Grenze International Journal of Engineering and Technology, Jan Issue

Event Driven Stock Prediction

Prateek N Kamath[1], Moksha Pradhan[2], Jayanth K[3] and Dr. Jyothi R[4]
[1-4]PES University/Computer Science, Bangalore, India
Email: prateek.kamath02@gmail.com, {moksha.pradhan, jayanthk714}@gmail.com, {jyothir@pes.edu}

Abstract— **The method for event-driven stock prediction presented in this work combines machine learning, sentiment analysis, and natural language processing techniques. The main objective is to create an advanced technique that might accurately predict stock market movements using these approaches. By looking at news stories and financial data and giving each a weight that is particularly significant, the suggested methodology primarily captures sentiment and relevancy. To depict the rather intricate correlations between events and stock price changes, a predictive model, like a recurrent neural network, is generally trained using weighted data. The results greatly outperform conventional time-series forecasting methods in a setting where events primarily affect the stock market. These insights are beneficial to traders, investors, and financial institutions.**

Index Terms— **Machine learning, Sentiment analysis, Natural language processing, Time-series forecasting**

I. INTRODUCTION

Event-driven stock prediction is a method that accurately anticipates stock price fluctuations by leveraging the impact of significant events and news releases. Traditional time-series forecasting methods face challenges in today's financial markets, necessitating innovative approaches that incorporate news and events into stock prediction models. Machine learning techniques like recurrent neural networks (RNNs) help study non-linear relationships and model complex patterns within data. This study aims to demonstrate the growing need for creative methods that incorporate news and events in predicting stocks, potentially altering how stock market forecasts are generated. The applications of event-driven stock prediction include increased precision and dependability of stock price forecasts, better risk management, investment opportunities, and optimization of trading strategies based on the anticipated impact of specific events. It also offers insights into individual stock performance, well-diversified portfolios, asset allocation optimization, and potential risks associated with specific events. Additionally, event-driven models can aid in market surveillance and regulation, detecting unusual trading patterns and potentially mitigating market manipulation.

II. METHODOLOGY

A. Event Correlation and Feature Engineering

1. Time Series Analysis: Time series analysis is a statistical technique used in various domains, including finance, economics, sales forecasting, and weather forecasting. It involves visualizing data to uncover patterns and trends, using smoothing techniques to remove noise, and decomposing time series into its constituent parts like trend,

Grenze ID: 01.GIJET.10.1.147

seasonality, and residual. Forecasting models like GARCH, ARIMA, and exponential smoothing use historical patterns and relationships to generate predictions for future values. To ensure reliability and effectiveness, time series models are evaluated using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

2. Sentiment Analysis: Sentiment analysis is a technique utilized in an event-driven stock prediction that aims to examine and comprehend the sentiment or emotional tone present in textual information related to events or news affecting stock markets significantly. To gauge the sentiment surrounding certain events and their potential impact on stock prices, subjective data must be obtained from a varied range of sources, including news articles, social media posts, financial reports and other relevant sources.

B. Forecasting Models

On the basis of past data, forecasting models use statistical or machine-learning approaches to forecast future values or trends. Time series models, econometric models, judgemental forecasting models, and Delphi approaches are the four basic categories into which forecasting models can be divided. Time series forecasting has been mostly used in this study. These models are excellent for jobs like sales forecasting, weather forecasting, and anomaly detection because they can capture complicated patterns and non-linear correlations in data. Additionally, Vector Autoregression (VAR) models are useful for forecasting multiple interrelated time series variables, commonly employed in macroeconomic forecasting and financial analysis.

1) Gradient Boosting: This machine learning method handles the classification and regression tasks. Gradient boosting basic principle is to repeatedly train new models that can fix the mistakes made by the older models. The residual errors of the ensemble are trained into each new model, and the predictions of all models are then integrated to get the final forecast.

2) Recurrent Neural Networks (RNN): are a subclass of neural networks that use the results of one phase to guide the next. They differ from modern neural networks, which have independent inputs and outputs. RNNs use a Hidden Layer to address this issue, with its Hidden state holding specialized sequence-related data. The network uses the same parameters for each input, making the parameter set simpler than for other neural networks.

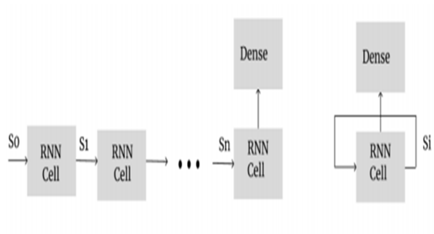

Fig. 1. Recurrent Neural Network

C. Measurement and Evaluation

1) Mean Absolute Error (MAE): The statistical notion of mean absolute error (MAE) is used to assess the correctness of matched data that reflect the same occurrence.

$$MAE = 1 / n \, (X \, |Y_i - Y\hat{}_i|) \qquad (1)$$

where:

The Mean Absolute Error, or MAE, The total number of data points or observations is n. Yi is the target variable's actual value, whereas i is its anticipated value.

Precision(P) among all the positive predictions the model makes, is the percentage of correct positive predictions. This demonstrates the potential accuracy of optimistic predictions.

Recall(R) often referred to as sensitivity, is the proportion of precise positive predictions made from all actual positive events in the dataset. It seeks to compile as many uplifting instances as it can.

F1 score(F1) is the harmonic average of precision and recall. A single number is generated that takes precision and recalls into account when false positives and false negatives occur. This graph is helpful for assessing how well the model is working.

$$F1 = 2 * (P * R)/(P + R) \qquad (2)$$

D. Techniques planned to be used

The project "Event-driven Stock Price Prediction" uses a number of machine learning (ML) and data analysis techniques. In our project, our motive is to use the following methods:

1) Techniques for Machine Learning (ML)

a) Regression Analysis: Regression analysis is a statistical method used to examine the relationships between independent variables and a dependent variable.

$$Y = a + bX + E \tag{3}$$

b) Deep Learning: This is a branch of machine learning that uses artificial neural networks. It can find complex relationships and patterns in data. Explicit programming is not required for deep learning. It is based on artificial neural networks (ANNs), also referred to as deep neural networks (DNNs), and has recently gained prominence as a result of improvements in processing power and the accessibility of big data sets.

2. Techniques for evaluation and validation

Cross-Validation It is a machine-learning approach used to evaluate a model's performance on untested data. The remaining folds are utilized to train the model while one of the given folds or subsets is used as a validation set. Every time this method is applied, which is frequently, the validation set is modified. To obtain a more accurate estimate of the model's performance, we merge the results of each validation phase.
Performance Metrics We use metrics to evaluate the performance of the machine learning based on the input data. The model's performance can be improved by modifying the input dataset's attributes or the hyper-parameters.
Accuracy(A) discusses the portion of predictions made by a categorization model that came true. Binary classification occurs when a model divides the input data into two categories. It is describe as:

$$A = (T.P + T.N)/(N) \tag{4}$$

where T.P denotes True Positive **T.N** denotes True Negative, **N** denotes the Total number of data samples
Precision(P) provides information on the proportion of predicted data items (predictions) that actually materialized as true or accurate:

$$P = (T.P) / (T.P + F.P) \tag{5}$$

where T.P denotes True Positive **F.P** denotes False Positive
Recall(R) provides information on the number of data points in a dataset that were truly relevant:

$$Recall = (T.P) / (T.P + F.N) \tag{6}$$

where T.P denotes True Positive F.N denotes False Negative
F1 Score(F1) It is obtained as the harmonic average of precision and recall. A single number is generated that takes precision and recalls into account when false positives and false negatives occur. This graph is helpful for assessing how well the model is working.

$$F1 = 2 * (P * R)(P + R) \tag{7}$$

III. LITERATURE SURVEY

[5] The study uses a deep convolution neural network and a unique neural tensor network to detect event embedding and forecast stock prices. Results show that embedding-based document representations outperform discrete event-based approaches and deep convolution neural networks outperform basic feed-forward neural networks in capturing longer-term news event consequences.[1] Future stock trend predictions are primarily based on market data, specifically news and fundamental research. The purpose of this study is to offer a tensor representation technique to represent or at least get as close to a multi-modal market.[4] The study developed a two-layer LSTM-RNN to detect latent structures in news events, enabling the forecasting of market volatility. The skip-thought model was used to represent semantic meanings. The model's performance and robustness were evaluated, and its efficacy was confirmed through experimental results. [3] This study examines the relationship between excess returns, relative returns, and stock price movements in nonferrous metals, computers, real estate, and healthcare industries. It reveals that stock prices influence current events, and the L1 logistic regression model is the best for short-term stock forecasting. Future research will explore novel events and combine components for better estimates of excess returns and long-term stock price movements. [2] A stock price trend/pattern prediction

model was developed using news events of multiple domains, pre-defined attributes, SVM models, and neural networks to understand the association between stock price movements and specific news or events.

IV. Experiment

A. BI-LSTM: To create bidirectional recurrent neural networks, two distinct RNNs are simply joined. Because of the way this structure is constructed, the networks attempt to access forward and backward information about the sequence at each step.The bidirectional LSTM process inputs in two directions from the present to the future and from the future to the present. By combining the two secret states, you may constantly store information about the past, present, and future. Unlike some other unidirectional options, this approach saves future information in the LSTM that flows backwards.

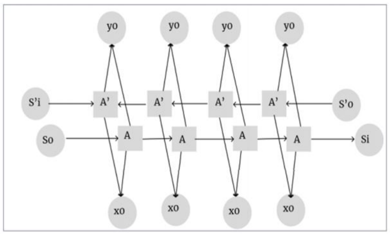

Fig. 2. Bi-directional Long Short Term Memory

Formulas:

Input Gate (i): $i(t) = (W\,i * [h(t-1), x(t)] + bi)$ (8)

Wi: weight matrix associated with the input gate, [h(t-1), x(t)]: Concatenates the current input x(t) and the previously hidden state h(t-1) bi: bias term

Forget Gate (f): $f(t) = (W\,f * [h(t-1), x(t)] + bf)$ (9)

The forget gate chooses how much of the prior concealed state should be forgotten, much like the input gate does.

Output Gate (o): $o(t) = (W\,o * [h(t-1), x(t)] + bo)$ (10)

The output gate can control how much data should be sent out from the cell.

Cell State Update (C): $C(t) = f(t)*C(t-1)+i(t)*tanh(W\,c*[h(t-1), x(t)]+bc)$ (11)

The forget gate, input gate, and a fresh candidate value acquired via a tanh activation function are used to update the cell state.

Hidden State (h): $h(t) = o(t) * tanh(C(t))$ (12)

The hidden state is determined by applying the output gate to the cell state using a Tanh activation function. Bidirectional long short-term memory (BI LSTM) is used for event-driven stock prediction. The process involves data preparation, feature extraction, data grouping, deep learning framework development, evaluation, and fine-tuning. The model enhances risk management in stock trading and investing strategies.

B. Decision Tree Regression

Machine learning algorithm regression predicts continuous numerical values by partitioning input data recursively according to attributes, creating a tree-like model with low variance. The algorithm divides a data set into subgroups using feature and threshold reduction, assigning a predicted value to each leaf node. It then traverses the decision tree to produce regression predictions based on input data features. Decision tree regression captures complex correlations between features and target variables, requires minimal data preprocessing, and is easy to read. However, it is susceptible to overfitting due to deep tree growth or dataset noise. Implementing a Decision Tree Regression model involves collecting stock price data and labeling it with associated stock price movement. Data splitting is used to split the dataset into training and testing sets, with hyperparameters tweaked to balance model complexity and predictive power.

C. XG Boost

XGBoost is a machine learning method that uses gradient boosting to solve classification and regression problems. It is efficient and accurate due to parallelization and regularization techniques, and supports a wider range of hyperparameters. It offers faster training time, regularization to avoid overfitting, and more control over the model. To use XGBoost in stock prediction, historical stock price data and events affecting stock prices must be pre-processed to remove outliers and convert into a model that XGBoost can understand. The model is trained on historical data, evaluated on a holdout dataset, and deployed to production for real-time stock value predictions.

D. Ridge Regression

Ridge regression is a statistical technique used to correct for multicollinearity in linear regression models by increasing the sum of squared coefficients with a penalty. It reduces coefficients and stabilizes the model by increasing the sum of squared coefficients. The hyperparameter alpha regulates shrinkage, with higher values causing greater shrinkage and lower values causing less. Ridge regression improves accuracy, reduces standard errors, and is simple to implement. However, it can reduce interpretability and be sensitive to the choice of hyperparameter alpha. To use ridge regression in stock prediction, collect historical data, choose features, prepare the data model, train it, test it, and use it to predict future stock prices.

E. GloVe Embedding

GloVe is a statistical model that extracts word embeddings from text, focusing on syntactic and semantic connections. It uses word co-occurrence in a corpus to infer semantic similarity. GloVe is effective in natural language processing applications like text categorization, sentiment analysis, and question answering. It is simple, effective, and available as a pre-trained model. However, it is static and not as accurate as other methods like ELMo. GloVe's advantages include its simplicity, extensive training on text, and ease of use.

V. RESULT

A. ML Models

Accuracy in classification and regression tasks depends on the quality of the training dataset and the complexity of the model. The XGBoost model, trained on time series data, achieved high accuracy and low MSE. The BI LSTM model, trained on property prices data, also showed precision in forecasting. Cross-validation optimization optimized the alpha hyperparameter, indicating the model's precision in predicting the target variable.

TABLE I. MSE AND ACCURACY FOR TELECOM SECTOR

Stock	Model	MSE	Accuracy
Idea	XG Boost	8.00e+05	0.5235
	Ridge Regression	3.310e-04	0.9998
	BI LSTM	5.51e+15	0.7750
	Decision Tree	8.18e+14	0.773
Reliance	XG Boost	3.55e+05	0.3149
	Ridge Regression	9.1315	0.9690
	BI LSTM	1.07e+13	0.7399
	Decision Tree	1.964e+10	0.9205
BHARATI Airtel	XG Boost	2.30e+05	0.6765
	Ridge Regression	19.82e+00	0.97
	BI LSTM	5.413e+13	0.6365
	Decision Tree	3.873e+12	0.1410

TABLE II MSE AND ACCURACY FOR RAILWAY SECTOR

Stock	Model	MSE	Accuracy
Texmaco	XG Boost	3.984e+04	0.9675
	Ridge Regression	3.300e-01	0.9896
	BI LSTM	1.259e+12	0.5640
	Decision Tree	1.541e+09	0.3341
BEML	XG Boost	3.251e+04	0.9847
	Ridge Regression	1.553e+02	0.9888
	BI LSTM	2.528e+11	0.372
	Decision Tree	4.457e+09	0.4641
Concor	XG Boost	8.681e+04	0.9485
	Ridge Regression	1.505e+01	0.986
	BI LSTM	1.130e+12	0.46
	Decision Tree	2.809e+10	0.2765

TABLE III MSE AND ACCURACY FOR IT SECTOR

Stock	Model	MSE	Accuracy
INFOSYS	XG Boost	4.419e+05	0.8146
	Ridge Regression	7.400e+01	0.9950
	BI LSTM	1.707e+13	0.407
	Decision Tree	1.205e+11	0.3190
LTI MINDTREE	XG Boost	7.458e+03	0.738
	Ridge Regression	1.208e+03	0.9790
	BI LSTM	3.35e+10	0.814
	Decision Tree	9.826e+08	0.8051
WIPRO	XG Boost	6.599e+05	0.5687
	Ridge Regression	4.194e+00	0.9796
	BI LSTM	2.436e+13	0.38
	Decision Tree	1.041e+11	0.2633
ORACLE	XG Boost	1.384e+05	0.962
	Ridge Regression	5.090e-01	0.987
	BI LSTM	3.358e+13	0.431
	Decision Tree	7.830e+09	0.305

TABLE IV MSE AND ACCURACY FOR FINANCE SECTOR

Stock	Model	MSE	Accuracy
HDFC BANK	XG Boost	2.088e+05	0.9742
	Ridge Regression	5.344e+01	0.9622
	BI LSTM	1.978e+13	0.57
	Decision Tree	8.590e+10	0.207
KOTAK BANK	XG Boost	8.344e+04	0.1158
	Ridge Regression	1.338e+01	0.9875
	BI LSTM	2.432e+12	0.7804
	Decision Tree	2.616e+10	0.9672
ICICI BANK	XG Boost	1.765e+05	0.571
	Ridge Regression	1.338e+00	0.9937
	BI LSTM	4.922e+03	0.7289
	Decision Tree	2.115e+12	0.8293

TABLE V MSE AND ACCURACY FOR HEALTHCARE SECTOR

Stock	Model	MSE	Accuracy
LUPIN	XG Boost	1.376e+05	0.9665
	Ridge Regression	1.508e+01	0.9866
	BI LSTM	1.466e+12	0.159
	Decision Tree	1.844e+10	0.3052
SUNPHARMA	XG Boost	3.273e+05	0.8159
	Ridge Regression	2.726e+01	0.9687
	BI LSTM	1.424e+13	0.212
	Decision Tree	1.574e+11	0.5707
BIOCON	XG Boost	1.483e+05	0.8304
	Ridge Regression	2.730e+00	0.9874
	BI LSTM	4.950e+10	0.425
	Decision Tree	8.980e+10	0.1926
APOLLO	XG Boost	3.114e+04	0.8828
	Ridge Regression	1.049e+03	0.9716
	BI LSTM	2.155e+10	0.645
	Decision Tree	1.894e+ 09	0.7496

B. Our Model

In this model, XGBoost and Bidirectional LSTM models are used to demonstrate a novel stock price prediction method. The program predicts stock prices with accuracy and spots possible financial market movements. It makes use of collective learning and detects data's past and future sequential patterns. Target factors and stock market characteristics are included in the dataset.
Accuracy 0.85, Macro Avg 0.63 and Weighted Avg 0.83

TABLE VI CONFUSION MATRIX OF OUR MODEL

Class	Precision	Recall	F1-Score	Support
0.0	0.52	0.25	0.34	2229
1.0	0.88	0.96	0.91	12210

Validation Results: Accuracy: 0.849 Mean Squared Error (MSE): 0.151
Test Results: Accuracy: 0.842 Mean Squared Error (MSE):0.157

TABLE VII CONFUSION MATRIX OF OUR MODEL AFTER CROSS VALIDATION

Class	Precision	Recall	F1-Score	Support
0.0	0.52	0.23	0.32	2317
1.0	0.87	0.96	0.91	12123

The model's performance was assessed using the exact and root mean squared error metrics. The K-fold cross-validation method guarantees robustness across various data partitions. The model significantly contributes to applications like stock price forecasting and other time series forecasting due to its remarkable performance and interpretability.

Cross-Validation Results: Accuracy: 0.847 Mean Squared Error (MSE): 0.153

VI. CONCLUSION

The proposed method for predicting stock prices using XGBoost and Bidirectional LSTM models outperforms simple machine learning and regression models. It accurately predicts stock prices and identifies financial market movements, demonstrating its ability to manage complex time-series data. The model uses ensemble learning strategies to recognize complex patterns and improves performance by considering temporal dependencies. Its higher performance is compared to conventional regression models, and its value in stock price forecasting is increased by its accuracy and fewer prediction mistakes. K-Fold Cross-Validation ensures the model's generalizability across multiple datasets.

REFERENCES

[1] Q. Li, J. Tan, J. Wang and H. Chen, "A Multimodal Event-Driven LSTM Model for Stock Prediction Using Online News," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 10, pp. 3323-3337, 1 Oct. 2021, doi: 10.1109/TKDE.2020.2968894.

[2] Y. Lei, K. Zhou and Y. Liu, "Multi-Category Events Driven Stock Price Trends Prediction," 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China, 2018, pp. 497-501, doi: 10.1109/CCIS.2018.8691392.

[3] S. -S. Luo, Y. Weng, W. -W. Wang and W. -X. Hong, "L1- regularized logistic regression for event-driven stock market predic tion," 2017 12th International Conference on Computer Science and Education (ICCSE), Houston, TX, USA, 2017, pp. 536-541, doi: 10.1109/ICCSE.2017.8085550.

[4] P. Lin, X. Mo, G. Lin, L. Ling, T. Wei and W. Luo, "A News Driven Recurrent Neural Network for Market Volatility Prediction," 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), Nanjing, China, 2017, pp. 776-781, doi: 10.1109/ACPR.2017.35.

[5] Ding, X., Zhang, Y., Liu, T. and Duan, J., 2015, June. Deep learning for event-driven stock prediction. In Twenty-fourth international joint conference on artificial intelligence.