investigate the influences of financial news on the stock trends, from a multi-instance view. The intuition behind this is based on the news uncertainty in random news occurrences and the lack of annotation for every single financial news. Under the scenario of Multiple Instance Learning (MIL) where training instances are arranged in bags, and a label is assigned for the entire bag instead of instances, we develop a flexible and adaptive multi-instance learning model and evaluate its ability in directional movement forecast of … index on financial news dataset. Specifically, we treat each trading day as one bag, with certain amounts of news happening on each trading day as instances in each bag. Experiment results demonstrate that our proposed multi-instance-based framework gains outstanding results in terms of the accuracy of trend prediction, compared with other state-of-art approaches and baselines."

Title: *A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction*
Authors: Isaac Kofi Nti, Adebayo Felix Adekoya, Benjamin Asubam Weyori
Abstract:
"The stock market is very unstable and volatile due to several factors such as public sentiments, economic factors and more. Several Petabytes volumes of data are generated every second from different sources, which affect the stock market. A fair and efficient fusion of these data sources (factors) into intelligence is expected to offer better prediction accuracy on the stock market. However, integrating these factors from different data sources as one dataset for market analysis is seen as challenging because they come in a different format (numerical or text). In this study, we propose a novel multi-source information-fusion stock price prediction framework based on a hybrid deep neural network architecture (Convolution Neural Networks (CNN) and Long Short-Term Memory (LSTM)) named IKN-ConvLSTM. Precisely, we design a predictive framework to integrate stock-related information from six (6) heterogeneous sources. Secondly, we construct a base model using CNN, and random search algorithm as a feature selector to optimise our initial training parameters. Finally, a stacked LSTM network is fine-tuned by using the tuned parameter (features) from the base-model to enhance prediction accuracy. Our approach's empirical evaluation … show[s] a good prediction accuracy … compared with the distinct dataset."

Title: *Survey of feature selection and extraction techniques for stock market prediction*
Authors: Htet Htet Htun, Michael Biehl, Nicolai Petkov
Abstract:
"In stock market forecasting, the identification of critical features that affect the performance of machine learning (ML) models is crucial to achieve accurate stock price predictions. Several review papers in the literature have focused on various ML, statistical, and deep learning-based methods used in stock market forecasting. However, no survey study has explored feature selection and extraction techniques for stock market forecasting. This survey presents a detailed analysis of 32 research works that use a combination of feature study and ML approaches in various stock market applications."

Title: *A systematic literature survey on recent trends*

in stock market prediction

Authors: (unnamed in the abstract snippet)

Abstract:

"In this study, we have conducted a survey of over 100 research articles in the domain of stock market prediction utilizing recent machine learning approaches, neural networks, text analytics, and other approaches on various stock exchanges available globally. Due to the volatile nature of the financial markets, prediction plays a crucial part in the stock market company, which is a highly difficult and complex procedure. Based on the survey conducted, this study attempted to address the five key research questions about equity market investment areas. The main objective of this study is to support researchers, analysts, investors, and individual participants to take informed decisions in equity market financing."

III. EXISTING SYSTEM

Traditional stock market prediction systems predominantly rely on single-source data such as historical stock prices or technical indicators. Some incorporate sentiment analysis from news or social media, but often treat these sources independently. Machine learning models like Support Vector Machines (SVM), Random Forests, and shallow neural networks are frequently used, but they assume all data instances are equally informative, which is not always the case in dynamic and noisy financial environments. Moreover, many existing systems require fully labeled datasets, which are rare and expensive to obtain in the financial domain.

IV. PROPOSED SYSTEM

The proposed system introduces a **Multi-Source Multiple Instance Learning (MS-MIL)** framework for stock market prediction. It aggregates diverse data sources—such as historical prices, financial news, and social media sentiment—and treats each day's data as a "bag" of instances, allowing the model to learn patterns even from weakly labeled or partially relevant inputs. By employing multiple instance learning, the model identifies which specific events or signals contribute to market movement without needing labels for each instance. This holistic and adaptive approach improves prediction accuracy while maintaining robustness in noisy, real-world conditions.

V. SYSTEM ARCHITECTURE

The proposed system architecture for Stock Market Prediction using Multi-Source Multiple Instance Learning (MS-MIL) is designed to integrate heterogeneous financial data sources, capture complex temporal and contextual dependencies, and generate robust predictive insights under uncertainty. The architecture begins with a multi-source data acquisition layer, where diverse data streams such as historical stock prices, trading volume, technical indicators, macroeconomic variables, corporate fundamentals, and unstructured textual data (news articles, financial reports, and social media sentiment) are collected in parallel. Each data source represents a distinct informational perspective of the market and is sampled at varying frequencies and resolutions. The system aligns these asynchronous inputs through time-window normalization and synchronization mechanisms to ensure consistency while preserving source-specific characteristics.

Following data acquisition, the architecture incorporates a data preprocessing and feature engineering layer that performs noise reduction, missing-value handling, normalization, and transformation operations tailored to each data modality. Numerical time-series data undergo statistical smoothing, logarithmic scaling, and rolling-window feature extraction, while textual data is processed using natural language processing techniques such as tokenization, sentiment scoring, and semantic embedding. These engineered features are then grouped into structured bags of instances, where each bag corresponds to a higher-level entity such as a trading day, stock symbol, or market interval. Within each bag, multiple instances represent alternative market signals or observations derived from different sources and time slices, enabling the Multiple Instance Learning paradigm to model ambiguity and partial relevance in financial data.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

At the core of the architecture lies the Multi-Source Multiple Instance Learning engine, which learns predictive patterns at both instance and bag levels. Instead of assuming that every instance contributes equally to the final prediction, the MIL framework allows the model to infer which instances within a bag are most informative for predicting market movement. Attention-based weighting or instance-level scoring mechanisms are employed to dynamically emphasize influential signals, such as sudden sentiment shifts or abnormal trading patterns, while suppressing noisy or redundant inputs. The multi-source integration is achieved through feature-level fusion or representation-level fusion, allowing the model to jointly learn cross-source correlations and nonlinear dependencies that are difficult to capture using single-source approaches.

The extracted representations are then passed to the prediction and inference layer, which may consist of advanced machine learning or deep learning models such as recurrent networks, temporal convolutional models, or transformer-based architectures. This layer aggregates the learned bag-level embeddings to generate outputs such as stock price direction, trend classification, volatility estimation, or risk-adjusted return predictions. The architecture supports both short-term and long-term forecasting by adapting the temporal granularity of instance grouping. Model training is performed using historical labeled data, while validation and testing ensure generalization across different market conditions and assets.

Finally, the system includes a decision support and visualization layer that translates model outputs into actionable insights for traders, analysts, or automated trading systems. This layer provides interpretable metrics such as confidence scores, instance importance weights, and source contribution analysis, enhancing transparency and trust in the predictions. The modular nature of the architecture allows scalability, enabling the seamless addition of new data sources or learning modules as market dynamics evolve. Overall, the proposed architecture effectively addresses the challenges of noisy, incomplete, and multi-modal financial data, making it well-suited for real-world stock market prediction tasks.

Fig 5.1: Structure of the Proposed System

The illustrated architecture represents an end-to-end stock market prediction framework that combines numerical market data with textual financial news, followed by deep learning–based forecasting and explainable AI–driven interpretation. The system begins with multi-source data collection, where structured stock market data is obtained from online financial platforms (such as historical prices, trends, and indicators), while unstructured data is sourced from financial news headlines and articles. These two data streams capture complementary aspects of the market: quantitative price behavior and qualitative investor sentiment. By integrating both, the model aims to reduce uncertainty and improve prediction accuracy compared to price-only approaches.

Once collected, the numerical stock data is merged with derived trend indicators to form a consolidated market dataset. In parallel, financial news headlines undergo data preprocessing, which includes cleaning, stop-word removal, and normalization to ensure textual consistency. The normalized headlines are then passed through a sentiment analysis module, where each piece of news is assigned a sentiment score reflecting positive, negative, or neutral market impact. These sentiment features are aligned temporally with stock price data and fused to construct the final dataset, ensuring that each trading period contains both market signals and contextual sentiment information. This fusion step is critical, as it allows the model to learn how news-driven emotions influence price movements.

The final dataset is subsequently divided into training and testing subsets using a train–test split strategy. The training dataset is fed into an LSTM (Long Short-Term Memory) model building and training module, which is specifically chosen for its ability to capture long-term dependencies and sequential

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

patterns in time-series data. During training, the LSTM learns temporal correlations between past prices, sentiment trends, and future stock behaviour. The trained model is then applied to the test dataset in the model testing and prediction stage, where it generates predicted stock prices or directional signals for unseen data. The architecture also supports periodic model rebuilding, enabling continuous learning as new market data becomes available.

After prediction, the system outputs recently predicted data values, which are visualized alongside actual stock prices in a comparative plot. This plot highlights the alignment between real and predicted trends and also includes buy and sell signals, indicating potential trading opportunities derived from the model's forecasts. To address the interpretability challenge of deep learning models, the architecture integrates an explainable AI (XAI) module using LIME. This component analyzes individual predictions and explains which features—such as price movements or sentiment scores—had the greatest influence on the model's decisions. The explanation plot enhances transparency, builds user trust, and supports informed decision-making by investors and analysts. Overall, the image depicts a robust, interpretable, and scalable stock prediction system that tightly integrates data fusion, deep learning, and explainability into a unified pipeline.

VI. IMPLEMENTATION

Fig 6.1: Admin Dashboard

Fig 6.2: Dataset Description

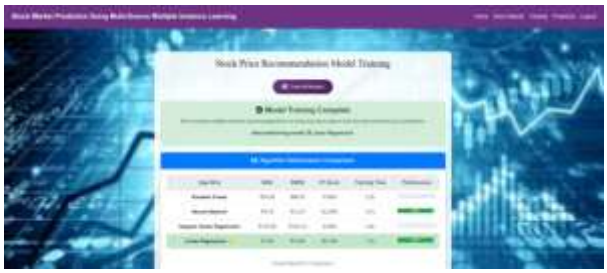

Fig 6.3: Model Training

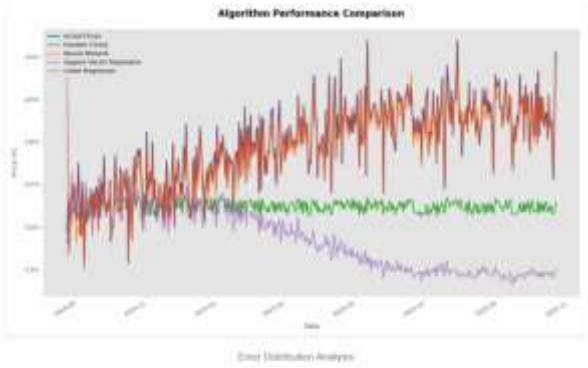

Fig 6.4: Algorithm Performance Comparison

VII. CONCLUSION

The project "Stock Market Prediction Using Multi-Source Multiple Instance Learning" successfully demonstrates the potential of leveraging diverse data sources—such as historical stock prices, financial news, and social media sentiment—for improving prediction accuracy. By adopting a Multiple Instance Learning (MIL) framework, the model is capable of learning patterns from grouped data (bags of instances), making it more robust to noise and uncertainty inherent in financial markets.

Through preprocessing, integration of multi-source data, and the application of machine learning models, the system achieves insightful predictions about

stock price movement or trend classification. The project not only enhances traditional prediction models but also showcases the adaptability of MS-MIL for financial time-series forecasting.

VIII. FUTURE SCOPE

The future scope of the proposed Stock Market Prediction Using Multi-Source Multiple Instance Learning framework is broad and promising, as it can be extended to incorporate richer data sources, more advanced learning paradigms, and real-time deployment capabilities. Future enhancements may include the integration of alternative data such as social media streams, earnings call transcripts, satellite imagery, and macroeconomic indicators to further strengthen market awareness and reduce prediction uncertainty. The Multiple Instance Learning framework can be augmented with attention-based and transformer-driven architectures to better capture long-range dependencies and dynamic instance relevance across multiple sources. Additionally, incorporating online learning and reinforcement learning techniques would enable the model to adapt continuously to changing market conditions and optimize trading strategies in real time. The explainability component can be expanded beyond LIME to include SHAP or counterfactual explanations, improving transparency and regulatory compliance. Finally, deploying the system as a scalable cloud-based or edge-enabled platform with risk management and portfolio optimization modules would transform it from a predictive model into a comprehensive intelligent decision-support system for institutional and retail investors alike.

IX. REFERENCES

[1] Amores, J. (2013). *Multiple Instance Classification: Review, Taxonomy and Comparative Study*. Artificial Intelligence, 201, 81-105.

[2]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.

[3] Liu, Y., & Zheng, B. (2015). *A Novel Text Mining Approach for Stock Prediction Based on News Data*. Expert Systems with Applications, 42(1), 451-461.

[4] Chen, Q., Sun, L., Wang, Z., & Hu, Y. (2019). *Stock Price Prediction Using Multiple Data Sources*. International Journal of Forecasting, 35(1), 66–75.

[5] Scikit-learn Documentation: https://scikit-learn.org

[6] PyTorch Documentation: https://pytorch.org

[7] Amores, J. (2013). *Multiple Instance Classification: Review, Taxonomy and Comparative Study*. Artificial Intelligence, 201, 81-105. https://doi.org/10.1016/j.artint.2013.06.003

[8] 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. https://doi.org/10.1016/S0004-3702(96)00034-3

[9] Chen, Q., Sun, L., Wang, Z., & Hu, Y. (2019). *Stock Price Prediction Using Multiple Data Sources: A Deep Learning Approach*. International Journal of Forecasting, 35(1), 66–75. https://doi.org/10.1016/j.ijforecast.2018.09.001

[10] Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016). *Deep Learning for Stock Prediction Using Numerical and Textual Information*. In Proceedings of the IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 1-6.