Stock Movement Prediction

Aishini Bhattacharjee
221AI007
Email: aishini.221ai007@nitk.edu.in

Arushi Biswas
221AI011
Email: arushi.221ai011@nitk.edu.in

Gunjan Das
221AI020
Email: gunjandas.221ai020@nitk.edu.in

Rohil Sharma
221AI033
Email: rohil.221ai033@nitk.edu.in

***Abstract*—The problem of stock movement prediction is a widely discussed one in academia. While there have been many complex models to process the data, there have been relatively lesser attempts at manipulating the data itself to give good results. The objective of our approach is to mine effective features from a causal graph and to observe how they have been influencing predictions when fed to a classifier. The analysis of the results does reveal the machine's relative inability to detect 'neutral' trends as opposed to 'upward' or 'downward' trends, which is why we attempt to use explainable AI to our advantage to see what feature is responsible for what kind of outcome.**

I. Introduction

Through this project, we aim to address the problem of stock movement prediction. As we are aware, financial modeling is a topic of frequent discussion, and the process itself is extremely high-risk, due to which we are proposing a method which employs a data-driven approach leading to decision-making in this regard.

We have come across many research endeavors which have employed graph attention networks and many other complex models which have improved accuracy over the years. [6] In this article, we have attempted to come up with a more data driven approach [1], where we extract features from a causal graph and financial data, and construct a comprehensive dataset, which we later apply to simpler classification models to magnify results.

Explainable AI has also been an explored area under financial modeling. In our methodology, we have used explainable AI to get deeper insights on the nature and influences of the engineered features in predicting stock movement.

A. Applications of the Project

Besides stating whether the investor should trade a particular stock or not (based on financial news), our approach helps to focus on the various features and technical indicators that affect the prediction. This can equip traders with objective signs to look out for before investing. We are achieving this by combining financial news sentiments with technical indicators available in numerical stock data.

The remaining sections in this article proceed as follows:

Section II covers the literature review. Section III covers methodology. Section IV covers the implementation. Section V covers results and analysis, and section VI covers conclusion and future endeavors.

II. Literature Review

A. Classification Models

Various classification models have been discussed before, which can probably address the problem of stock movement prediction, including support vector machines and random forests. [5] Yu Ma [10] has proposed a Multi-source aggregated classification by pre-training an embedding feature generator by fitting financial news to real stock price movements. Saetia and Yokrattanasak [15] have used stock market data and Google trends keywords to map technical indicators with those keyword, followed by modelling.

B. Causal Analysis

K. Nam and N. Seong [12] presented an approach that took into consideration stocks of a particular index, as well as company-specific stocks. They extracted news concerning those events by means of news crawling, and extracted the features of all causal firms to be fed into a machine learning model. Y. Deng [4] has proposed an approach that involves clustering of similar expressions in financial news, and forms a causal graph of the clusters, following which, a Hierarchical Attention Network is implemented for predicting stock trends.

C. Explainable AI

In explainable AI, there is a very popular approach called LIME [14] that shows how, instance by instance, various features are responsible for the outcome of a machine learning model. It assigns positive or negative probabilities to features responsible for increasing or decreasing the chances of that particular prediction, respectively.

These academic dissertations have served as background for the approach that we developed. There have also been many other approaches, including Graph Attention Models, semantic graphs, graph neural networks and community detection, which have been summarised in Table I.

III. Methodology

A. Dataset Pre-processing

We used two datasets:

- A pre-curated financial news dataset containing data from January 2020 to April 2024. [13] It contained attributes like date of publishing, headline, description, articlebody,

TABLE I
SUMMARY OF LITERATURE SURVEY

Paper	Merits	Limitations	Suggestions
ML-GAT:A Multilevel Graph Attention Model for Stock Prediction (2022) [6]	Novel GNN-based attention network to capture relationships between stocks with higher accuracy.	Architecture is complex, leading to risk of over- fitting.	Usage of dimensionality reduction and regularization could mitigate overfitting.
FLAG: Fusing Logic and Semantic Graphs of Financial News (2022) [8]	Usage of clauses enhances granularity in the text. Attention mechanism focuses on relevant parts of the news	High computational cost, Limited exploration of temporal modeling approaches.	Integration of temporal modeling to capture the dynamic nature of stocks.
KG and DL with a stock price prediction network (2022) [16]	ConvLSTM: For spatio- temporal data, Piecewise loss function to improve model robustness.	Complexity and Scalability becomes an issue. Accurate identification of mutation points can be difficult.	Use dynamic graph learning techniques, Integrate explainable AI techniques like feature importance analysis.
Forecasting stock prices changes using LSTM neural network with symbolic genetic programming (2024) [7]	Significant enhancement in prediction accuracy for stock returns.Outperforms major Chinese stock indexes and utilizes a large dataset of 4500 stocks, providing robust analysis.	Requires substantial computational resources for data processing and model training.High dimensionality and extensive features increase the risk of overfitting.	Automate feature selection to minimize manual intervention.
Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets (2022) [3]	Uses state-of-the-art deep learning models for improved prediction accuracy.Hybrid Approach of combining sentiment analysis with financial data provides more accurate predictions.	High computational complexity, Predictions heavily rely on the quality and accuracy of sentiment data, Deep learning models can overfit.	Implement better filtering techniques to ensure high-quality sentiment data.Expand data sources.
NLP in Stock Market Prediction: A Review (2022) [1]	Effectively integrates financial news, social media, and historical price data,Strong emphasis on sentiment analysis, using VADER and TextBlob.	Potential biases in the textual data sources are not fully addressed. The models' adaptability to sudden market crashes or extreme volatility is not explored in depth.	Develop methods to mitigate biases in sentiment analysis from social media sources.Integrate risk management strategies within the predictive models to account for market volatility.
Multi-feature fusion stock prediction based on knowledge graph (2024) [9]	KGs facilitate the discovery of potential correlations between related stocks, which traditional models often overlook.By incorporating sentiment analysis of news and investor comments, KGs help in understanding investor behavior.	As the size of the knowledge graph grows, managing and querying the data becomes increasingly challenging. KGs rely on predefined relationships and entities, which might not capture emerging trends or new relationships that could impact stock prices.	Implement more efficient data management and querying techniques to address scalability issues,Expand the model to analyze and predict trends across different markets (e.g., commodities, bonds).
A knowledge graph–GCN–community Detection integrated model for large-scale stock price prediction (2022) [18]	Community detection within the graph is beneficial as it helps in finding groups of stocks that behave similarly or are influenced by similar market factors.	Computationally intensive, especially when dealing with very large graphs. Risk of overfitting.	Combining GCNs with other models, such as LSTM networks, could help capture the temporal dynamics of stock prices and the relationships between different stocks.
ChatGPT informed graph neural network for stock movement prediction. (2023) [2]	Incorporation of Domain Knowledge. Better Performance on Sparse Data.	The approach heavily relies on the availability and accuracy of domain knowledge.	Automated knowledge Integration. Cross-Domain Knowledge Transfer.
Temporal and Heterogeneous Graph Neural Network for Financial Time Series Prediction (2023) [19]	Novel Approach to Temporal Graphs. Multi-Aspect Data Handling.	The model's architecture is complex. Limited applicability in resource-constrained environments due to high computational requirements.	Enhanced Interpretability: incorporating attention mechanisms that highlight which parts of the graph or which time points are most influential in making predictions. Optimization for Efficiency.
Natural Language Processing and Multimodal Stock Price Prediction (2024) [17]	Use of Percentage Change. Focus on Sector-Specific Data.	Neglect of Temporal Factors. Relatively small dataset for stock prediction models, limiting generalizability and robustness.	Incorporating additional features like trading volumes, earnings reports, or social media sentiment. Combining the BERT model with other predictive models, like LSTM or GRU networks.

corresponding company and author. The training dataset contained 1884 such instances.

- A numerical stock price dataset corresponding to those same dates. We used Yahoo Finance API to extract the

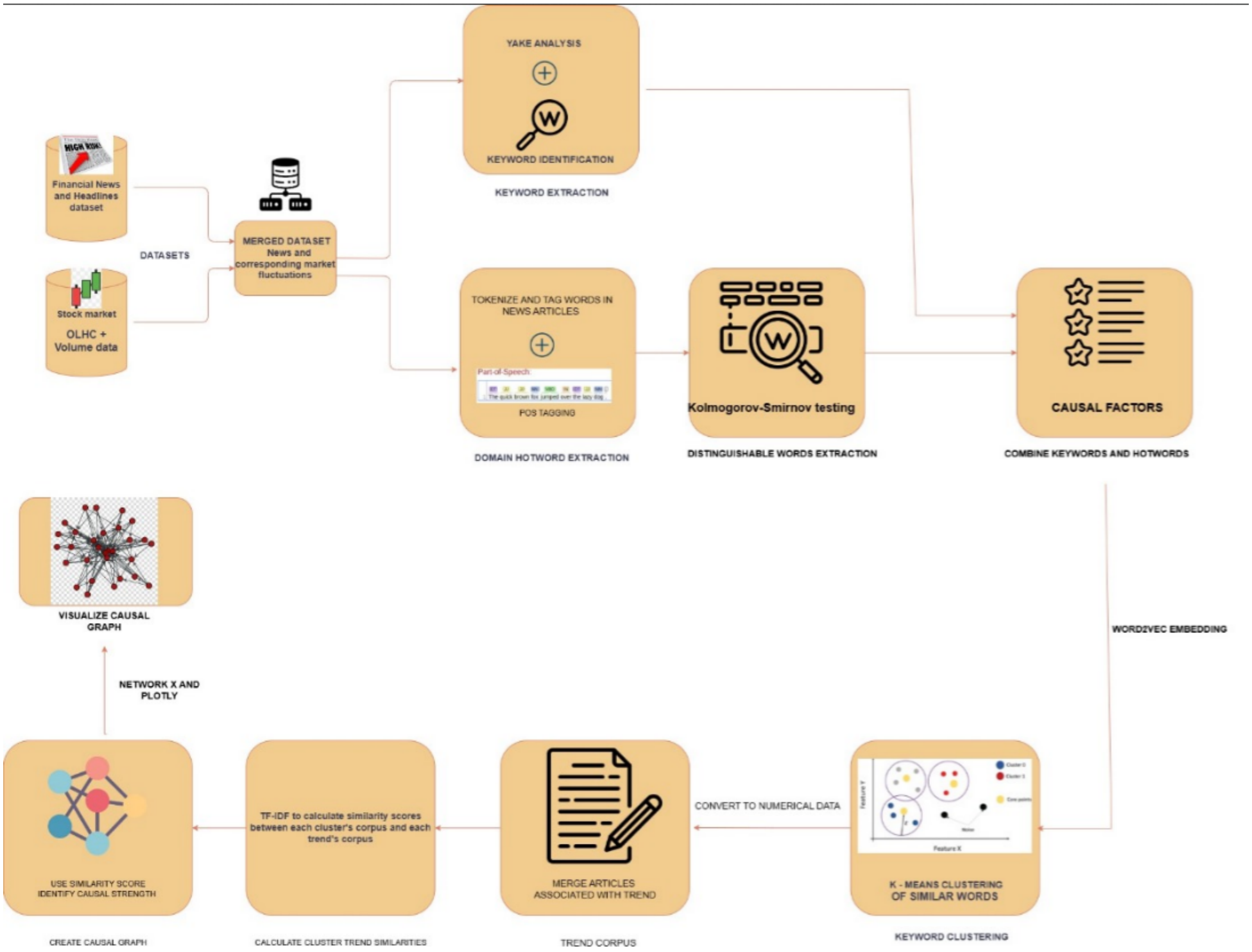
YAKE ANALYSIS
KEYWORD IDENTIFICATION
KEYWORD EXTRACTION
Financial News and Headlines dataset
DATASETS
Stock market
OLHC + Volume data
MERGED DATASET News and corresponding market fluctuations
TOKENIZE AND TAG WORDS IN NEWS ARTICLES
Part-of-Speech
The quick brown fox jumped over the lazy dog
POS TAGGING
DOMAIN HOTWORD EXTRACTION
Kolmogorov-Smirnov testing
DISTINGUISHABLE WORDS EXTRACTION
CAUSAL FACTORS
COMBINE KEYWORDS AND HOTWORDS
WORD2VEC EMBEDDING
VISUALIZE CAUSAL GRAPH
NETWORK X AND PLOTLY
USE SIMILARITY SCORE IDENTIFY CAUSAL STRENGTH
CREATE CAUSAL GRAPH
TF-IDF to calculate similarity scores between each cluster's corpus and each trend's corpus
CALCULATE CLUSTER TREND SIMILARITIES
MERGE ARTICLES ASSOCIATED WITH TREND
TREND CORPUS
CONVERT TO NUMERICAL DATA
Feature Y
Feature X
K - MEANS CLUSTERING OF SIMILAR WORDS
KEYWORD CLUSTERING

Fig. 1. Methodology for Causal Graph Construction

data. This dataset contained attributes like opening and closing prices, highest and lowest prices, and the volume of stocks exchanged during that time period, for the National Stock Exchange Index.

Using the numerical dataset, some other features were introduced, like:

- **Return:** The percentage change in the closing prices of the current date and the previous date.
- **Trend:** This feature is based on the return. if return is less than 0.1 per cent, the trend is set to 'Down'. If return is greater than 0.1 per cent, the trend is set to 'Up'. It is 'Neutral' otherwise. The proportion of each of these categories in the training dataset is displayed in Fig. 3.

The following points are to be noted in this regard:

- The distribution between data points among the three trends are uneven as can be seen in the figure. We decided to keep this data untouched because historical stock data represents a form of time series, and removing any data points or augmenting the data in any way might ultimately disrupt decision-making.
- We chose 0.1 per cent difference between daily returns to be the threshold because daily closing prices are of the order of 10^4. Therefore, the difference in return at 0.1 per cent is expected to be of the order of 10. This seems a reasonable order of difference in closing price per stock, thus magnifying the effect when thousands of such stocks are purchased.

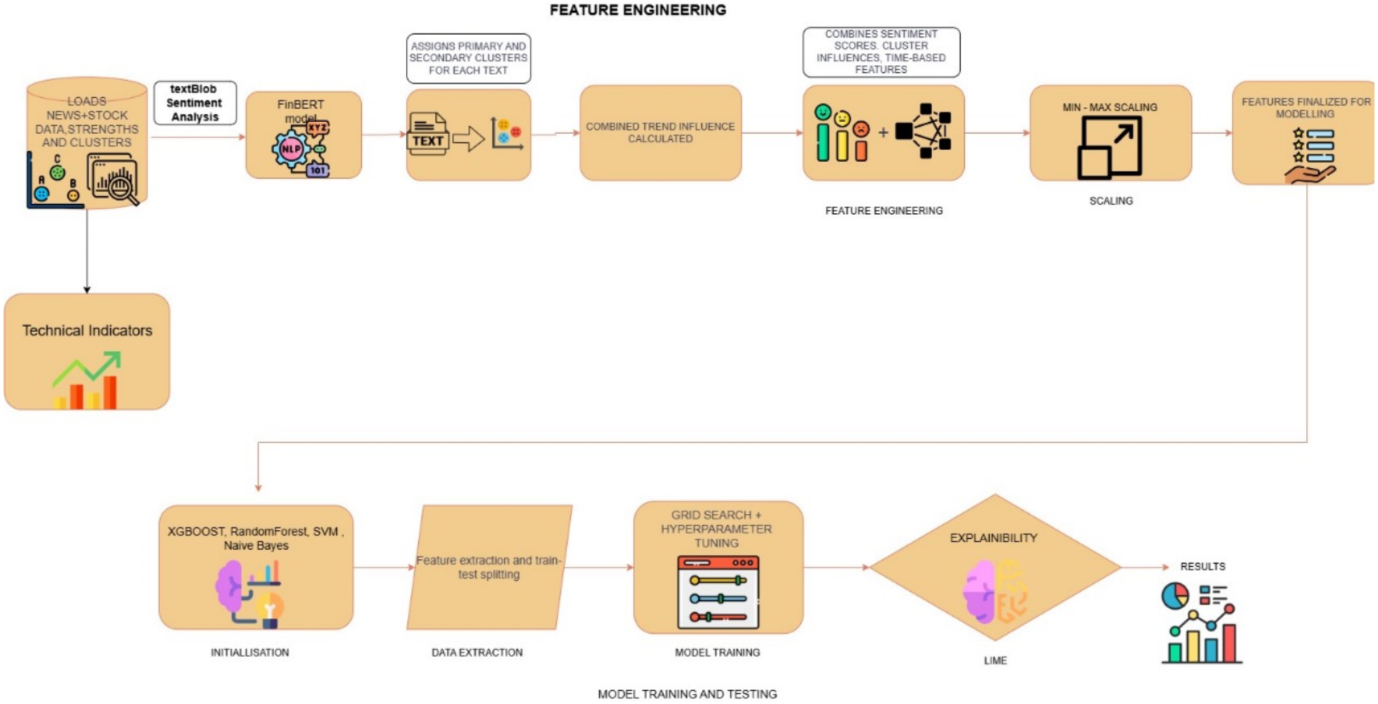
FEATURE ENGINEERING
LOADS NEWS+STOCK DATA,STRENGTHS AND CLUSTERS
A
B
C
textBlob Sentiment Analysis
FinBERT model
XYZ
NLP
101
ASSIGNS PRIMARY AND SECONDARY CLUSTERS FOR EACH TEXT
TEXT
COMBINED TREND INFLUENCE CALCULATED
COMBINES SENTIMENT SCORES. CLUSTER INFLUENCES, TIME-BASED FEATURES
FEATURE ENGINEERING
MIN - MAX SCALING
SCALING
FEATURES FINALIZED FOR MODELLING
Technical Indicators
XGBOOST, RandomForest, SVM , Naive Bayes
INITIALLISATION
Feature extraction and train-test splitting
DATA EXTRACTION
GRID SEARCH + HYPERPARAMETER TUNING
MODEL TRAINING
EXPLAINIBILITY
LIME
RESULTS
MODEL TRAINING AND TESTING

Fig. 2. Methodology for Feature Set Building and Training

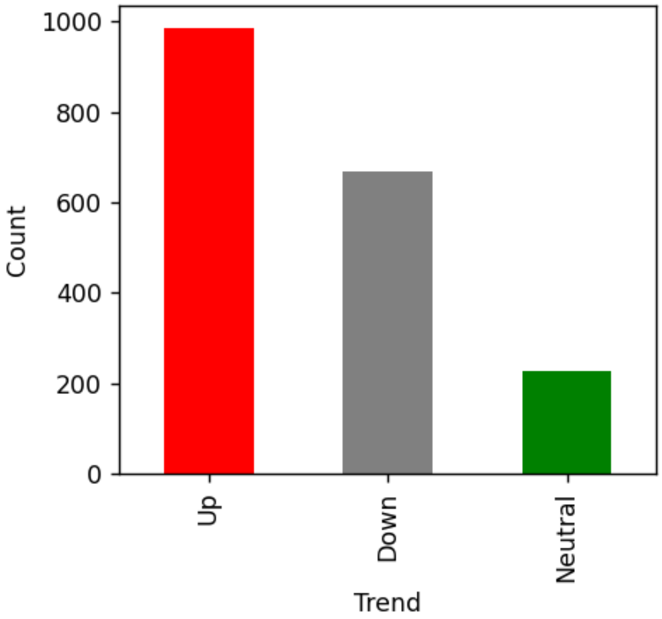

Fig. 3. Distribution of Financial News Data

B. Causal Graph Construction

The causal graph was constructed using a method partially in compliance with Y. Deng's methodology [4]. This process can be summarised as follows:

- From each piece of financial news, stopwords are omitted and individual words are tokenized, also labeling their parts of speech. The rest of the terms are deemed as keywords. Hotwords are extracted by filtering verbs and adverbs considered indicative of significant actions or qualities. After filtering, hotwords for each trend are pooled together, and the most frequently occurring words for each trend are selected and kept as representative of that trend.
- Words which are capable of distinguishing between trends are identified by comparing their relative differences in frequencies across the different trends using the Kolmogorov-Smirnov (KS) test [11].
- Keywords are clustered using K-means clustering using Word2Vec embeddings. The keyword vectors generated are grouped together based on their similarities.
- A trend specific corpus is created by pooling all the terms corresponding to that particular trend. TF-IDF vectorization is used to vectorize the words in clusters as well as the trend corpora, and cosine similarity is used to calculate the similarities between each cluster and a trend.
- Each cluster is assigned the trend node to which it bears the most similarity.
- The graph is constructed by making three subgraphs for each trend, with most similar clusters being associated with each trend. The strength between two nodes is equal to the similarity between a cluster and a trend in accordance with the previous steps.

C. Feature Engineering

In this section, sentiment analysis is first carried out on each text. TextBlob is used for generating a basic sentiment polarity score, whereas FinBERT is used for tailoring the results to the specific domain. In our use case, the aggregate sentiment score had 70 per cent contribution from FinBERT and 30 per cent from TextBlob. Tags present for each financial news are leveraged to find out how similar the news is to the causal clusters generated. The clusters are sorted by a match score to assign a primary and secondary cluster to each text. The causal strengths of the primary clusters to all 3 trends are noted, and a weighted sum of the primary and secondary cluster are calculated to come up with the influence for each trend for a given text. The secondary cluster carries lesser weight as compared to the primary cluster. The combined trend influece is calculated in the following way:

$$\begin{aligned} \text{combined\_trend\_influence} = {} & \text{influence\_Trend\_Up} \\ & - \text{influence\_Trend\_Down} \\ & + 0.5 \times \text{influence\_Trend\_Neutral} \end{aligned}$$

For all 3 trends, sentiment interaction features are calculated using the following equation:

$$\begin{aligned} \text{sentiment\_trend\_interaction} = {} & \text{sentiment\_score} \\ & \times \text{influence\_Trend} \end{aligned}$$

1) Technical Indicators: Technical indicators calculated from the stock data have been added to the already present financial data. They are:

- **Relative Strength Indicator (RSI):** It is a momentum oscillator that measures the speed and change of price movements. It is calculated using the following equation:

  $$\text{RSI} = 100 - \frac{100}{1 + \text{RS}}$$

  RS or relative strength is the ratio of the average of up closing prices to down closing prices over a certain period. RSI was calculated using the inbuilt $RSIIndicator$ class.
- **Moving Average Convergence Divergence (MACD):** MACD is a trend-following momentum indicator that shows the relationship between two moving averages of a security's price. It is used to identify potential buy/sell signals based on crossovers, divergence, and rapid rises/falls. It was calculated using the inbuilt $MACD$ function of the $trend$ class.
- **Returns:** Returns represent the percentage changes in closing price over one or many days.
- **Simple Moving Averages (SMA):** SMA helps to smoothen out the data and to find the direction of trend by calculating the mean of closing prices over a sliding window of a certain number of days.

2) Dataset Building: The dataset to be passed to the classification model is constructed using the features described above. They became easy to process and feed to a model because they were all numeric in nature.

IV. Implementation

A. Models Used

The feature matrix for training the classification model was constructed in part C of the third section, and the labels for each row in the dataset has already been constructed as the $Trend$, as described in part A of that section.

The models used were:

- **XGBoost:** It build a sequence of decision trees where each subsequent tree tries to correct the error by the previous tree.
- **Random forests:** A collection of decision trees are constructed by training them on different random subsets of the data.
- **Support Vector Machines:** This model attempts to construct hyperplanes between data points to ensure the most efficient and accurate division between classes.
- **Naive Bayes Classifier:** Based on Bayes' theorem, its predictions are based on the calculation of likelihood of the features given in every class.
- **Convolutional Neural Network:** CNNs for numerical data can capture local patterns and interactions between numerical features.

B. Explanation of Outcomes

As an add-on, after predictions are made by a model, the prediction can be examined using LIME, that lists down the probabilities of each feature influencing for or against a particular outcome. This approach is employed to understand what features to emphasize on while using any model for the classification of stock movement trends.

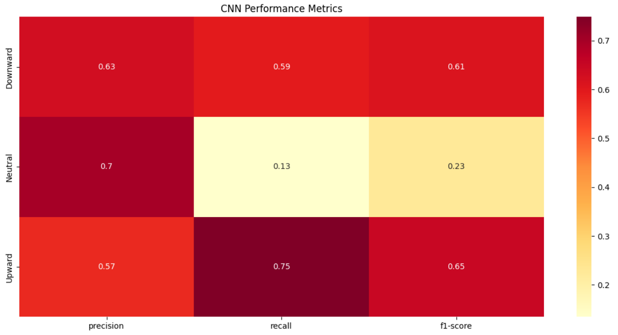

Fig. 8. Metrics for CNN

V. Results and Analysis

The evaluation metrics used are:

- **Precision:** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

  $$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall:** Recall (also known as Sensitivity or True Positive Rate) is the ratio of correctly predicted positive observations to all the observations in the actual class.

  $$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1-Score:** The F1 score is the harmonic mean of precision and recall. It provides a single metric that

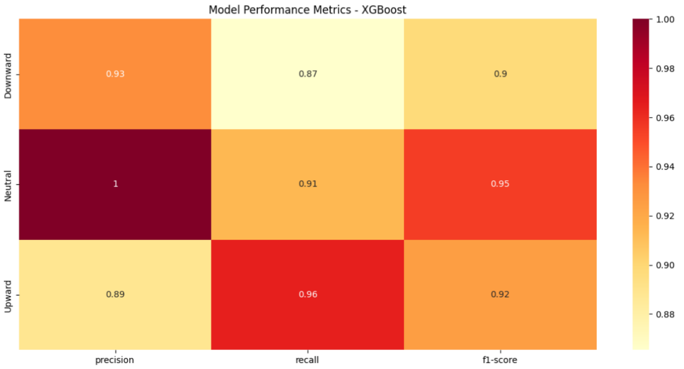

Fig. 4. Metrics for XGBoost

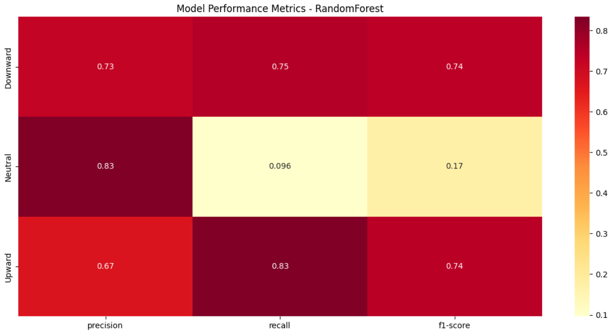

Fig. 5. Metrics for Random Forests

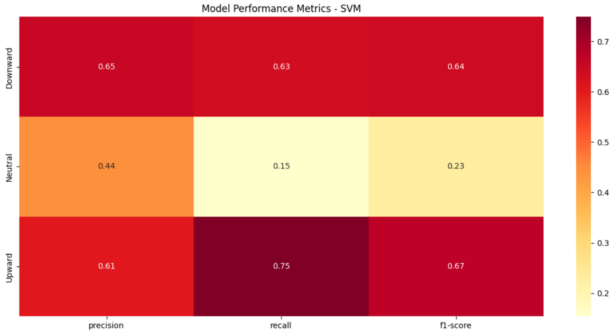

Fig. 6. Metrics for Support Vector Machine

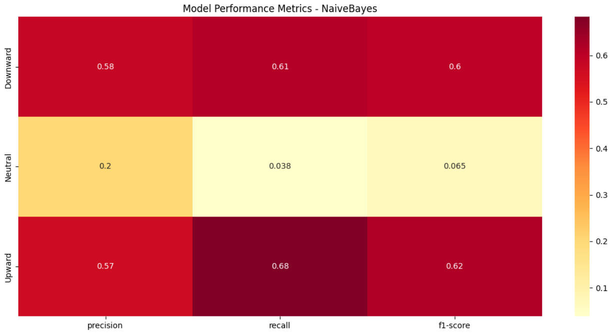

Fig. 7. Metrics for Naive Bayes Classifier

balances both precision and recall. It is an indication of how balanced the dataset is.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Accuracy:** Accuracy measures the proportion of correct predictions out of the total predictions. This result was tabulated even though the previous three metrics are more applicable for unbalanced datasets like the one that we used.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Cross validation score:** It estimates the model's performance by training and evaluating it on different subsets of the data, thus helping mitigate the risk of overfitting. Accuracy is used as the 'score' in this case.

$$\text{CV}_k = \frac{1}{k} \sum_{i=1}^{k} \text{Score}_i$$

The accuracies have been tabulated in Table II, and the performance metrics have been visualised in figures 4, 5, 6, 7 and 8. As can be seem from the visualised and tabulated results, XGBoost significantly outperformed the other models, not only in terms of accuracy, but also in terms of precision and recall. One noticeable inconsistency is the significantly lower value for most metrics in the Neutral category. This might happen because there are less instances of financial news in the neutral category. Another important reason is semantic ambiguity. Since the causal graph clustering takes in action words as 'domain hotwords', many verbs belonging in the neutral category might be in the other two categories because neutral action words are fairly common in everyday speech. Only exception to this is the XGBoost model, which gave exceptionally high precision and recall even for the neutral category, thus leading to the inference that this model might be immune to the imbalance in the training dataset.

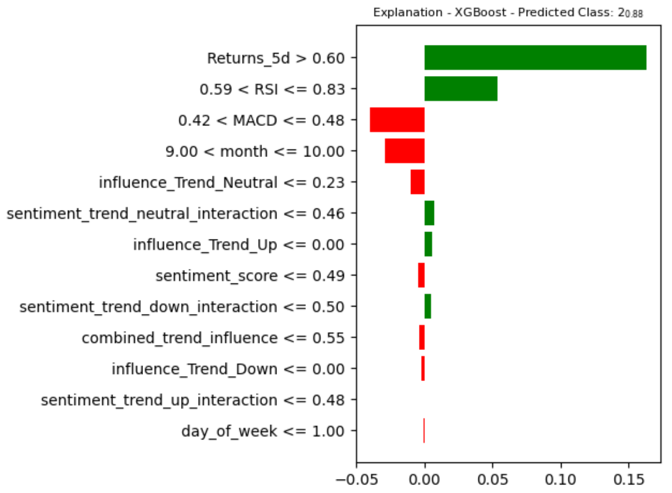

Fig. 9. LIME explanations for one instance

To gain insights on what actually affected a prediction, LIME was tried out on instances in the dataset. As can be seen from Fig. 9, we observe that the predicted class id 'Up', denoted by 2. What influenced the upward trend the most was a 5-day return of more than 0.6 and an RSI of 0.59 to 0.83. However, an MACD of 0.42 to 0.48 seemed to play a role in attempting to make the prediction 'Neutral' or 'Down'. Interestingly, the same effect was observed if the month of trading was October (10). By analysing a group of such features in a similar way, we can get better insights on which technical indicators to look for while trading a stock, even the preferred days and months of the year when returns might be promising.

TABLE II
TEST SET METRICS OF THE MODELS

Model	**Accuracy**	**Best CV Score**
XGBoost	0.9191	0.8822
Random Forests	0.6801	0.6712
SVM	0.6324	0.6243
Naive Bayes	0.5294	0.5
CNN	0.6342	0.612

VI. Conclusion

In this article, we proposed a data-driven approach for stock movement prediction wherein, after constructing a causal graph from similar features in financial news, we carry out feature engineering on the stock data and the causal clusters to produce a training dataset to which a classifier can be trained. While the results for upward and downward trends seem justifiable, scores for neutral trends have been significantly lower due to textual ambiguity and a general lack of notion on what counts as 'neutral' or 'not neutral' when considering action words like verbs or adverbs. To get into the depth of predictions, we implemented LIME on test instances to see what factors drove the ML model to predict a particular class. In future, we aim to use explainability on all test cases and carry out a statistical analysis on what defines a 'good' feature set for any kind of stock trend, aiming to potentially help investors in decision-making.

References

[1] Rodrigue Andrawos. Nlp in stock market prediction: A review. 06 2022.

[2] Zihan Chen, Lei Zheng, Cheng Lu, Jialu Yuan, and Di Zhu. Chatgpt informed graph neural network for stock movement prediction. *SSRN Electronic Journal*, 2023.

[3] Narayana Darapaneni, Anwesh Reddy Paduri, Himank Sharma, Milind Manjrekar, Nutan Hindlekar, Pranali Bhagat, Usha Aiyer, and Yogesh Agarwal. Stock price prediction using sentiment analysis and deep learning for indian markets, 2022.

[4] Yiqi Deng, Yuzhi Liang, and Siu-Ming Yiu. Towards interpretable stock trend prediction through causal inference. *Expert Systems with Applications*, 238:121654, 2024.

[5] H. H. Htun and M. Biehl. Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation*, 2023.

[6] Kun Huang, Xiaoming Li, Fangyuan Liu, Xiaoping Yang, and Wei Yu. Ml-gat:a multilevel graph attention model for stock prediction. *IEEE Access*, 10:86408–86422, 2022.

[7] Kamaruddin N. Yuhaniz S.S. et al. Li, Q. Forecasting stock prices changes using long-short term memory neural network with symbolic genetic programming. *Sci Rep 14, 422*, 2024.

[8] Angela Lil, Jiduan Liu, Yuyong Lil, and Fan Meng. FLAG: Stock Movement Prediction via Fusing Logic and Semantic Graphs of Financial News . In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 190–197, 2022.

[9] Zhenghao Liu, Yuxing Qian, Wenlong Lv, Yanbin Fang, and Shenglan Liu. Multi-feature fusion stock prediction based on knowledge graph. *The Electronic Library*, 42, 06 2024.

[10] Yu Ma, Rui Mao, Qika Lin, Peng Wu, and Erik Cambria. Multi-source aggregated classification for stock price movement prediction. *Information Fusion*, 2023.

[11] Frank J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[12] KiHwan Nam and NohYoon Seong. Financial news-based stock movement prediction using causality analysis of influence in the korean stock market. *Decision Support Systems*, 2019.

[13] Chaithra National Institute of Technology Karnataka. Nifty 50 financial news, 2024.

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[15] Kittipob Saetia and Jiraphat Yokrattanasak. Stock movement prediction using machine learning based on technical indicators and google trend searches in thailand. *International Journal of Financial Studies*, 2022.

[16] Meiyao Tao, Shanshan Gao, Deqian Mao, and Hong Huang. Knowledge graph and deep learning combined with a stock price prediction network focusing on related stocks and mutation points. *Journal of King Saud University - Computer and Information Sciences*, 34(7):4322–4334, 2022.

[17] Kevin Taylor and Jerry Ng. Natural language processing and multimodal stock price prediction, 2024.

[18] Ting Wang, Jiale Guo, Yuehui Shan, Yueyao Zhang, Bo Peng, and Zhuang Wu. A knowledge graph–gcn–community detection integrated model for large-scale stock price prediction. *Applied Soft Computing*, 145:110595, 2023.

[19] Sheng Xiang, Dawei Cheng, Chencheng Shang, Ying Zhang, and Yuqi Liang. Temporal and heterogeneous graph neural network for financial time series prediction. In *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*. ACM, October 2022.