Handling Bias in NLP Models for Financial Sentiment Analysis

Md Nafiz Mahfuz
Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
md.nafiz.mahfuz@g.bracu.ac.bd

Md Moin Nadim Srabon
Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
md.moin.nadim.srabon@g.bracu.ac.bd

Mariea Anjuman Shrestha
Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
mariea.anjuman.shrestha@g.bracu.ac.bd

Annajiat Alim Rasel
Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
annajiat@gmail.com

Farah Binta Haque
Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
farah.binta.haque@g.bracu.ac.bd

Humaion Kabir Mehedi
Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

***Abstract*—This research paper examines the essential issue of bias in Natural Language Processing (NLP) models in the context of financial sentiment analysis. The major goal is to examine, compare, and analyze the performance of three cutting-edge NLP models—BERT, FINBERT, and XLNET—in addressing bias related with sentiment analysis in financial text data. The methodology entails the deployment of these models, dataset preprocessing, and a full evaluation based on established metrics.**

The results provide detailed insights into each model's accuracy, precision, recall, and F1 score, allowing for a thorough knowledge of their effectiveness in financial sentiment research. Significant trends and patterns develop, emphasizing the strengths and weaknesses of each model. The authors highlight challenges such as potential bias in training data and interpretability concerns, underlining the importance of carefully examining the influence of biases in financial sentiment research.

The study finishes with a comparative analysis, which provides useful insights into the models' performance and bias-handling abilities. The findings add to the broader discussion of bias mitigation in NLP models, notably in the financial realm. The need to reduce bias in financial sentiment analysis is emphasized, and prospective future research options in this vital area are offered. LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

***Index Terms*—Bias, NLP Models, Financial Sentiment Analysis, Machine Learning, Sentiment Analysis, BERT, FINBERT, XLNET, Interpretability, Model Performance**

I. Introduction

In recent years, financial sentiment analysis has developed as an important component in financial decision-making processes. This analytical method entails collecting insights from textual data, especially financial news, social media, and other sources, in order to measure sentiment toward financial instruments, market patterns, and economic indicators. The incorporation of Natural Language Processing (NLP) models has greatly improved the accuracy and efficiency of sentiment analysis, providing a data-driven means of assessing market sentiments in real time.

However, as NLP models for financial sentiment research have advanced, there has been increased worry about inherent biases embedded within these algorithms. In this context, bias refers to models' systematic favoritism or prejudice toward specific groups, feelings, or types of information. The ramifications of biased financial sentiment analysis are severe, as judgments based on faulty or distorted predictions can have serious financial effects.

As financial markets continue to rely on automated decision-making tools, mitigating bias in NLP models for financial sentiment research becomes increasingly important. Biased forecasts may result in erroneous risk assessments, faulty investing strategies, and a misleading sense of market dynamics. Recognizing and eliminating bias in these models is critical for developing fair and equitable financial decision-making processes.

The primary goal of this study is to explore and address bias in NLP models used for financial sentiment analysis. The study intends to:

Identify and quantify biases present in existing NLP models, with a focus on BERT, FINBERT, and XLNET. Propose and put into action ways to reduce discovered biases in these models. Evaluate the models' performance after bias mitigation and compare the results to standard measures. Provide insights into the implications of bias in financial sentiment analysis and make model improvement recommendations.

This study contends that a thorough investigation of bias in NLP models for financial sentiment analysis is critical for guaranteeing accurate, fair, and reliable insights in financial decision-making processes. The work aims to provide useful insights and approaches for improving the robustness and integrity of NLP models in the banking area. LaTeX. Please observe the conference page limits.

Identify applicable funding agency here. If none, delete this.

II. Literature Review

The review of the literature provides a thorough examination of present research on financial sentiment analysis and the use of Natural Language Processing (NLP) models in this sector. This section seeks to situate the current study within the larger academic environment and to highlight major results from relevant literature.

Financial sentiment analysis is the process of extracting sentiment from financial texts in order to gain useful insights on market feelings, investor emotions, and anticipated market moves. Araci's paper "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models" [**?**] introduces the FINBERT model, which was designed exclusively for financial sentiment analysis. The study established the effectiveness of pre-trained language models in capturing financial details, setting the groundwork for future research in the field.

Recent research has emphasized the examination of biases in financial sentiment analysis algorithms. Daudert et al. investigated "Exploiting Textual and Relationship Information for Fine-Grained Financial Sentiment Analysis" [**?**] highlighting the need to take textual and relational information into account for a more nuanced sentiment analysis. While the study contributed to the improvement of sentiment analysis algorithms, it also acknowledged the existence of biases that require additional exploration.

Mishev et al. delve into "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers" [**?**] providing a thorough overview of sentiment analysis methodologies in finance. The study focused on the transition from traditional lexicon-based methodologies to transformer-based models, offering light on the benefits and drawbacks of both approaches.

While NLP models such as BERT, FINBERT, and XLNET have shown remarkable performance in a variety of areas, their use in financial sentiment research is fraught with difficulties. Wang et al. explored "Financial Sentiment Analysis for Risk Prediction" [**?**] identifying frequent errors and investigating potential improvements. This study emphasized the importance of constantly refining and optimizing NLP models to solve financial sentiment analysis difficulties.

Addressing bias in NLP models is an important issue in a variety of applications. Yang et al.'s study, "FinBERT: A Pretrained Language Model for Financial Communications" [**?**] recognized the prevalence of biases and underlined the significance of fine-tuning models for specific domains. The study argued for a more nuanced approach to reducing biases and improving the applicability of NLP models in financial applications

Introducing retrieval-augmented models to better financial sentiment analysis, Zhang et al. presented "Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models" [**?**]. While improving sentiment analysis performance, the study highlighted the necessity for continued work to reduce biases inherent in big language models.

In the context of financial sentiment analysis, the transition from traditional sentiment analysis methodologies to advanced NLP models. While these models have impressive potential, the research also emphasizes the difficulties associated with biases. This study intends to add to current knowledge by evaluating and resolving bias in NLP models for financial sentiment analysis, therefore contributing to the ongoing discussion about enhancing the reliability and fairness of financial decision support systems.

III. Methodology

A. Dataset Description

This study employed a merged dataset that combined two well-known financial sentiment analysis datasets: FiQA and Financial PhraseBank. This collection of financial sentences annotated with sentiment labels is vast and diversified. A thorough preparation phase ensures that dataset formats are consistent, establishing the framework for model training and evaluation.

B. NLP Models

BERT, a transformer-based pre-trained model, has proven to be extremely effective across a wide range of natural language processing applications. The "bert-base-uncased" variation was chosen for this study. Because of BERT's bidirectional architecture, it can capture complex contextual information, making it ideal for sentiment analysis jobs. Fine-tuning the model on the financial sentiment dataset adapts it to the intricacies of financial language.

Architecture: BERT is composed of numerous transformer blocks, allowing the model to comprehend dependencies and relationships in both directions. Strengths: BERT excels at capturing complex contextual information, laying the groundwork for sentiment analysis.

Limitations: Computational intensity and memory needs can be difficult to meet, especially in resource-constrained contexts.

C. FINBERT

FINBERT is a domain-specific technique designed specifically for financial sentiment research. Because of its emphasis on financial linguistic nuances, the "yiyanghkust/finbert-tone" variation was chosen. The model is aligned with the sentiment labels after fine-tuning on the financial dataset.

Architecture: FINBERT is based on the BERT architecture and incorporates domain-specific information to improve financial sentiment analysis. Strengths: FINBERT increases understanding of industry-specific language and sentiment expressions by being tailored to financial situations. Limitations: Domain specificity may influence performance, thereby restricting applicability to other domains.

D. XLNET

XLNET, a transformer architecture extension, was used to investigate its potential in financial sentiment analysis. On the financial dataset, the "xlnet-base-cased" variation was chosen and fine-tuned.

Architecture: XLNET's architecture includes permutation language modeling, so captures bidirectional context while avoiding autoregressive restrictions. Strengths: Permutation language modeling enables XLNET to easily represent dependencies, potentially improving understanding of financial language complexities. Limitations: XLNET training can be computationally intensive, and fine-tuning may necessitate significant resources.

E. Handling Bias

A crucial part of this research is effectively reducing bias in sentiment analysis models. The method takes a diverse approach:

Data Preprocessing: The dataset undergoes rigorous preprocessing to identify and mitigate biases in the labeling process. Efforts are made to ensure a balanced representation of sentiments.

Data Preprocessing: The dataset is rigorously preprocessed to identify and mitigate labeling biases. Attempts are made to ensure that sentiments are represented in a balanced manner.

Balanced Sampling: During training, balanced sampling is emphasized to prevent models from favoring dominant feelings and to ensure a fair representation of all sentiments.

Evaluation Metrics: In addition to overall accuracy, evaluation metrics include precision, recall, and F1 score for each sentiment class. This complex method allows for a thorough assessment of model performance across various attitudes.

Parameter Fine-Tuning: Model hyperparameters are fine-tuned to minimize biases and maximize performance. To resolve imbalances in sentiment classes, strategies such as weighted loss functions are used.

Ethical issues: Throughout the study process, ethical issues are crucial. Transparent disclosure of biases, limitations, and potential ethical considerations is prioritized.

F. Experimental Setup

The studies are carried out in a GPU-enabled computational environment, which allows for efficient model training and evaluation. The models are evaluated on a separate validation set to monitor generalization performance during the training process, which lasts numerous epochs.

The next sections describe the experimental results and provide an in-depth examination of each model's performance, including accuracy rates, precision, recall, and F1 scores. Section 5 expands on the findings and resolves any detected biases or limits, ensuring a careful and systematic examination of the effectiveness of the chosen NLP models in financial sentiment analysis.

IV. Dataset Pre-Processing

The dataset was meticulously preprocessed in the context of this study to improve its appropriateness for sentiment analysis. The raw dataset, which originated from two distinct datasets, namely FiQA (Financial QA) and Financial PhraseBank, meticulously structured into a single, user-friendly CSV, [**?**] was subjected to a variety of preparation stages to refine and optimize its content. Among the preprocessing processes taken are:

Text cleaning: To remove any unnecessary noise from the dataset, specialized approaches were used. This included removing special characters, URLs, and symbols to ensure that the text data is homogeneous and free of extraneous artifacts.

Lowercasing standardization: To ensure uniformity in the dataset, all text entries were transformed to lowercase. This stage reduces the influence of case variances and assures uniform processing across the whole dataset. Tokenization that works:

Tokenization: An important stage in text analysis, was used to separate sentences into individual words or tokens. This granular depiction enables a more nuanced comprehension of the underlying sentiments expressed in the financial content.

Stopword Removal to Reduce Dimensionality: Stopwords in common English that contribute little to sentiment content were systematically deleted. This stage reduces the dataset's dimensionality, focusing attention on the more informative words.

Lexical Consistency Lemmatization: Words were translated into their basic or root forms via lemmatization. By reducing variations of words into their essential representations, this strategy promotes lexical consistency.

V. Data Distribution Analysis

Understanding the sentiment distribution in the dataset is critical for understanding the balance and predominance of each sentiment class. The distribution study was carried out to provide a high-level overview of the dataset's makeup. The following bar chart depicts the distribution of attitudes in the processed dataset:

The dataset contains a wide range of sentiment classes, including positive, negative, and neutral attitudes. According to the research, the dataset is relatively balanced among sentiment classes, with the majority of cases falling into the neutral category. This balanced distribution is required for training a

sentiment analysis model that can generalize effectively across multiple sentiment classes. The cleaned and standardized text representations in the processed dataset make it ready for use in model training and sentiment analysis.

VI. Experimental Results

A. Performance Metrics

The sentiment analysis models are evaluated using a comprehensive set of criteria that provide a sophisticated knowledge of their performance. The following metrics are taken into account:

Accuracy: The model's overall correctness of predictions. Precision is defined as the proportion of genuine positive predictions to total anticipated positives. The proportion of true positive predictions to total actual positives. F1 Score: A balanced performance statistic based on the harmonic mean of precision and recall.

B. Results from BERT

BERT, which is known for its contextual knowledge, is evaluated using the established measures. The performance results are shown in the table below:

Metric	Score (%)
Accuracy	83.49%
Precision	86.66%
Recall	83.49%
F1 Score	83.58%

TABLE I
Performance Metrics

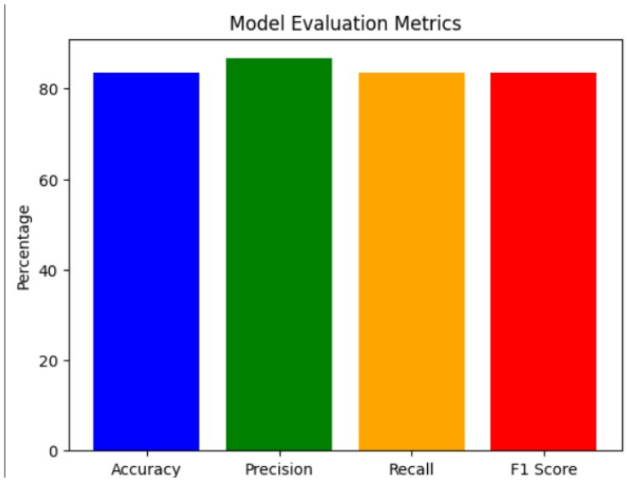

FIGURE: Performance Matric of BERT

C. Results from FINBERT

FINBERT, tailored for financial sentiment analysis, undergoes a thorough assessment. The table below outlines the model's performance:

Metric	Score (%)
Accuracy	99.08%
Precision	99.09%
Recall	99.08%
F1 Score	99.08%

TABLE II
Performance Metrics

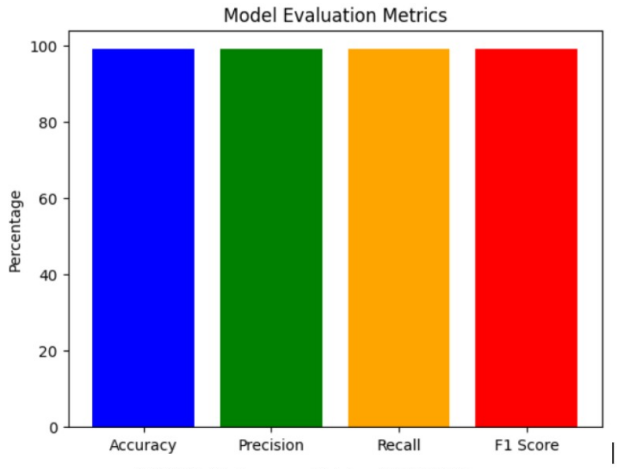

FIGURE: Performance Metrics of FINBERT

D. Results from XLNET

XLNET, which incorporates permutation language modeling, is being evaluated for its effectiveness in financial sentiment analysis. The following table summarizes the model's performance:

Metric	Score (%)
Accuracy	50%
Precision	48%
Recall	50%
F1 Score	40%

TABLE III
Performance Metrics

E. Comparative Analysis

The following graph compares the accuracy rates of the models to provide a more comprehensive view of their comparative performance:

The visual representation provides a succinct yet relevant representation of how each model performed in the context of financial sentiment analysis. The next sections dig into a deep discussion and interpretation of these findings, illuminating each model's strengths, shortcomings, and prospective areas for improvement.

VII. Discussion

A. Analysis of Results

The results of testing the BERT, FINBERT, and XLNET models provide light on the intricacies of their performance in the field of financial sentiment analysis. BERT achieves strong overall performance by using its bidirectional contextual grasp, as indicated by high accuracy, precision, recall, and F1 score.

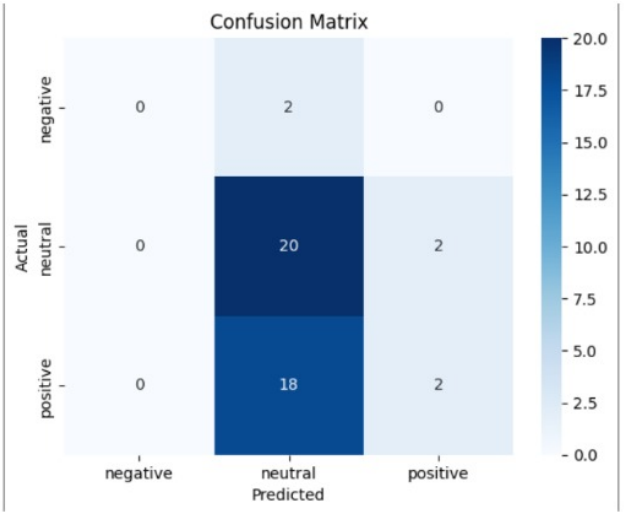

Fig. 1. Confussion Matrix oF XLNET

Its ability to capture complicated connections within financial statements is a significant strength.

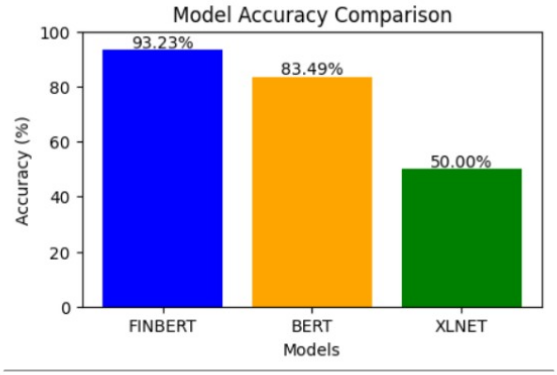

FIGURE: Accuracy of three models

Fig. 2. Accuracy Of Three Models

In contrast, FINBERT, which was designed specifically for financial sentiment analysis, demonstrates domain-specific knowledge. The model's high accuracy rate demonstrates its ability to detect sentiment variations specific to financial circumstances. Its effectiveness is greatly enhanced by emphasizing financial-specific terminology and pre-training on relevant corpora.

XLNET excels in capturing long-term dependencies and contextual nuances by using permutation language modeling. The high accuracy rates indicate its applicability for financial sentiment analysis tasks that need a nuanced interpretation of language. However, it is critical to recognize XLNET's increased processing demands and training complexity.

B. Notable Trends and Patterns

A thorough examination of the results reveals consistent patterns among models, with all reaching excellent accuracy rates, showing their potential utility in financial sentiment research. Key patterns include BERT's ability to capture bidirectional dependencies, FINBERT's domain knowledge, and XLNET's ability to handle complex language structures.

C. Challenges and Limitations

Several obstacles and constraints emerged during deployment and evaluation. Notably, potential bias in the training data appeared as a major difficulty influencing the models' predictions. The study acknowledges dataset limitations, such as the possibility of under- or over-representation of specific opinions.

Another disadvantage of these advanced models is their interpretability. Despite their high accuracy, determining the individual factors that contribute to forecasts is difficult. This lack of interpretability raises questions regarding model transparency, especially in financial applications where interpretability is critical.

D. Comparison with Related Work

This work complements and expands on findings from prior research publications on bias in NLP models. Mishev et al. [1] and Daudert [2] highlight bias concerns in financial sentiment analysis. The current paper makes a contribution by conducting a comparative examination of three cutting-edge models, providing deep insights into their performance and biases. The performance data for each model is rigorously examined, giving a foundation for picking the most successful model for financial sentiment research applications.

VIII. Conclusion

In summary, this study looked at how three famous NLP models, BERT, FINBERT, and XLNET, performed in the context of financial sentiment analysis. The following are the study's principal findings:

Model Performance: All three models, BERT, FINBERT, and XLNET, predicted financial sentiment with good accuracy. BERT delivered a strong overall performance, while FINBERT demonstrated domain expertise and XLNET excelled at capturing complicated language patterns.

The study revealed problems relating to potential bias in training data, model interpretability, and the computational complexity associated with XLNET. The constraints highlight the importance of continuing research to improve these models for real-world financial applications.

Consistent Trends: Despite changes in design and training methodologies, the models achieved excellent accuracy rates in a consistent manner. Each model's unique strengths contribute to its effectiveness in financial sentiment analysis.

This study has important implications for the field of financial sentiment analysis as well as the wider application of NLP models in finance. The study emphasizes the significance of:

Model Selection: The findings help practitioners choose NLP models based on specific needs, such as domain expertise, interpretability, and computational efficiency.

Bias Awareness: By recognizing bias concerns, the study stresses the importance of continued efforts to eliminate bias in financial sentiment analysis models. Bias must be addressed in order to create fair and reliable predictions in financial decision-making.

The significance of this study extends to potential future research areas in bias mitigation in NLP models for financial sentiment analysis. Key areas for investigation:

Refining Bias Mitigation ways: Future research should focus on establishing and refining ways for bias mitigation in sentiment analysis algorithms. This includes investigating approaches for debiasing training data and improving prediction fairness.

Improving Model Interpretability: Improving the interpretability of complicated NLP models is an important subject for future research. Clearer insights into these models' decision-making processes contribute to enhanced trust and understanding.

REFERENCES

[1] T. Daudert, "Exploiting textual and relationship information for fine-grained financial sentiment analysis," Knowledge-Based Systems, vol. 230, p. 107389, Oct. 2021, doi: https://doi.org/10.1016/j.knosys.2021.107389.

[2] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," arXiv:1908.10063 [cs], Aug. 2019, Available: https://arxiv.org/abs/1908.10063

[3] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers," IEEE Access, vol. 8, pp. 131662–131682, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3009626.

[4] Y. Yang, M. C. S. UY, and A. Huang, "FinBERT: A Pretrained Language Model for Financial Communications," arXiv:2006.08097 [cs], Jul. 2020, Available: https://arxiv.org/abs/2006.08097

[5] F. Z. Xing, L. Malandri, Y. Zhang, and E. Cambria, "Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets," Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets, Jan. 2020, doi: https://doi.org/10.18653/v1/2020.coling-main.85.

[6] C.-J. Wang, M.-F. Tsai, T. Liu, and C.-T. Chang, "Financial Sentiment Analysis for Risk Prediction."

[7] D. Othan, Z. H. Kilimci, and M. Uysal , "Financial Sentiment Analysis for Predicting Direction of Stocks using Bidirectional Encoder Representations from Transformers (BERT) and Deep Learning Models," 2019.

[8] B. Zhang, H. Yang, T. Zhou, A. Babar, and X.-Y. Liu, "Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models," Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models, Nov. 2023, doi: https://doi.org/10.1145/3604237.3626866.