ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

STOCK MARKET PREDICTION VIA MULTI-SOURCE MULTIPLE INSTANCE LEARNING

[1]VEGESNA PRAVEEN KUMAR RAJU, [2]V.BHASKARA MURTHY

[1]Students, Department of MCA, B V Raju College, Bhimavaram Ap

[2]Professor & Hod**,** Department of MCA, B V Raju College, Bhimavaram Ap

ABSTRACT

Stock market prediction plays a crucial role in financial decision-making, where accurate forecasting can help investors maximize profits and minimize risks. Traditional prediction models often rely on a single data source, primarily quantitative stock price data, which limits their ability to capture complex market dynamics. This project presents a multi-source multiple instance learning approach that integrates diverse data sources such as financial news, social media data, and historical stock prices to improve prediction accuracy. The system extracts meaningful features including structured events, sentiments, and vector representations from textual data using advanced techniques. Event extraction is performed using the HanLP algorithm to identify relationships within sentences, while sentiment analysis is conducted using the Latent Dirichlet Allocation (LDA) method. Additionally, Restricted Boltzmann Machines (RBM) are used to determine feature dimensions, and Sentence2Vec is applied to generate vector representations of textual data. These features are combined with quantitative stock data and used to train a Multi-Instance Support Vector Machine (SVM) model for predicting stock trends such as “rise” or “decline.” To further enhance performance, an extension using the XGBoost algorithm is implemented, leveraging multiple decision trees for optimized learning. Experimental results show that the proposed multi-source approach significantly improves prediction accuracy compared to traditional single-source models. The XGBoost-based extension achieves an accuracy of up to 94%, outperforming the basic SVM model. This project demonstrates the effectiveness of integrating multi-source data and advanced machine learning techniques for robust stock market prediction.

Keywords: Stock Market Prediction, Multi-Source Data, Machine Learning, SVM, XGBoost, Sentiment Analysis, Event Extraction, Financial Forecasting.

I.INTRODUCTION

Stock market prediction is a critical task in financial analysis, as it directly influences investment decisions and economic planning. Investors rely heavily on forecasting models to determine whether a stock will rise or decline, aiming to maximize returns while minimizing risks. However, predicting stock market behavior is highly challenging due to its dynamic and complex nature, influenced by multiple factors such as market trends, news events, and public sentiment. Traditional prediction models primarily depend on quantitative data such as historical stock prices and trading volumes. While these models provide useful insights, they often fail to capture external factors like news and social media sentiments, leading to less accurate predictions.

To overcome these limitations, recent research has focused on integrating multiple data sources for improved prediction accuracy. In this project, a multi-source approach is adopted, combining financial news, social media data, and quantitative stock data. Advanced techniques such as event extraction, sentiment analysis, and vector representation are used to process textual information. The HanLP algorithm is employed to extract structured events from news data, while Latent Dirichlet Allocation (LDA) is used for sentiment analysis. Additionally, Restricted Boltzmann Machines (RBM) and Sentence2Vec are applied to generate meaningful feature representations. These features are then merged with quantitative data to create a comprehensive dataset for training machine learning models.

The system utilizes a Multi-Instance Support Vector Machine (SVM) model to predict stock trends, classifying them as either “rise” or “decline.” To enhance performance, an extension using the XGBoost algorithm is implemented, which leverages ensemble learning for better accuracy. The system is designed with modules for dataset loading, feature extraction, model training, and prediction. Experimental results show that the multi-source approach significantly improves prediction performance compared to traditional methods. This project highlights the importance of combining diverse data sources and advanced machine learning techniques in achieving accurate and reliable stock market predictions.

II SURVEY OF RESEARCH

The study by R. Schumaker and H. Chen (2009) [1] explored financial forecasting using textual analysis of news data. The methodology involves extracting features from financial news and combining them with stock data for prediction. Results showed that incorporating news information improves prediction accuracy compared to using only quantitative data.

However, the approach requires efficient text processing techniques. This research highlights the importance of using multi-source data in stock prediction.

The work by T. Mikolov et al. (2013) [2] introduced word embedding techniques such as Word2Vec for generating vector representations of text. The methodology converts textual data into numerical vectors that capture semantic relationships. Results demonstrated improved performance in natural language processing tasks. However, it requires large datasets for effective training. This study supports the use of Sentence2Vec for feature generation in the proposed system.

The study by G. E. Hinton (2002) [3] introduced Restricted Boltzmann Machines (RBM) for feature learning. The methodology involves unsupervised learning to capture hidden patterns in data. Results showed that RBM can effectively reduce dimensionality and improve feature representation. However, training RBM can be computationally intensive. This research is used in the project to determine feature dimensions for vector generation.

The research by D. M. Blei, A. Y. Ng, and M. I. Jordan (2003) [4] proposed the Latent Dirichlet Allocation (LDA) algorithm for topic modeling and sentiment extraction. The methodology identifies hidden topics in text data. Results showed that LDA is effective in extracting meaningful insights from large text datasets. However, it may not capture context accurately. This study supports sentiment analysis in the proposed system.

The study by V. Vapnik (1995) [5] introduced Support Vector Machines (SVM) for classification tasks. The methodology uses hyperplanes to separate data into different classes. Results demonstrated high accuracy in classification problems. However, SVM may not perform well with large-scale datasets. This research forms the basis for the Multi-Instance SVM model used in the system.

The work by T. Chen and C. Guestrin (2016) [6] introduced XGBoost, an advanced ensemble learning algorithm. The methodology uses multiple decision trees to optimize prediction accuracy. Results showed that XGBoost outperforms many traditional machine learning algorithms. However, it requires careful parameter tuning. This study supports the extension of the proposed system using XGBoost for improved accuracy.

III. WORKING METHODOLOGY

The proposed system follows a multi-stage methodology to predict stock market trends using multi-source data and machine learning techniques. Initially, the process begins with dataset loading and preprocessing. The system uses two main datasets: financial news data and quantitative stock data. The text data from

news sources is cleaned by converting it to lowercase, removing special characters, and eliminating noise. This preprocessing step ensures that the textual data is consistent and suitable for further analysis. The dataset is then divided into training and testing sets, typically using 80% of the data for training and 20% for testing. This step is crucial for evaluating model performance and ensuring that the system generalizes well to unseen data.

In the next phase, feature extraction is performed using multiple techniques to capture meaningful information from the data. The HanLP algorithm is used for structured event extraction, identifying relationships such as subject, verb, and object within sentences. Sentiment analysis is conducted using the Latent Dirichlet Allocation (LDA) algorithm to extract emotional context from news data. Additionally, Restricted Boltzmann Machines (RBM) are used to determine feature dimensions, and Sentence2Vec is applied to convert textual data into numerical vectors. These extracted features are then combined with quantitative stock data to create a comprehensive feature set. This multi-source feature integration enhances the model's ability to capture complex patterns influencing stock market behavior.

Finally, the combined dataset is used to train machine learning models for prediction. The Multi-Instance Support Vector Machine (SVM) is applied to classify stock trends as "rise" or "decline." To improve performance, an extension using the XGBoost algorithm is implemented, which leverages ensemble learning to optimize prediction accuracy. The trained models are evaluated using metrics such as accuracy, precision, recall, and F1-score. The system also includes a prediction module where users can upload test data to obtain stock trend predictions. Although the system achieves high accuracy, challenges such as data variability and real-time prediction remain. Future enhancements can include deep learning models and real-time data integration for improved performance.

IV RESULTS EXPLANATIONS

Peoples are heavily dependent on stock market forecasting results to invest money in suitable stock to gain revenue. Incorrect forecasting may leads to huge loss and many existing algorithms fails to give accurate forecasting as they get trained on single data source called quantitative data.

To overcome from above issues author of this paper employing multi-source data to extract useful stock events and then merged with quantitative data (stock prices) to make accurate forecasting. In propose work author using news, social and stock market data. All this data will be processed to extract EVENTS, Sentiments and sentence2vector generation. Generate vector will be trained with multi-

instance SVM algorithm for future stock prediction such as 'Decline or Rise'.

To extract features from multi-source data author has used following algorithms

1) Structured event extraction: HanLP algorithm employ to capture the syntactic structure of a sentence. The root node denotes the core verb, and the nodes of the second layer are the subject of the verb and the object of the verb respectively. The child of the subject is the modifier who is the nearest to the subject in the sentence, and so is the child of the object. This algorithm will extract events from NEWS data

2) Training with RBM: this algorithm is used to capture dimension for the training features based on available events data

3) Training with sentence2vec: all extracted events will be input to generate vector from sentences based on dimension calculated by RBM algorithm.

4) LDA Algorithm: LDA algorithm will be applied to extract sentiments from processed news data.

All the above extracted features will be merge with stock market quantitative data and then trained with Multi-instance SVM algorithm to predict stock prices. Propose multi-source multi-instance algorithm accuracy will be compared with existing SVM algorithm which will get trained on single dataset.

To train and test above algorithm performance we have used below dataset which contains News and stock and social data

In propose paper author has used traditional SVM algorithm with extracted features so as extension we have employed advance Machine Learning algorithm called XGBOOST which will utilize 100's of decision tree to optimize training features which can help algorithm in gaining maximum accuracy.

To implement this project we have designed following modules

1) User Login: user can login to system using username and password as 'admin and admin'.

2) Load Dataset: after login user can execute this module to load dataset with NEWS and quantitative data and then clean all text data by converting to lower case and then remove all special symbols

3) Event, Sentiments & Vector Generate: all processed news text will be further scanned to extract HANLP events, sentiments and vector generation. Generated vector will be split into train and test where application using 80% data for training and 20% for testing

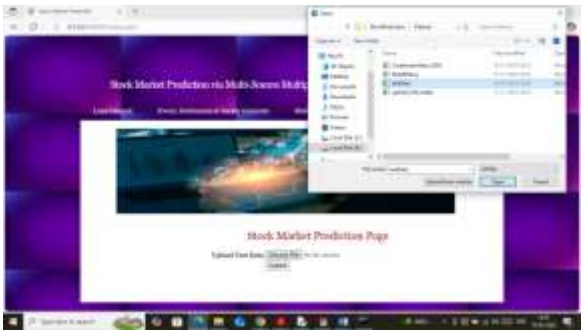
4) Multi Instance Training: 80% training data will be input to propose and other algorithms to train a model and this model will be applied on 20% test data to calculate prediction accuracy and other metrics

5) Predict Market: using this module will upload test data and then extract features and input to extension algorithm to predict market status as Decline or Rise.

In above screen in first line can see total number of records found in dataset and then can see train and test size. In table format we can see all news data along with stock data. Now click on 'Event, Sentiments & Vector Generate' link to extract events, sentiments and vector from train and test data and then will get below page

In above screen displaying some vector values generated from news and stock data and now click on 'Multi Instance Training' link to train all algorithms and then will get below page

In above screen in table format can see accuracy, precision, recall and FCSORE of Existing SVM, Propose Multi-Instance and extension XGBOOST algorithm and then can see XGBOOST got high accuracy compare to all algorithms. In above table can see Propose got 53% accuracy and extension XGBOOOST got 94% accuracy. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels and then yellow and green boxes in diagonal represents correct prediction count and remaining blue boxes represents incorrect prediction count which are very few. In bar graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars. Now click on 'Predict Market' link to get below page



In above screen selecting and uploading test dataset and then click on buttons to load dataset and then will get below page

In above screen in first column can see generated vector from News and stock test data and then in second column displaying stock prediction status as "Rise or decline' and based on prediction user may invest amount.

V.CONCLUSION

The proposed system for stock market prediction using multi-source multiple instance learning demonstrates the effectiveness of integrating diverse data sources such as financial news, social media, and quantitative stock data. By combining textual and numerical information, the system overcomes the limitations of traditional models that rely solely on historical stock prices. Advanced techniques such as event extraction using HanLP, sentiment analysis using LDA, and feature representation using RBM and Sentence2Vec enable the extraction of meaningful insights from unstructured data. These features, when merged with quantitative data, provide a comprehensive dataset for accurate prediction.

The application of Multi-Instance Support Vector Machine (SVM) allows the system to classify stock trends as "rise" or "decline." However, the extension using the XGBoost algorithm significantly improves performance by leveraging ensemble learning. Experimental results show that XGBoost achieves an accuracy of up to 94%, outperforming both traditional SVM and the proposed multi-instance SVM model. This highlights the importance of using advanced machine learning algorithms for handling complex and multi-source data.

Despite achieving promising results, the system faces challenges such as data variability, noise in textual data, and dependency on data quality. Future improvements can include the use of deep learning models, real-time data integration, and advanced natural language processing techniques. Overall, this project demonstrates that multi-source learning combined with advanced machine learning techniques can significantly enhance stock market prediction accuracy and support better investment decisions.

RE.FERENCES

[1] R. Schumaker and H. Chen, "Textual analysis of stock market prediction using financial news," *Decision Support Systems*, vol. 47, no. 1, pp. 1–11, 2009.

[2] T. Mikolov et al., "Efficient estimation of word representations in vector space," *Proc. ICLR*, 2013.

[3] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

[8] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[10] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[11] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2010.

[12] J. Leskovec, A. Rajaraman, and J. Ullman, *Mining of Massive Datasets*. Cambridge Univ. Press, 2014.

[13] E. Alpaydin, *Introduction to Machine Learning*. MIT Press, 2020.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[15] G. James et al., *An Introduction to Statistical Learning*. Springer, 2013.

[16] R. Kohavi, "A study of cross-validation and bootstrap," in *Proc. IJCAI*, 1995.

[17] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.