# A Sentiment-Augmented Machine Learning Approach to Forecasting IHSG Prices Using XGBoost

1st Muhammad Rizki Wiratama
School of Computing
Telkom University
Bandung, Indon esia
rizkiwirat@student.telkomuniversity.ac.id

2nd Narita Aquarini
Ecole Doctorale Science Economics
Universit´e de Poitiers Intervenant Finance
La Rochelle, France
aquarinin@excelia-group.com

3rd Putu Harry Gunawan
CoE HUMIC, School of Computing
Telkom University
Bandung, Indonesia
phgunawan@telkomuniversity.ac.id

***Abstract*—Stocks are a commonly used investment instrument, representing ownership in a company, and offering opportunities for investors to gain profits through the appreciation of stock value as well as dividend distribution. As one of the main financial assets, stocks are also influenced by various external factors, such as economic conditions, government policies, and market sentiment. All of these factors play a crucial role in determining stock price movements. This study integrates sentiment analysis with the XGBoost algorithm to predict IHSG stock prices. By utilizing historical stock data and sentiment derived from financial news, the study evaluates the impact of sentiment data integration on prediction accuracy. Three types of returns (absolute, relative, and logarithmic) and five sentiment scenarios were employed to assess the contribution of sentiment features to the prediction model. The results indicate that sentiment integration consistently improves the predictive performance of the model compared to using historical data alone. Among the tested scenarios, Scenario 2, 4, and 5 demonstrated the best performance, with an RMSE value of 0.009163 and an MAE value of 0.007432, using the logarithmic return type. These findings suggest that incorporating sentiment features into predictive models can enhance the accuracy of stock price predictions and highlight the potential of Natural Language Processing (NLP) and Large Language Models (LLMs) in stock market analysis.**

***Keywords*—XGBoost algorithm, stock prediction, sentiment analysis, large language models**

## I. Introduction

Investment involves the allocation of capital into one or more assets with the objective of generating profits [1]. In the financial domain, one of the most common forms of investment is stocks. Stocks are securities that represent ownership in a company, granting shareholders rights to dividends or other benefits provided by the company to its shareholders [2]. By owning stocks, shareholders are entitled to dividends and can participate in electing the board of directors as well as in making corporate decisions. When stock values increase, they present investors with numerous profit opportunities.

Shares are issued by companies for various strategic purposes, including raising the necessary funds to support business expansion, conserving cash reserves, and leveraging high market valuations. The issuance of shares is one of the primary methods companies use to attract investments from both individual and institutional investors. Each share issued represents one unit of ownership in the company, granting the shareholder the right to receive a portion of the company's earnings as profits [3].

There are two types of risks in the stock market: systematic risk and unsystematic risk. Systematic risk refers to risks that affect a large segment of the market. These risks are unavoidable as they stem from external factors that impact the entire economy, such as interest rates, economic recessions, and government policies. Systematic risk influences three key areas: macroeconomics, financial crises, and market conditions [4], [5].

Unsystematic risk, on the other hand, is specific to a particular company or industry. This type of risk can be mitigated through diversification, which envolves spreading investments across different assets or sectors to reduce exposure to any single entity. It encompasses various factors, including management performance and financial conditions [6].

Since the government implemented regulations in the financial and banking sectors, including capital markets, Indonesia's capital market has experienced rapid growth. The Jakarta Composite Index (IHSG) is the most commonly used indicator to track the development of Indonesia's capital market, as it reflects the overall movement of stock prices and impacts traders and investors [7]. The Indonesia Stock Exchange (IDX) has the authority to include or exclude listed companies from the IHSG calculation to better represent market conditions [7].

Large Language Models (LLMs) like ChatGPT have garnered significant attention for their accuracy in recognizing and understanding text in sentiment analysis. Recent studies highlight their potential, including the challenges associated with their application. ChatGPT has demonstrated excellent performance in sentiment analysis, even without additional training [8]. However, the implementation costs and the necessity of refining prompts remain important considerations in practical use.

Research indicates a significant correlation between market sentiment and stock price movements, making sentiment analysis a valuable addition to stock price prediction models for deeper insights and improved accuracy. LLMs like

ChatGPT offer distinct advantages, including the ability to simplify complex financial insights and enhance sentiment analysis through robust natural language processing (NLP) capabilities, providing greater depth in financial contexts [9].

The prediction model utilizes XGBoost and incorporates sentiment analysis as a feature. XGBoost is a highly effective machine learning algorithm for processing data. It offers features for classification, regression, and ranking tasks [10]. The capabilities of XGBoost, such as pruning trees, parallel tree building, built-in cross-validation, handling missing data, and feature awareness, make it well-suited for prediction tasks [11]. This research aims to determine whether the addition of sentiment analysis can improve the accuracy of stock price predictions compared to using historical data alone.

This paper is structured as follows: The next section discusses the methodology employed in this study. It covers details about the dataset, sentiment analysis techniques, the model utilized, and the evaluation metrics (RMSE and MAE). The third section presents the experimental findings, model evaluation, and its advantages. The final section provides the conclusions derived from this study's findings. Additionally, it offers recommendations for further development and suggests potential future research directions to explore various approaches.

## II. METHODOLOGY

### A. Flow Design

The flow design of this study integrates stock data and sentiment data, where the combined data is used to predict IHSG stock prices based on returns. The process begins with the collection of stock data, during which a new variable, Return, is created using three targets: absolute, relative, and logarithmic returns. simultaneously, news data is collected through web scraping and undergoes preprocessing to clean and structure the dataset. After preprocessing, sentiment analysis is performed using a ChatGPT-based LLM to extract sentiment information from the news articles. The sentiment data is then merged with the stock data to create a comprehensive dataset for supervised learning, incorporating feature lagging to capture temporal patterns. To improve the robustness of the model, five sentiment scenarios are implemented and tested. Finally, predictions are made using the XGBoost model, and the results are analyzed and evaluated to assess the model's performance in predicting IHSG stock prices.

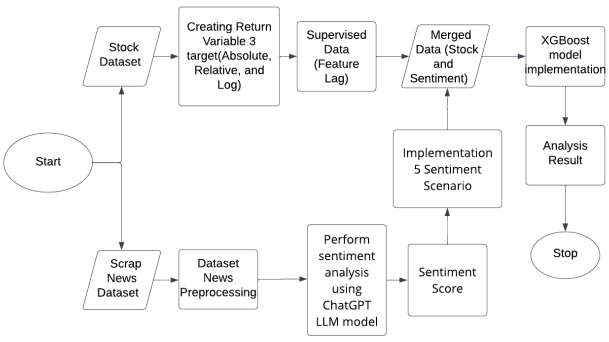


Fig. 1. Flow System Design

### B. Datasets

In this study, two datasets are utilized to support the analysis.

*1) **News Dataset**:* This dataset consists of a collection of news articles, The result of sentiment analysis will be incorporated as additional features for predicting stock prices. From the data presented in Table I.

TABLE I
DATASET IHSG STOCK

| Tanggal | Judul | Link | Isi |
|---|---|---|---|
| 21/05/2013 | IHSG Bakal... | https://mone... | IHSG Bakal... |
| 21/05/2013 | Proyeksi IHSG.. | https://mone.. | Proyeksi IHSG ... |
| 21/05/2013 | IHSG Dibuka.. | https://mone.. | IHSG Dibuka ... |

*2) **Stock Dataset**:* The Second dataset contains historical stock price data from several companies in indonesia, which serves as independent variables in the prediction model. the data is presented in Table II, and and plots illustrating Stock Price Movement Trends are shown in Fig. 2.

TABLE II
DATASET IHSG STOCK

| Date | Close | Open | High | Low | Volume |
|---|---|---|---|---|---|
| 01/01/2014 | 4274.18 | 4240.39 | 4274.18 | 4232.58 | NaN |
| 02/01/2014 | 4327.27 | 4294.49 | 4327.27 | 4287.81 | 2,31B |
| 03/01/2014 | 4257.66 | 4297.72 | 4298.23 | 4247.99 | 2,19B |
| 06/01/2014 | 4202.81 | 4259.58 | 4263.62 | 4188.38 | 1,97B |
| 07/01/2014 | 4175.81 | 4206.30 | 4212.32 | 4175.81 | 2,38B |


Fig. 2. Stock Movement Over Time

In addition to the historical stock data, three types of return variables relative return, absolute return, and logarithmic return are calculated to measure stock price changes over time, which play a crucial role in the development of the prediction model. Relative return is more optimal and stable, capable of identifying patterns that may not be evident in absolute return. Absolute return, on the other hand, is valuable for assessing risk, particularly during crisis periods. Logarithmic return, with its additive properties, simplifies statistical analysis and provides more optimal long-term prediction outputs [12]. The calculation of these returns is performed by applying the following formulas:

**Absolute Return**: This measures the difference between the closing price on day $t$ and the closing price on day $t-1$.

$$R_{\text{abs}} = P_t - P_{t-1} \tag{1}$$

**Relative Return**: This calculates the percentage change between the closing prices on day $t$ and $t-1$.

$$R_{\text{rel}} = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100 \tag{2}$$

**Logarithmic Return**: This measures the price change using the natural logarithm of the ratio of closing prices on day $t$ and $t-1$.

$$R_{\log} = \ln\left(\frac{P_t}{P_{t-1}}\right) \tag{3}$$

### C. Sentiment Analysis

In conducting this sentiment analysis, the ChatGPT LLM is utilized for its advanced natural language processing capabilities to analyze news content and sentiment analysis involves examining the sentiment conveyed in news data to understand stock market movements [13].

The study uses an LLM and batch processing to analyze financial news articles, extracting key entities, concise summaries, and frequently occurring keywords. The structured format helps analyze patterns and linkages within financial narratives. To adapt sentiment analysis for the financial domain, prompt engineering is applied. In this case, the model is guided to act as a professor of economics and stock markets, using the following prompt:

> "As a stock market analyst, review the following news article and classify the sentiment as Positive, Neutral, or Negative, focusing on its potential impact on the performance of the stock market."

The LLM generates sentiment scores, which are numerical values representing the degree of positivity, neutrality, or negativity. These scores, as illustrated in Table II, are subsequently transformed into a structured dataset.

TABLE III
SENTIMENT SCORES BY DATE

| Date | Sentiment Score |
|---|---|
| 2013-05-21 | -0.2 |
| 2013-05-22 | 1.0 |
| .. | .. |
| 2024-08-02 | -0.333 |
| 2024-08-04 | 1.0 |

The formula for calculating the sentiment score generated by the LLM model (such as ChatGPT) is as follows [14]:

$$S = \frac{\sum_{i=1}^{N} p(h_i) - \sum_{i=1}^{N} n(h_i)}{\sum_{i=1}^{N} p(h_i) + \sum_{i=1}^{N} n(h_i)} \tag{4}$$

- $h_i$ represents the $i$-th news headline of the day.
- $p(h_i)$ is a function that assigns a value of 1 if the headline is positive, and 0 if it is not.
- $n(h_i)$ is a function that assigns a value of 1 if the headline is negative, and 0 if it is not.
- $N$ indicates the total count of headlines.

### D. Sentiment Scenarios

Following the preliminary model construction, sentiment data is incorporated into the stock dataset as an auxiliary feature. Diverse scenarios are constructed to assess various methodologies for using sentiment data, including time series analysis, with the objective of identifying the most effective technique to enhance predictive accuracy. This technique intends to investigate the extent to which sentiment data may operate as an extra feature in the stock prediction model. This study seeks to determine the most effective integration technique for depicting the link between market mood and stock movements by examining five scenarios. The following is the scenarios used in this study :

**X = Sentiment**

*1) **Scenario 1**:*

$$\{X_{t-1}, X_{t-2}, \ldots, X_{t-n}\} \tag{5}$$

Utilizing comprehensive historical sentiment data to identify sentiment trends.

*2) **Scenario 2**:*

$$\sum_{i=1}^{n} \frac{X_{t-i}}{2} \tag{6}$$

Employs a straightforward average of previous emotion values to encapsulate patterns.

*3) **Scenario 3**:*

$$X_{t-1} \tag{7}$$

Utilizes the latest sentiment value ($X_{t-1}$) as the primary indicator.

*4) **Scenario 4**:*

$$\sum_{i=1}^{n} X_{t-i} \tag{8}$$

Aggregates all prior sentiment levels to ascertain the cumulative impact.

*5) **Scenario 5**:*

$$\sum_{i=1}^{n} \frac{X_{t-i}}{e^i} \tag{9}$$

Implement exponential weighting to prioritize the most current sentiment.

### E. Model

Chen and Guestrin (2016) proposed XGBoost to perform predictions [10].XGBoost is a versatile and efficient machine learning method that excels in handling large datasets, managing sparse data, and providing robust performance in various applications, making it a popular choice among data scientists [10].Each tree in XGBoost focuses on learning the weaknesses of the data and improving the overall prediction [10]. In addition, the XGBoost model is popular because XGBoost is a model that has very good learning accuracy and efficiency [15].

The XGBoost model improves regulation on its learning target based on the traditional gradient boosting framework, the formula for XGBoost is as follows:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{10}$$

Formally $\hat{y}_i^{(t)}$ is the prediction of the $i$-th iteration at the $t$-th iteration [9].

### F. Performance Metrics

The evaluation of the XGBoost model in this study utilized two performance metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). RMSE calculates the square root of the average of the squared differences between the predicted and actual values [16]. the formula for RMSE is as follow:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

where $y_i$ represents the actual values, $\hat{y}_i$ represents the predicted values, and $n$ is the number of observations [17]. On the other hand, MAE measures the average of the absolute errors between the predicted and actual values, as given by the formula:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

where $y_i$ is the actual value, $\hat{y}_i$ is the predicted value, and $n$ is the number of observations [17].

RMSE places greater emphasis on larger errors, making it more sensitive to outliers. This is particularly useful in scenarios where large errors significantly impact analysis or decision-making, such as in financial contexts. In contrast, MAE treats all errors equally, making it less affected by the presence of outliers [18].Using both RMSE and MAE provides a comprehensive evaluation, as RMSE highlights the impact of larger errors while MAE gives an average error magnitude. Relying solely on RMSE may lead to biased analysis in datasets with outliers, as it amplifies the effect of large errors. Combining RMSE and MAE ensures a balanced and accurate assessment of model performance. In this study, XGBoost outperformed other methods, such as ARIMA, based on lower RMSE and MAE values [19].

## III. Result and Discussion

This study develops a data-driven analytical approach that integrates the Return variable as the primary feature. The Return variable is initialized in three forms: absolute return, relative return, and log return. The main objective of utilizing these three types of Return is to evaluate which form is the most effective in improving prediction accuracy. The Return variable not only provides an overview of stock value changes but also offers a unique perspective on the volatility and dynamics of the stock market.

This study incorporates sentiment analysis as a complementary element to the stock dataset to enhance prediction accuracy. Sentiment analysis is conducted using the ChatGPT LLM model, which can generate sentiment results with high sentiment score accuracy based on relevant texts, such as news datasets. The processed stock dataset is then merged with the sentiment dataset, creating an enriched dataset with more comprehensive information. This approach is designed to explore the impact of integrating sentiment analysis on prediction accuracy, particularly in different scenarios.

The dataset is divided into training and testing subsets to objectively evaluate the model's performance. Subsequently, hyperparameter tuning is performed to ensure the model operates efficiently and delivers optimal prediction accuracy. The hyperparameters used include n_estimators = 3000, max_depth = 7, min_child_weight = 2, and learning_rate= 0.48. These hyperparameters are selected based on previous experiments demonstrating that this combination achieves an optimal balance between accuracy and training efficiency.

After the modeling process, the evaluation results demonstrate that the integration of sentiment into the predictive model consistently improves prediction performance compared to models that do not incorporate sentiment. Table III shows the evaluation results for the five scenarios tested across three returns types: relative, absolute, and log.

TABLE IV
Evaluation Results for Relative Return Type

| Scenario | RMSE | MAE |
|---|---|---|
| No Sentiment | 0.008398 | 0.006492 |
| Scenario 1 | 0.00872 | 0.006612 |
| Scenario 2 | 0.008644 | 0.006667 |
| Scenario 3 | 0.009656 | 0.007325 |
| Scenario 4 | 0.008644 | 0.006667 |
| Scenario 5 | 0.008644 | 0.006667 |

TABLE V
Evaluation Results for Log Return Type

| Scenario | RMSE | MAE |
|---|---|---|
| No Sentiment | 0.009407 | 0.007348 |
| Scenario 1 | 0.009746 | 0.007 |
| Scenario 2 | 0.009163 | 0.007432 |
| Scenario 3 | 0.010973 | 0.008121 |
| Scenario 4 | 0.009163 | 0.007432 |
| Scenario 5 | 0.009163 | 0.007432 |

TABLE VI
Evaluation Results for Absolute Return Type

| Scenario | RMSE | MAE |
|---|---|---|
| No Sentiment | 65.684481 | 50.579228 |
| Scenario 1 | 67.679263 | 51.985494 |
| Scenario 2 | 66.105451 | 50.616032 |
| Scenario 3 | 70.792413 | 54.523370 |
| Scenario 4 | 66.105451 | 50.616032 |
| Scenario 5 | 66.142305 | 50.616032 |

Table III presents the evaluation results, indicating that the model without sentiment delivers the most optimal performance compared to the other scenarios, including those that incorporate sentiment. For the No Sentiment scenario, the RMSE is 0.009407 and the MAE is 0.007348. Table IV shows that Scenario 2, 4, and 5 exhibit better performance compared to the other scenarios, with Scenario 2, 4, and 5 having an RMSE of 0.009163 and MAE of 0.007432. Scenario 3 has the highest evaluation results in the Log return type. Table V indicates that No Sentiment provides the most favorable evaluation results with an RMSE of 65.684481 and MAE of 50.579228, which are lower than those of the other scenarios.

Among the three returns types relative, absolute, and log, the log return type was found to be the most optimal for predictive performance. Scenario 2, 4, and 5, which incorporate the log return type, demonstrated superior results with an RMSE of 0.009163 and MAE of 0.007432, which

are lower than those of the other scenarios. A detailed comparison was conducted between Scenario 2 and the model without sentiment features, highlighting the significant role of sentiment in enhancing the prediction accuracy of the IHSG.

To provide a clearer understanding of the impact of sentiment in modeling, two plots are presented comparing the prediction results without sentiment integration and with sentiment.

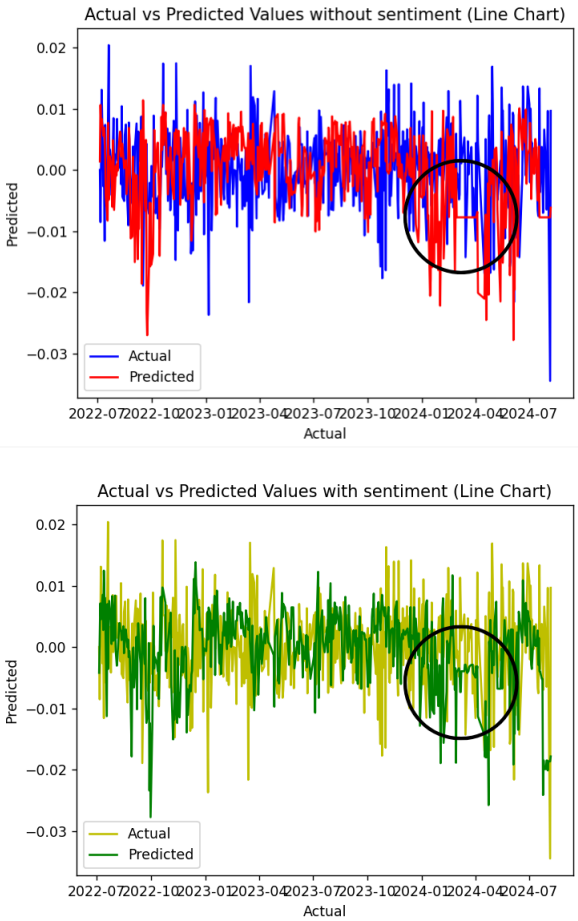


Fig. 3. Log Return Predictions

Based on Fig. 3. it is evident that the model incorporating sentiment (the bottom plot) demonstrates better performance compared to the model without sentiment integration (the top plot). This indicates that adding sentiment as a feature to the model contributes significantly to improving prediction accuracy. Although the predictive results in the bottom plot are not entirely perfect, the inclusion of this feature enables the model to capture more complex and dynamic market patterns, resulting in more accurate outcomes compared to a model relying solely on return data.

Table VI presents a comparative analysis of performance metrics between Random Forest and XGBoost models. The comparison employs evaluation metrics of RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) to demonstrate the predictive capabilities of both models.

TABLE VII
COMPARISON OF RMSE AND MAE FOR MODELS WITH AND WITHOUT SENTIMENT (LOG RETURN TYPE)

| Model | RMSE | MAE |
|---|---|---|
| XGBoost (No Sentiment) | 0.009407 | 0.007348 |
| XGBoost (Sentiment) | 0.009163 | 0.007432 |
| Random Forest (No Sentiment) | 0.011421 | 0.007593 |
| Random Forest (Sentiment) | 0.010821 | 0.007327 |

## IV. CONCLUSION

This study successfully integrates sentiment analysis with a stock prediction model based on the XGBoost algorithm to forecast the movement of HSG. By combining historical stock data with sentiment analysis derived from financial news, this research identifies complex patterns within stock market dynamics. The study employs three types of returns (absolute, relative, and logarithmic) and five sentiment scenarios to evaluate the contribution of sentiment features to prediction performance. The evaluation results consistently show that models incorporating sentiment demonstrate improved performance compared to those without sentiment integration.

From the experiments, the log return type was found to be the most optimal for enhancing prediction accuracy. Scenario 2, 4, and 5, which employ the log return type, produced the best evaluation results with an RMSE value of 0.009163 and an MAE value of 0.007432, the lowest among all scenarios. The addition of sentiment data also enables the model to capture more dynamic market movement patterns, as observed in the comparison between models with and without sentiment integration. Further analysis reveals that although the predictions are not entirely perfect, integrating sentiment data significantly enriches the dataset, allowing the model to become more optimal and adaptive to rapid market changes.

Consequently, this study underscores the critical role of market sentiment in stock price prediction, particularly when combined with predictive algorithms such as XGBoost. The findings contribute to the development of more accurate financial prediction models and provide insights into the application of technologies like NLP and LLMs for analyzing stock markets.

Future research can focus on developing models that utilize alternative data sources, such as social media or corporate financial reports, to enrich sentiment datasets. Additionally, exploring hybrid or ensemble methods that combine XGBoost with other algorithms, such as LSTM or Transformer, could enhance the model's ability to capture complex patterns in the stock market. Further development may also involve integrating global data to analyze the impact of international sentiment on local markets. As a practical application, this research could be implemented in the form of a web-based system or application to enable real-time stock market predictions.

## REFERENCES

[1] T. S. J. Wijaya and S. Agustin, "Faktor-faktor yang mempengaruhi nilai ihsg yg terdaftar di bursa efek indonesia," *Jurnal Ilmu dan Riset Manajemen (JIRM)*, vol. 4, no. 6, 2015.

[2] D. Arista and A. Astohar, "Analisis faktor–faktor yang mempengaruhi return saham," *Jurnal Ilmu Manajemen dan Akuntansi Terapan (JIMAT)*, vol. 3, no. 1, 2012.

[3] J. Wang, "The analysis of the financial market in china," *Academic Journal of Business & Management*, 2021.

[4] C. Wang, "Practical significance of distinguishment between systematic/non-systematic risks," *BCP Business & Management*, 2023.

[5] S. Chen, "The differences between systematic and non-systematic risk and alternative approaches to understanding risk," *Journal of Education, Humanities and Social Sciences*, 2023.

[6] V. Vongphachanh and K. Ibrahim, "The effect of financial variables on systematic risk in six industries in thailand," *Asian Business Consortium Journal of Applied Research*, vol. 9, pp. 63–68, 2020.

[7] B. Jange, "Prediksi indeks harga saham gabungan (ihsg) menggunakan prophet," *JOTIKA Journal in Management and Entrepreneurship*, vol. 1, no. 2, pp. 53–59, 2022.

[8] T. Ouyang, H.-Q. Nguyen-Son, H. Nguyen, I. Echizen, and Y. Seo, "Quality assurance of a gpt-based sentiment analysis system: Adversarial review data generation and detection," *2023 30th Asia-Pacific Software Engineering Conference (APSEC)*, pp. 450–457, 2023.

[9] B. Lefort, E. Benhamou, J.-J. Ohana, D. Saltiel, B. Guez, and D. Challet, "Can chatgpt compute trustworthy sentiment scores from bloomberg market wraps?" *Social Science Research Network*, Jan 2024.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[11] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, and E. Salwana, "Deep learning for stock market prediction," *Entropy*, vol. 22, no. 8, p. 840, 2020.

[12] A. Ultsch, "Is log ratio a good value for measuring return in stock investments?" in *Advances in Data Analysis, Data Handling and Business Intelligence*, A. Fink, B. Lausen, W. Seidel, and A. Ultsch, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 505–511.

[13] V. Padmanayana and K. Bhavya, "Stock market prediction using twitter sentiment analysis," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, vol. 7, no. 4, pp. 265–270, 2021.

[14] B. Lefort, E. Benhamou, J.-J. Ohana, D. Saltiel, B. Guez, and D. Challet, "Can chatgpt compute trustworthy sentiment scores from bloomberg market wraps?" *Social Science Research Network*, Jan 2024.

[15] L. Xie, J. Liu, S. Lu, T.-H. Chang, and Q. Shi, "An efficient learning framework for federated xgboost using secret sharing and distributed optimization," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 5, pp. 1–28, 2022.

[16] M. Sharaf, E. E. Hemdan, A. El-Sayed, and N. A. El-Bahnasawy, "Stockpred: a framework for stock price prediction," *Multimedia Tools and Applications*, vol. 80, pp. 17 923 – 17 954, 2021.

[17] T. Chai and R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, pp. 1247–1250, 2014.

[18] W. Dong, Y. Huang, B. Lehane, and G. Ma, "Xgboost algorithm-based prediction of concrete electrical resistivity for structural health monitoring," *Automation in Construction*, vol. 114, p. 103155, 2020.

[19] M. Noorunnahar, A. Chowdhury, and F. A. Mila, "A tree based extreme gradient boosting (xgboost) machine learning model to forecast the annual rice production in bangladesh," *PLOS ONE*, vol. 18, 2023.