Innovative Sentiment Analysis and Prediction of Stock Price Using FinBERT, GPT-4, and Logistic Regression: A Data-Driven Approach

Olamilekan Shobayo * , Sidikat Adeyemi-Ionge , Olusogo Popoola , Bayode Ogunleye

Posted Date: 13 September 2024

doi: 10.20944/preprints202409.1089.v1

Keywords: FinBERT Model; Logistic Regression; FinBERT; Optuna; Timeseries cross validation

Article

Innovative Sentiment Analysis and Prediction of Stock Price Using FinBERT, GPT-4, and Logistic Regression: A Data-Driven Approach

Olamilekan Shobayo [1,*], Sidikat Adeyemi-Longe [1], Olusogo Popoola [1] and Bayode Ogunleye [2]

[1] School of Computing and Digital Technologies, Sheffield Hallam University, Sheffield, S1 2NU, UK
[2] Department of Computing & Mathematics, University of Brighton, Brighton BN2 4GJ, UK
* Correspondence: o.shobayo@shu.ac.uk

Abstract: This study explores the comparative performance of FinBERT, GPT-4, and Logistic Regression for sentiment analysis and stock index prediction using the NGX All-Share Index dataset. By leveraging advanced language models like GPT-4 and FinBERT, alongside a traditional machine learning model, Logistic Regression, we aim to classify market sentiment, generate sentiment score, and predict market price movements. The models were assessed using metrics such as accuracy, precision, recall, F1 score, and ROC AUC. Results indicate that Logistic Regression outperformed both FinBERT and GPT-4, with an accuracy of 81.83% and a ROC AUC of 89.76%. GPT-4 predefined approach exhibited a lower accuracy of 54.19% but demonstrated strong potential in handling complex data. FinBERT, while offering more sophisticated analysis, was computationally demanding and yielded a moderate performance. Hyperparameter optimization using Optuna and cross-validation techniques ensured the robustness of the models. This study highlights the strengths and limitations of these approaches in stock market prediction and presents Logistic Regression as the most efficient model for this task, with FinBERT and GPT-4 showing promise for future exploration.

Keywords: FinBERT model; logistic regression; FinBERT; Optuna; timeseries cross validation

1. Introduction

The prediction of stock market movements has been a focal point for researchers and investors due to the financial market's complexity and volatility. The ability to make precise stock or market predictions can result in improved decision-making, reduction of risks, and increased profitability. Traditional statistical techniques often fail to identify the complex patterns in stock data, most especially when affected by external variables like news and market sentiment. Recent advancements in machine learning and deep learning have provided more sophisticated tools to address this problem. Sentiment analysis has become a crucial tool in stock market prediction, as investor sentiment significantly impacts market trends. Natural Language Processing (NLP) models, such as FinBERT and GPT-4, have proven to be effective in analysing unstructured textual data like financial news headlines [1]. These models can classify sentiment (e.g., positive, negative, neutral) and predict how such sentiment may influence stock prices. Moreover, simple and classic models like Logistic Regression remain effective for sentiment classification and market prediction when adjusted properly.

In this study, we aim to evaluate the performance of FinBERT, GPT-4, and Logistic Regression in predicting the NGX All-Share Index (NGX). Each model was trained on historical NGX data labels and financial news, utilizing various sentiment analysis techniques in the process. The labeled data allows the models to associate input news features with specific outcomes, which leads to better prediction accuracy when analyzing news articles for tasks like sentiment analysis or topic classification [2]. Five key metrics i.e. Accuracy, Precision, Recall, F1 Score, and ROC AUC were to assess the performance of each model. Previous research have demonstrated the potential of FinBERT, a financial-domain-specific model, in understanding financial terminology and context

 doi:10.20944/preprints202409.1089.v1

with remarkable precision [3]. However, its resource-intensive nature can present challenges in terms of computational efficiency. GPT-4, a versatile language model, has remarkable capabilities in understanding and generating human-like text. This makes it ideal for processing unstructured news data. Nevertheless, the exploration of its predetermined and heuristic approach may restrict its precision in particular financial situations. Logistic Regression is simple, computationally efficient, and produces reliable results when properly optimized. The results show that Logistic Regression outperformed both FinBERT and GPT-4 across most metrics. Logistic Regression achieved the highest accuracy (81.83%) and ROC AUC (89.76%), while FinBERT and GPT-4, lagged behind in predictive accuracy. This highlights the effectiveness of traditional models when properly tuned, and the potential for hybrid approaches combining the strengths of all three methods.

To provide a clear understanding of the current state of research and position this study within the broad academic context, a comprehensive Literature Review was developed (Table 1). This table summarizes the different data analysis methods employed in previous stock price prediction studies, along with their respective strengths and weaknesses. Furthermore, the performance of the each model will be compared with existing approaches to demonstrate their ability to gauge the market's mood and predict stock price movements. Positive sentiment indicate rising stock prices, while negative sentiment signal potential declines, which makes it a useful tool in stock market prediction and decision-making

Table 1. Literature Review.

Literature Info	Data Processing Methods	Significance	Contrast References	Our Work
Liu Z. et al. [3]	FinBERT for Financial Text Mining	Enhanced financial sentiment analysis with domain-specific FinBERT model	Limit to only FinBERT for sentiment analysis	Associate FinBERT and other deep and machine learning models for better analysis.
Leippold M. [4]	GPT-3 for Financial Sentiment Analysis	Explored adversarial attacks on financial sentiment predictions	Limited transparency of GPT-3's decision-making	Compares GPT-4 with financial news data features to improve interpretability.
Yang J. et al. [5]	LASSO-LSTM with FinBERT	Combined technical indicators with FinBERT for stock predictions.	Feature extraction limitation	Feature extraction improvement using NGX label and financial news for stock market prediction & accuracy.
Sidogi T. et al. [6]	FinBERT with LSTM for Stock Price Prediction	Utilized FinBERT for financial sentiment analysis and LSTM for	Limit performance metrics to Root Mean Square Error (RMSE) and	Addition of more evaluation metrics and other deep learning GPT 4

 doi:10.20944/preprints202409.1089.v1

	predicting stock price movements based on sentiment.	mean absolute error (MAE) only	and classic machine learning for better perspective

Liu, Z developed a pre-trained FinBERT model for financial text mining and sentiment analysis. Their research demonstrated that FinBERT significantly outperformed other models in understanding financial language, providing more accurate sentiment classifications in financial reports and news [3]. However, the study was limited by its focus on FinBERT only without evaluating the performance with other models. This study builds upon their findings by applying and evaluating FinBERT with other AI models to ascertain its utility and performance on financial news text. Leippold, M explored the vulnerabilities of financial sentiment models to adversarial attacks on GPT-3. The research revealed that subtle manipulations in financial texts could alter sentiment predictions, which highlight GPT-3's sensitivity to adversarial inputs. Although GPT-3 showed great potential in financial text generation, its lack of interpretability remains a significant limitation [4]. The research extends this work by integrating explainable AI methods alongside GPT-4 to improve transparency in financial sentiment and predictions. This offers a more robust approach to understanding model decision-making. Yang, J combined LASSO, LSTM, and FinBERT to predict stock price direction using technical indicators and sentiment analysis [5]. Their model achieved high accuracy in predicting price movements based on market sentiment. However, their approach was limited by the feature extraction techniques employed, which may not fully capture non-linear relationships in financial data. This study addresses feature extraction issue by extracting strictly financial news and adding NGX labels for easy topic classification as part of input features. Sidogi, T used FinBERT and LSTM to analyze the impact of financial sentiment on stock prices. Their study focused on using LSTM for time series forecasting, with FinBERT providing sentiment features from financial news and reports. The study limited performance metrics to Root Mean Square Error (RMSE) and mean absolute error (MAE) only without evaluating the performance of FinBERT itself [6]. This research intend to evaluate all selected models to ascertain each models performance for better perspectives

2. Materials and Methods

This study uses News Headlines sentiment data scraped from Nairametric and Proshare websites using Data Miner to analyse the performance and direction of the Nigerian stock market. Nairametric is a Nigerian investment advocacy company, while Proshare is a professional practice firm that offers various services to connect investors and markets. The aim of using two organisations' news was to ensure accuracy, reduce errors, and boost trust. News headlines were chosen over social media for sentiment analysis due to their credibility, structured data, reduced bias, event-centricity, manageable volume, less manipulation, reliability, timeliness, source verification, journalistic standards, and focused content. The data cover from January 4, 2010 to June 7, 2024, with 3,573 observations. News labels are based on the stock index categorization, where:

"Class 1" implies daily share price gain and

"Class 0" signifies unchanged or fall in share price.

2.1. News Data Pre-Processing and Preparation

The News headlines were pre-processed using the Natural Language Toolkit (NLTK). The NLTK is a Python library used for dataset cleansing and natural language processing (NLP). The data cleaning methods include stopwords elimination, data conversion, concatenation, tokenization, noise abatement, normalization, and feature extraction.

- Stopwords (high-frequency words with limited semantic meaning) are re moved to improve accuracy.
- Data conversion converts text to lowercase letters,

 doi:10.20944/preprints202409.1089.v1

- Concatenation joins text strings for better feature engineering and financial data preparation.
- Tokenization breaks text into manageable tokens, which enhances learning model performance
- Noise abatement removes random or unnecessary data that masks market trends and reduces data analysis and machine learning model precision
- Normalization is a process that standardises text to improve speed and quality of text analysis. Stemming and lemmatisation are methods used to standardize words by removing suffixes and affixes to reveal their root form. Stemming algorithms use heuristic principles for efficiency and simplicity. Heuristic principles use pattern matching, rule-based simplifications, fixed-order operations, search space reduction, and statistical heuristics to guide problem-solving and decision-making [7]. Lemmatisation analyses context and grammatical components to generate a lemma (root word), which improves text analysis, accuracy, and clarity through contextual comprehension [8].
- Feature extraction is a method that transforms data into features for machine and deep learning algorithms [9]. It improves data interpretability, model performance, and dimensionality. BERT embeddings was used for BERT due to their contextual comprehension, transfer learning potential, and resilience in false news identification and sentiment analysis, while TF-IDF vectorisation was used for logistic regression as it performs, interprets, and efficiently handles sparse and dimensional data like news data.

The News headlines datasets were splitted into 85% for training and validation, and 15% for testing. The dataset is a chronological dataset and its temporal order was maintained throughout the model training and evaluation. It also used timeseriessplit cross validation (TSCV) to train and validate the model with folds (n=5) on the news dataset. This method maintains the time dependency between data points. TSCV closely mirrors how the model would perform in real-world applications, such as stock price prediction by simulating real-world scenarios with unseen future data in each fold. This approach enhances the model's ability to generalize, reduces overfitting and improves prediction accuracy [10]. Moreover, it avoids data leakage by ensuring that no future information influences the training phase, which results in a more reliable evaluation of the model's performance. This method uses all available data points across the folds for both training and validation, providing a comprehensive assessment of the model's robustness and accuracy. Besides, data label based on the stock index categorization were added to news dataset as input feature into the model. The labels help the model to differentiate news categories, reduce noise, and enable efficient model training through clear mappings between input features and the desired outcomes [11]

2.2. Algorithm Selection and Computation for Financial News

The study used FinBERT, GPT 4, and Logistic Regression model to find the optimal method because of their ability to handle complex language, domain-specific text, and provide interpretable results. These techniques assisted in understanding model complexity, performance, interpretability trade-offs, and enables informed algorithm selection for hybrid model use cases.

2.3. FinBERT Architecture, Development and Training

FinBERT, a specialized variant of BERT, is a pre-trained language model designed for financial text analysis, interpreting and analysing nuances of financial language, including finance and economics-specific jargon. It excels in financial sentiment analysis, market sentiment research, stock trading strategy formulation, and risk management. FinBERT-Base model (12 layers, 768 hidden size, 12 attention heads) and FinBERT-Large (24 layers, 1024 hidden size, 16 attention heads) are the variants but the study used the FinBERT-base because of its computational efficiency, lower memory requirement, sufficient performance, overfitting concerns and ease of use. FinBERT-base model training process started with ascertaining the data integrity through correct parsing of the date fields of the cleaned news data to prevent errors in temporal analyses. The cleaned data were tokenised using input IDs and attention masks, and converted into PyTorch tensors for data feeding into the model. Dataloaders were created for batching data during training, validation, and testing. News data was divided into training & validation, and test sets chronologically. Hyperparameter tuning

 doi:10.20944/preprints202409.1089.v1

was performed using Optuna, while automatic mixed precision (AMP) training was employed to fast-track the training process while maintaining model accuracy. Also, early stopping with a patience parameter of 5 was applied to prevent overfitting and preserve the model's generalization capabilities. The final trained model was evaluated with classification metrics on the test dataset.

2.4. GPT 4

GPT predefined approach was also used to classify, analyse and generate sentiment score on news data. The end to end processing of GPT is efficient as it analyses news data from unprocessed headlines to internal representations such as sending news data to GPT4 API, GPT4 processes text, analyses sentiment of news text, and ended with sentiment scores and output for evaluation.

2.5. Logistic Regression Architecture Development and Training:

Logistic regression is a statistical model that is designed for solving binary classification problems. The architecture of the logistic regression (LR) used for this study is defined by its key parameter components and hyperparameters. LR binary classification fits into listed stock labels, with "Class 1" representing daily share price gain and "Class 0" signifying price decline or unchanged price. This assisted in predicting the financial news input's class. The model architecture core is the 'C' parameter, penalty, and solver. The value of C balances fitting training data well and keeping the model simple to minimise overfitting. A smaller C discourages large coefficients, strengthening regularisation, whereas a bigger C weakens it. Also, L2 regularisation (penalty='l2') kept model coefficients minimal to prevent overfitting. 'Liblinear' was chosen over 'lbfgs' and 'saga' due to its resilience and speed. Liblinear optimises small datasets quickly and reliably. It also enables rapid computation for L2 penalised logistic regression. Sigmoid function, another key logistic regression component, translates all real numbers to values between 0–1. It used scikit-learn to create sentiment score and 0.5 as a decision threshold to classify inputs. This is based on logistic function probability and the mathematical formula for Sigmoid is:

$$\text{Sigmoid}\,(z) = \frac{1}{1+e^{-z}} \tag{1}$$

where:
z = the weighted sum of the input features
e = mathematical constant (~ 2.71828).
−z = negative of the input z

The LR model was trained using Optuna to automatically optimise "C" and solver hyperparameters. The objective function was defined and parameter ranges were suggested to maximise the F1 score: C: 0.0001 – 100; Solver: Liblinear, lbfgs; and Penalty: L2. Optuna explored multiple C values and tested various solvers to identify the optimal combination that yields the best F1 score on the validation data. Additionally, timeseries cross validation with n=5 was implemented to stabilise the selected hyperparameters across several time-based folds. The model was trained on the training dataset and evaluated on the validation dataset to calculate the F1 score. F1 was selected over other classification metric because of its suitability in scenarios with class imbalance. This process was repeated for each set of hyperparameters suggested by Optuna. The optimal hyperparameters were chosen and the LR model was retrained on the training and validation sets, and tested on a news testing set. These methods provide a reliable method for developing, optimising, and training this study's LR model.

3. Results

3.1. FinBERT

The optimal hyperparameter suggested by Optuna was used to train the FinBERT model on a chronological order and tested on 15% test dataset. The evaluation result is tabularised below:

 doi:10.20944/preprints202409.1089.v1

Table 2. Evaluation Result of FinBERT.

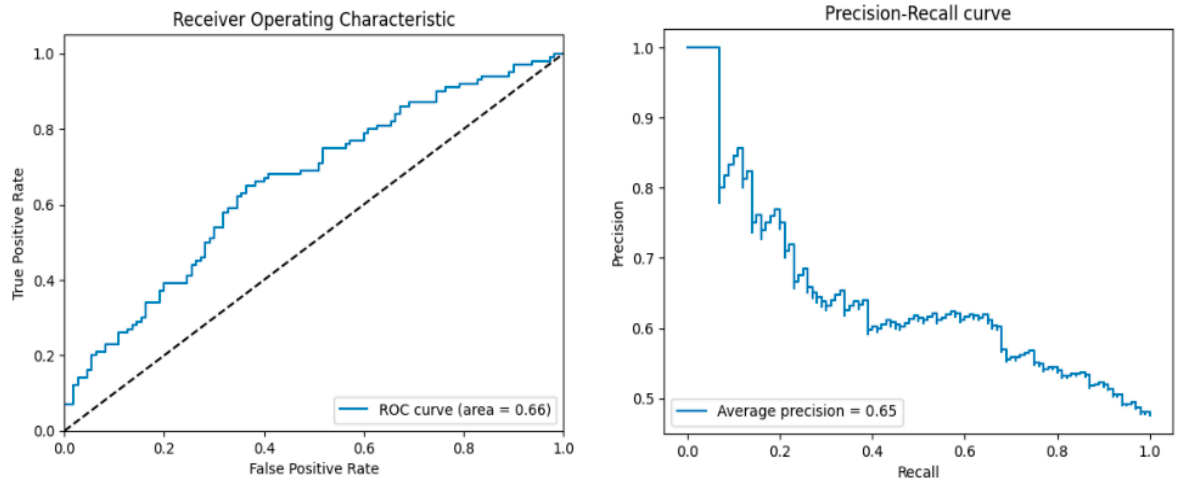
Best hyperparameters: learning_rate: 3.564937469182303e-05, batch_size': 16		
Test set metrics		**Test Metrics (Percentage)**
Accuracy	0.6333	63.33
Precision	0.6376	63.76
Test Recall	0.6333	63.33
Test F1 Score	0.6330	63.30
Test ROC AUC	0.6559	65.59

3.1.1. Model Evaluation

The result of the evaluation metric above shows that FinBERT correctly predicts the sentiment of financial news 63.33% of the time. This shows moderate performance. The complexity and fluctuating nature of financial terms and market sentiments can make achieving a high level of accuracy to be tough. The precision score shows when FinBERT predicts the news sentiment, it is 63.76% correct of the time. This performance is also modest. The prediction precision is very important as false positive sentiment can lead to incorrect market trading decisions. The recall result shows FinBERT has been able to identify 63.33% of all relevant instances of sentiment. This implies the model is moderately efficient as it might still miss some relevant market signal and sentiments in the financial data. The F1 score is a balance between precision and recall, and the score at 63.30% is fairly balanced but not strong enough. The ROC AUC score of 65.59% is a positive indicator of the model ability to differentiate between positive and negative sentiment. A higher ROC AUC value close to higher than 0.7 is considered good in complex field like financial news interpretation as it indicates better ability to distinguish between positive and negative sentiment. Chen, T empirically demonstrates benchmarking scores of existing methods and discusses specific models designed for financial news sentiment analysis [12]. The findings suggested that well-performing models in complex financial sentiment analysis often achieve ROC AUC scores close to 0.7 or higher. This confirms the research claim that FinBERT does moderately well in predicting financial news. Overall, the FinBERT prediction performance is within range given in to consideration the complexity of financial news and the fact that sentiments are hidden in technical language and also influenced by context. The studies of Kirtac, K.K and Varghese, R.R emphasize that while models like FinBERT can perform well on financial data; they often require specific adaptations to the financial datasets they analyse [13, 14]. The FinBERT scores indicate an effective but not highly reliable model for critical financial decisions.

3.1.2. Visual Inspection

The AUC of 0.66 of ROC below shows FinBERT's accuracy is modest. The precision versus recall curve also an average precision score of 0.65.



Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2024 doi:10.20944/preprints202409.1089.v1

Figure 1. (a) ROC Curve and (b) Precision-Recall Curve of FinBERT Model.

The daily sentiment score plot below shows varying trend in the news data while the distribution of predicted probabilities for each class, Class 0 vs. Class 1, indicates some uncertainty in predictions. This suggests that while the model distinguishes between classes, there is room for improvement in its confidence

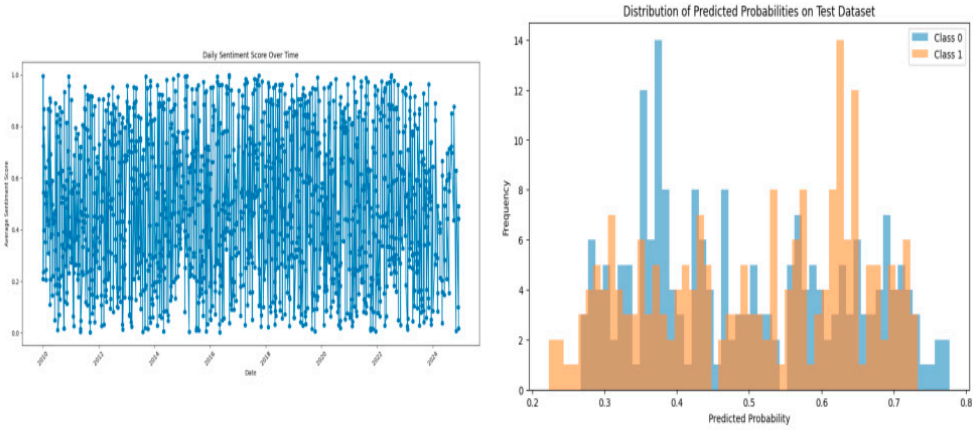


Figure 2. (a) Sentiment Score over Time and (b) Distribution of Predicted Probabilities of FinBERT Model.

Overall, the FinBERT model shows moderate effectiveness in classification, with some uncertainty in predictions. The sentiment score fluctuations provide insights into market sentiment over time, which can be valuable for trading decisions. Both the ROC and Precision-Recall curves suggest that while the model performs above random chance, there is potential for further refinement to improve its accuracy and reliability.

3.2. GPT

The findings reveals that GPT 4 can perform sentiment analysis and evaluation using classic machine learning models such as predefined approach, Naïve Bayes, linear regression, etc. However, this study explores the predefined sentiment approach of GPT. The process started with uploading the News headline csv file with the instruction of using GPT to classify, evaluate and perform sentiment analysis on the uploaded data as shown in figure below

Financial Data

	Date	label	text
1	07/06/2024	1	world bank mandates hire security consultant billion loan project nigeria despite
2	06/06/2024	0	afrexim bank disburse another million nnpc billion loan n trillion provision fuel
3	05/06/2024	1	executive order tinubu stop

use chat GPT4 to classify, and do sentiment analysis on the uploaded

Figure 3. The Financial News Data Input into GPT.

The response from GPT itemised the steps of executing the instruction as data loading, text inspecting, text data pre-processing using "re" library for predefined approach, splitting the data into

 doi:10.20944/preprints202409.1089.v1

training (2501 samples), validation (536 samples) and testing (536 samples), performing sentiment analysis, classification, and evaluating the model on the validation and test sets. The predefined approach based the evaluation on predefined sentiment labels.

3.2.1. Model Evaluation:

The table below shows the result of the evaluation metrics using GPT 4.

Table 3. Evaluation Metrics Result of GPT.

Evaluation Metrics	Predefined Sentiment Set Metrics	Predefined Sentiment Set Metrics (Percentage)
Validation Accuracy:	0.5720	57.20
Test Evaluation		
Accuracy	0.5419	54.19
Precision	0.7266	72.64
F1 Score	0.4509	45.09
Recall (Sensitivity):	0.3269	32.69
AUC-ROC	0.6537	65.37

The accuracy of 57.20% on the validation set indicates that the model correctly classifies sentiment more than half of the time. This shows moderate performance and suggests room for improvement in capturing the nuances of sentiment in financial news. The test accuracy of 54.19% is slightly lower than the validation accuracy. This suggests that the model may struggle to generalize on unseen data. This could be due to inherent complexity of financial sentiment. The test precision of 72.66% is relatively high. This shows that when the model predicts a positive sentiment, it is often correct. The news with positive sentiment that has been selected is expected to genuinely reflect confidence about the state of the stock market. The low test recall of 32.69% suggests that the model misses many actual positive sentiment news. This indicates that many potential significant positive news items might not be recognized. This leads to underrepresentation of positive sentiment. The F1 Score of 45.09% is a balanced measure that shows the overall performance. It demonstrates the difference resulting from increased accuracy and decreased completeness. The Area Under the Curve - Receiver Operating Characteristic. (AUC-ROC) of 65.37% is moderate and it suggests the model has a fair ability to differentiate between positive and negative sentiments. This is important for predicting the impact of news on stock movements. The predefined approach of GPT can reliably identify positive sentiments when they occur. This can be useful for stockbrokers or investors that are focusing on signals for bullish market conditions. However, the low recall means many positive opportunities might be missed. This potentially may lead to conservative trading strategies. The moderate AUC-ROC shows the predefined approach can capture some trend but many might go unnoticed. This affects market prediction accuracy. Overall, the model performance suggests it should not be solely relied upon for stock trading decisions.

3.2.2. Visual Inspection

The ROC Curve below, with an AUC of 65.37%, shows moderate ability to differentiate between positive and negative market sentiments. However, it cannot be compared to the precision-recall curve, which fluctuates significantly in middle recall values. The precision-recall curve below starts high but drops as recall increases, indicating a trade-off between capturing more positive sentiments at the expense of accuracy. Senapaty, MK emphasise the importance of considering both ROC and Precision-recall curves to understand the trade-offs between sensitivity, specificity, and precision in practical applications [15].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2024 doi:10.20944/preprints202409.1089.v1

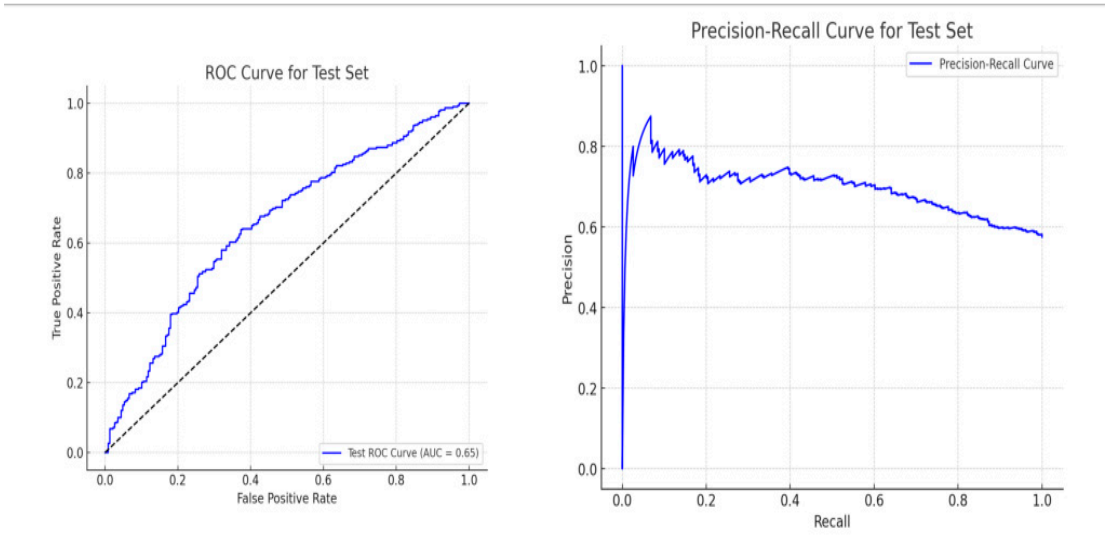


Figure 4. (a) ROC Curve and (b) Precision-Recall Curve of GPT.

The GPT model predicts stable daily sentiment scores, with a high number of positive sentiments on most days. This pattern provides insights into market sentiment trends and potential impacts on stock prices. The predicted probabilities are clustered around 0.5 thresholds, with a central tendency of 0.45 to 0.55 ranges. The model's low confidence level indicates uncertainty in classifying financial news as positive (close to 1) or negative (close to 0) sentiment in numerous instances.

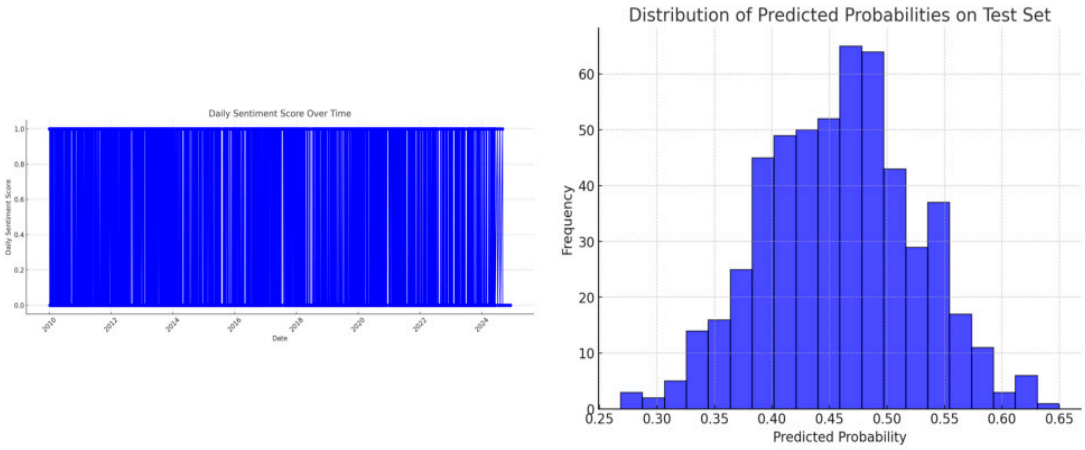


Figure 5. (a) Sentiment Score over Time and (b) Distribution of Predicted Probabilities of GPT.

3.3. Logistic Regression

The study used logistic regression to analyse sentiment in financial news headlines. The model was optimized using a logarithmic scale, with a fixed penalty type of 'l2'. The model outputs were used to derive sentiment scores, which indicate the positive or negative sentiment expressed in each headline. The optimal hyperparameter from Optuna was used for model training and testing. The daily sentiment score was generated and the evaluation result is tabularized below:

Table 4. Evaluation Result of Logistic Regression.

Best hyperparameters: {'C': 3.037005064126959, 'solver': 'liblinear', 'penalty': 'l2'}		
Training metrics Accuracy: 0.8093 = 80.93%		
Test set metrics		Metric %
Accuracy	0.8183	81.83

 doi:10.20944/preprints202409.1089.v1

Precision	0.8257	82.57
Test Recall	0.8115	81.15
Test F1 Score	0.8185	81.85
Test ROC AUC	0.8976	89.76

3.3.1. Model Evaluation

The test results from using logistic regression on news data show accuracy of 81.83%. This clearly shows a solid performance of distinguishing between positive and negative sentiments on financial news. Besides, the model test accuracy is very close to the training accuracy of 80.93% which indicates that the model generalizes well and there is no significant overfitting. The high precision of 82.57% implies that positive news is well-identified and out of all the positive predictions made by the model, 82.57% were actually positive. This shows the model is quite good when it predicts positive sentiment and has a relatively low rate of false positives. This is very important for investors that rely on positive market signal to make buying decision. The recall of 81.15% shows that out of all the actual positive cases, 81.15% were correctly identified by the model. This shows the model is doing well at capturing most of the actual positives. Although, a slight drop from precision indicates a small trade-off between precision and recall. For stock market applications, this suggests potential for investment opportunities. The ROC AUC score of 89.76% shows the model's ability to distinguish between the positive and negative classes. A score closer to 100% is ideal. The result show excellent performance and indicates that the model is very good at ranking positive cases higher than negative cases. This is especially useful in scenarios where you might want to adjust the decision threshold for different costs of false positives and false negatives. The F1 Score of 81.85% is the harmonic mean of precision and recall that balances the two metrics. A value of 81.85% shows a good balance between precision and recall. This suggests a well-balanced model that is effectively managing both false positives and false negatives. The model generalizes well. This is shown with the close alignment of training and test accuracy. Besides, there is a good balance between precision and recall. This means that the model is well-suited for stock market sentiment where both false positives and false negatives are costly. Overall, the model is robust in distinguishing between the classes. This makes it reliable for the study.

3.3.2. Visual Inspection

The ROC curve and Precision Recall curve below shows the performance of logistic regression model on the financial news.

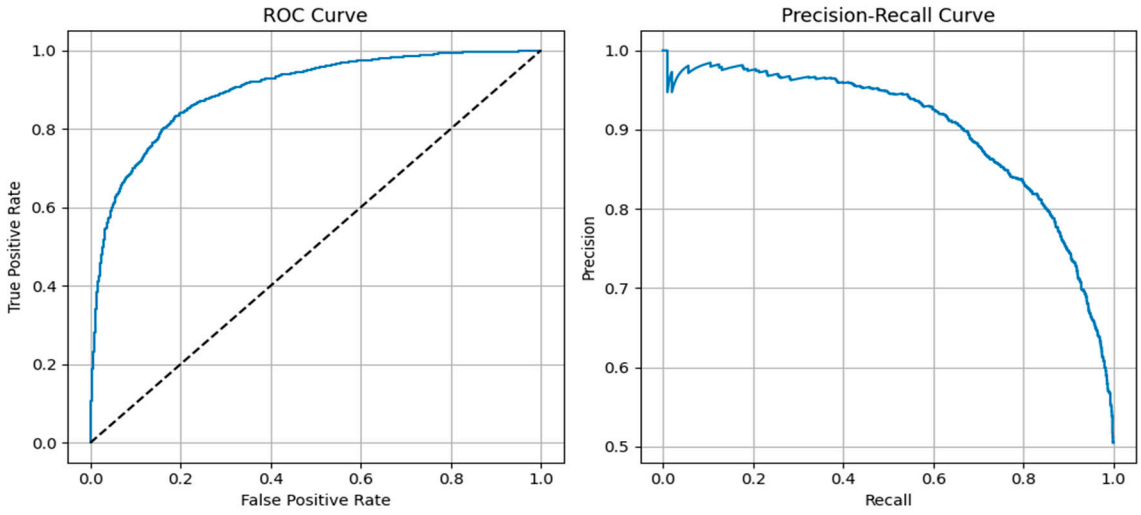


Figure 6. (a) ROC Curve and (b) Precision-Recall Curve of Logistic Regression Model.

The ROC curve above the diagonal indicates superior performance compared to random guessing. The elevated AUC value indicates robust model performance. The model maximises the true positive rate and maintains a low false positive rate. Logistic regression effectively separates

 doi:10.20944/preprints202409.1089.v1

between positive and negative classes, as observed in the ROC curve surpassing the diagonal line. The Precision-Recall Curve shows a positive correlation between precision and recall, with a decrease in precision as recall increases. Nevertheless, the model maintains high accuracy across different recall values, which indicates its performance across different thresholds. Furthermore, the plots below are for the daily sentiment score over time and the distribution of predicted probabilities on test set.

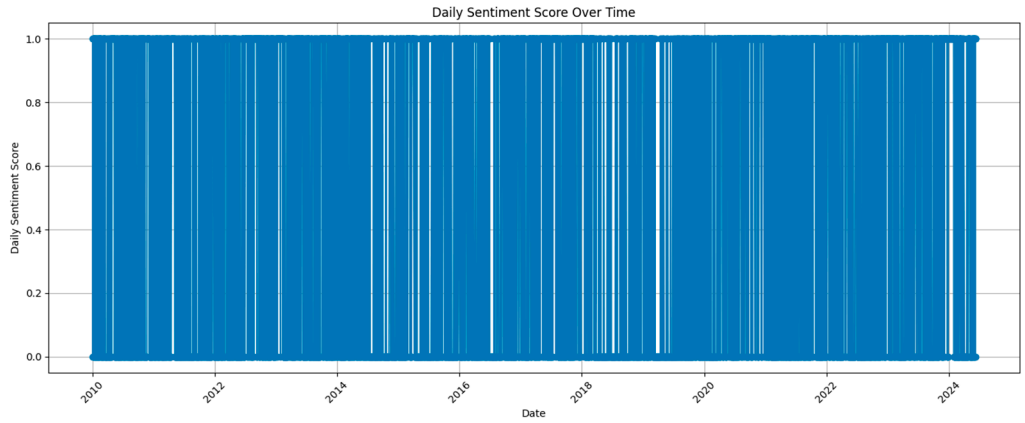


Figure 7. Sentiment Score over Time.

The NGX market's sentiment scores show significant volatility over time, with frequent fluctuations between 0 and 1. This indicates a high sensitivity to financial news and other exogenous variables. The repeated extreme fluctuations suggest a strong reaction to news. The understanding of these fluctuations is necessary for traders and investors, as sentiment-based trading algorithms can capitalize on the changes. Nevertheless, models must also be resilient to unforeseen sentiment fluctuations.

The histogram of predicted probabilities below shows an even distribution and a wide spread across the 0 to 1 range. This indicates uncertainty in market sentiment. It also suggests the market is often in flux, reacts to various factors, and often predict price movements with caution or moderate confidence. The sentiment near 0.5 suggests potential for either direction depending on subsequent financial news. This simply suggests adding more nuanced strategies like technical indicators and historical trends to account for market uncertainty, rather than relying solely on news sentiment.

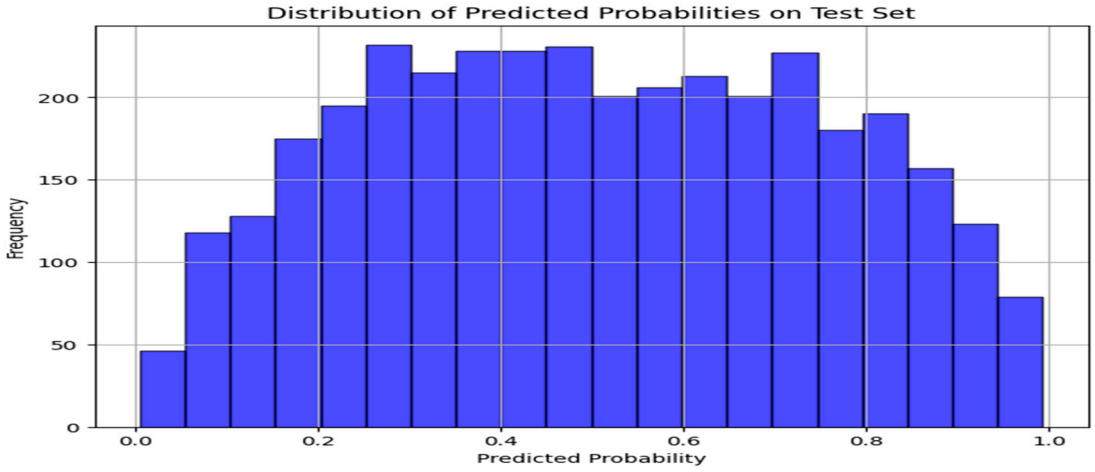


Figure 8. Distribution of Predicted Probabilities on Test Set.

The findings of Bagate, R on the application of sentiment analysis in algorithmic trading using various machine learning models, such as logistic regression, to predict stock market prices showed that logistic regression are prone to error risk but can be effective in the initial monitoring of price movement alone [16]. This aligns with the study findings as sole reliance on sentiment for market

 doi:10.20944/preprints202409.1089.v1

direction could lead to missed or incorrect predictions. Instead, strategies should consider the inherent uncertainty and potential for varied market reactions. The sentiment score daily volatility suggests a reactive market environment, while the model's cautious probability distribution indicates that it is well-calibrated to handle this volatility without making overly confident predictions. However, the model performs well but is not perfect. This shows LR can distinguish between positive and negative outcomes with some uncertainty. Therefore, investment strategies should incorporate this uncertainty rather than relying solely on market sentiment.

4. Discussion

The table below shows the classification results of the sentiment analysis on financial news

Table 5. Comparative Analysis of Sentiment Analysis Models.

Test set metrics	GPT Predefined Approach (%)	FinBert (%)	Logistic Regression (%)
Accuracy	54.19	63.33	81.83
Precision	72.66	63.76	82.57
Test Recall	45.09	63.33	81.15
Test F1 Score	32.69	63.30	81.85
Test ROC AUC	65.37	65.59	89.76

The GPT predefined approach has high precision (72.66%), moderate ROC AUC (65.37%) and exhibits weakness with low recall (32.69%) and F1 score (45.06%). FinBERT has a balanced performance with slightly better result than GPT predefined approach but with lower precision (63.66%). FinBERT requires large dataset, substantial computational resources and experience to fine-tune and implement. This is unlike Logistic regression that shows high accuracy (81.83%) and precision (82.57%) to reflect a strong overall performance. The model's F1 Score (81.85%) and ROC AUC (89.76%) are the highest, which indicates a good balance between precision and recall. Logistic regression is simple and interpretable, but may struggle with complex, non-linear relationships in the data. The discrepancy in the sentiment distribution of GPT predefined approach could be attributed to rules of simple assessments or heuristic methods associated to the predefined approach. The method often lack flexibility and adaptability to new data patterns. It lacks the capacity for improvement and learning of a machine learning model, which can learn and provide more accurate sentiment predictions. Although, the precision metric is high but the overall performance metrics shows that the predefined approach may not generalize well to unseen data or may inherit bias from training data. The operation like black boxes by GPT may lead to questioning the result generated. Paripati, L highlighted ethical issues such as privacy and data protection, accountability and governance, bias and fairness, etc. when utilising GPT models for data analysis [17].

Yang, H demonstrated how FinBERT outperformed conventional models in capturing market sentiment in sentiment analysis of financial texts [18]. This aligns with the study that confirmed FinBERT ability to process unstructured text and extract insights for financial prediction and analysis. However, the high computational power and processing time of FinBERT make it a barrier for practical deployment. Logistic regression having the best result out the three models examined, works well with structured data, can capture linear relationships, and is favoured for transparent models. Although, it has limitations with non-linear data, but its ability to perform well on the financial dataset may likely due to effective feature engineering and regularization. Kumar, R highlighted logistic regression effectiveness in financial risk prediction when combined with feature selection techniques [29]. The ease of use and implementation of logistic regression are important in financial applications. Logistic Regression is the best choice for the news dataset due to its high accuracy, precision, and F1 Score. FinBERT, a competitive metric and excellent sentiment analysis, may not be suitable for this dataset due to its complexity and resource demands. Although, FinBERT provided competitive metric result to logistic regression but its complexity and resource demands

 doi:10.20944/preprints202409.1089.v1

may not warrant its utilisation above logistic regression for this financial news dataset. The study did not rule out the consideration of FinBERT for future exploration on specific tasks that involve nuanced sentiment analysis in financial texts.

Based on the findings of this study, we recommend prioritizing Logistic Regression for NGX index sentiment prediction tasks, most especially when computational efficiency and model simplicity are important. Its high accuracy, combined with minimal computational resources, makes it a practical choice for real-time market forecasting. However, FinBERT and GPT-4 should not be overlooked. Their ability to analyse complex textual data and understand nuanced sentiment is valuable for comprehensive stock market sentiment prediction models. For future research, we recommend exploring hybrid models that combine the strengths of Logistic Regression's simplicity and accuracy with the depth of sentiment analysis provided by FinBERT and GPT-4. Furthermore, additional external data sources, such as macroeconomic indicators or geopolitical news, should be integrated into these models to improve predictive accuracy. This would allow for a more holistic view of the factors influencing stock market behaviour, which would lead to better forecasts. Finally, continued use of hyperparameter optimization tools such as Optuna is crucial for ensuring the models perform at their best across different datasets and market conditions.

5. Conclusions

This study provides a comprehensive analysis of the predictive capabilities of FinBERT, GPT-4, and Logistic Regression for stock index prediction using the NGX All-Share Index dataset. Our results indicate that Logistic Regression, despite being a simpler model, outperformed FinBERT and GPT-4 in terms of accuracy, precision, recall, F1 score, and ROC AUC. The robustness of Logistic Regression, particularly after hyperparameter tuning, made it the most efficient model for predicting stock market trends, achieving a high accuracy of 81.83% and ROC AUC of 89.76%. FinBERT, while better equipped to handle financial language, faced challenges in terms of computational demands and resource usage, which limited its practical application in real-time prediction scenarios. Although GPT-4 is powerful in general text analysis but it showed limitations when applied specifically to financial data.

These findings suggest that, while advanced NLP models offer promise in sentiment analysis, traditional models like Logistic Regression still provide strong performance with lower computational costs. However, FinBERT and GPT-4 offer avenues for future exploration, most especially when combined with other machine learning techniques in a hybrid approach.

Author Contributions: Conceptualization, S.A and O.S.; methodology, S.A. and O.S; software, S.A. and O.S.; validation, O.S., O.P., and B.O.; formal analysis, S.A. and O.S.; investigation, S.A.; resources, O.S.; data curation, S.A.; writing—original draft preparation, S.A..; writing—review and editing, O.S., O.P. and B.O.; supervision, O.S.; project administration, O.P., B.O. and O.S.; All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not Applicable.

Data Availability Statement: Data is available upon request

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fatouros G, Soldatos J, Kouroumali K. Transforming sentiment analysis in the financial domain with ChatGPT. Machine Learning with Applications. 2023. Available from: https://www.sciencedirect.com/science/article/pii/S2666827023000610
2. Shapiro AH, Sudhof M, Wilson DJ. Measuring news sentiment. Journal of Econometrics. 2022. Available from: https://www.sciencedirect.com/science/article/pii/S0304407620303535.

3. Liu Z, Huang D, Huang K, Li Z, Zhao J. FinBERT: A pre-trained financial language representation model for financial text mining. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2021. Available from: https://www.ijcai.org/proceedings/2020/0622.pdf.
4. Leippold M. Sentiment spin: Attacking financial sentiment with GPT-3. Finance Research Letter. 2023. Available from: https://www.sciencedirect.com/science/article/pii/S154461232300329X
5. Yang J, Wang Y, Li X. Prediction of stock price direction using the LASSO-LSTM model combining technical indicators and financial sentiment analysis. PeerJ Computer Science. 2022. Available from: https://peerj.com/articles/cs-1148.pdf
6. Sidogi T, Mbuvha R, Marwala T. Stock price prediction using FinBERT and LSTM. 2021 IEEE International Conference Systems Man and Cybernetics. 2021. Available from: https://ieeexplore.ieee.org/abstract/document/9659283
7. Gigerenzer, G. (2022). Simple heuristics to run a research group. PsyCh Journal, 11(2), 133-135. https://doi.org/10.1002/pchj.533
8. Bafitlhile, K. D. (2022). A Context-aware Lemmatization Model for Setswana Language using Machine Learning. Botswana International University of Science and Technology. http://repository. biust.ac.bw/handle/123456789/536
9. Taherdoost, H. (2022). What are different research approaches? Comprehensive Review of Qualitative, quantitative, and mixed method research, their applications, types, and limitations. Journal of Management Science & Engineering Research, 5(1), 53-63
10. Priyatno AM, Ningsih L, Noor M. Harnessing machine learning for stock price prediction with random forest and simple moving average techniques. Journal of Engineering and Science Application. 2024. Available from: https://jesa.aks.or.id/index.php/jesa/article/view/1.
11. Lin F, Cohen WW. Semi-Supervised Classification of Network Data Using Very Few Labels. IEEE Conf. 2010. Available from: https://ieeexplore.ieee.org/abstract/document/5562771/
12. Chen T, Zhang Y, Yu G, Zhang D, Zeng L, He Q. EFSA: Towards Event-Level Financial Sentiment Analysis. Computation and Language. arXiv preprint arXiv:2404.08681. Available from: https://arxiv.org/abs/2404.08681
13. Kirtac K, Germano G. Sentiment trading with large language models. Finance Research Letters. Available from: https://www.sciencedirect.com/science/article/pii/S1544612324002575
14. Varghese RR, Mohan BR. Dynamics of Nonlinear Causality: Exploring the Influence of Positive and Negative Financial News on the Indian Equity Market. In: Proceedings of Annual International Conference on Intelligent Systems and Signal Processing; 2023. Available from: https://ieeexplore.ieee.org/abstract/document/10420348/
15. Senapaty MK, Ray A, Padhy N. A Decision Support System for Crop Recommendation Using Machine Learning Classification Algorithms. Agriculture. 2024;14(8):1256.
16. Bagate R, Joshi A, Trivedi A, Pandey A, Tripathi D. Survey on algorithmic trading using sentiment analysis. In: Proceedings of the 6th International Conference on Advance Computing and Intelligent Engineering: ICACIE 2021; 2022 Sep; Singapore: Springer Nature Singapore. p. 241-252.
17. Paripati L, Hajari VR, Narukulla N, Prasad N, Shah J, Agarwal A. Ethical Considerations in AI-Driven Predictive Analytics: Addressing Bias and Fairness Issues. Darpan Int Res Anal. 2024;12(2):34-50.
18. Yang H, Ye C, Lin X, Zhou H. Stock Market Prediction Based on BERT Embedding and News Sentiment Analysis. In: Wang Z, Wang S, Xu H, editors. Service Science. ICSS 2023. Communications in Computer and Information Science. Vol. 1844. Springer; 2023. p. 334-348. https://doi.org/10.1007/978-981-99-4402-6_20.