School of Computing, Engineering & Digital Technologies
Middlesbrough TS1 3BA

CIS4055-N-FJ1-2023

Empowering Market Forecasting and Investment Decisions with Advanced Sentiment Analysis Techniques.

(A ML and NLP Approach).

Submitted in partial requirements for the degree of MSc Data Science

Chinedu Emmanuel Agwunobi

B1555134

Supervised by Chaimaa Tarzi

ABSTRACT

In the realm of financial decision-making, leveraging sentiment analysis driven by Machine Learning (ML) and Natural Language Processing (NLP) is vital. This dissertation explores advanced sentiment analysis techniques within the financial domain, building upon prior research in Financial Sentiment Analysis (FSA).

The study integrates traditional ML techniques like Logistic Regression (LR), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) with state-of-the-art models like Bidirectional Encoder Representations from Transformers (BERT) and its financial variant, FinBERT. The methodology aims to contribute to FSA by systematically addressing challenges and exploring opportunities highlighted in existing literature.

Through meticulous data preprocessing, feature engineering, and model training, the research evaluates the effectiveness of different techniques. Notably, ensemble techniques and model interpretability using Explainable AI methodologies, such as SHAP and LIME, are emphasized.

Findings reveal competitive performance from both traditional ML and NLP models, with Ensemble SVM and RoBERTa emerging as top performers, respectively. The study contributes valuable insights to the field of financial sentiment analysis, paving the way for future advancements.

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this dissertation. A special gratitude I give to my academic supervisor, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this dissertation.

Furthermore, I would also like to acknowledge with much appreciation the crucial role of my friends and family, who have supported me throughout entire process, both by keeping me harmonious and helping me putting pieces together. I am grateful for their aspiring guidance, invaluable constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

I express my warm thanks to my parents and siblings for their unfaltering support and encouragement throughout my study. This accomplishment would not have been possible without them. Thank you.

I would like to express my special gratitude to my uncles and aunties for their words of encouragement and constant motivation. Their support and love have been invaluable to me.

Finally, I wish to thank all those, who have contributed to this dissertation and supported me in any respect during the completion of the project.

ABBREVIATIONS

Abbreviations	Meaning
ML	Machine Learning
DL	Deep Learning
LR	Logistice Regression
SVM	Support Vector Machine
RF	Random Forest
DT	Decision Tree
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
XGBoost	Extreme Gradient Boosting

Table of Contents

1 INTRODUCTION

Sentiment analysis, a key discipline within Natural Language Processing (NLP), plays a vital role in identifying and categorizing sentiments within textual data. This process allows for the extraction of valuable insights from text, facilitating informed decision-making and personalized responses Liu (2022). Sentiment analysis techniques have evolved beyond basic text classification into positive, negative, or neutral sentiments, now capable of discerning specific emotions and nuanced aspects of sentiment Cambria et al. (2017).

In the financial domain, sentiment analysis takes on a significant role through Financial Sentiment Analysis (FSA). FSA studies investor sentiment and financial textual sentiment, providing essential insights for understanding market dynamics and supporting robust financial decision-making Kearney and Liu (2014). However, the unique challenges presented by FSA, such as the domain-dependent nature of financial sentiment indicators, necessitate specialized approaches.

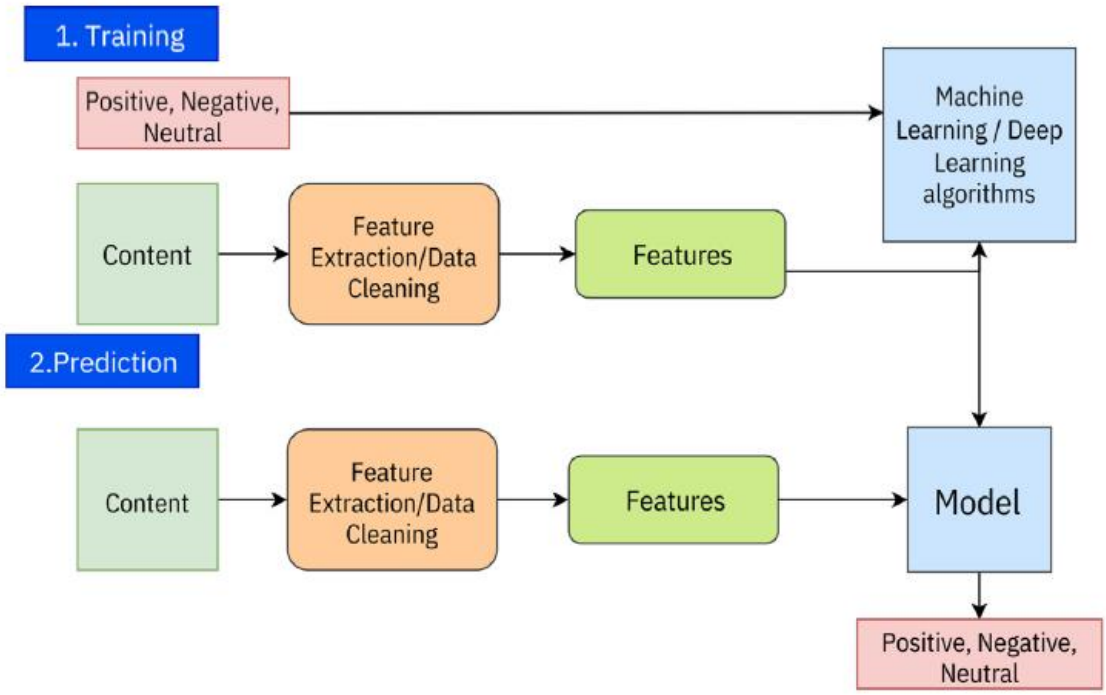


Fig. 1. Working Process of Sentiment Analysis Jim et al. (2024)

Recent advancements in machine learning (ML) and deep learning (DL) techniques have significantly enhanced the accuracy and scalability of sentiment analysis systems, leading to the development of sophisticated models for comprehensive sentiment analysis across various domains Revathy et al. (2022); Abdullah and Ahmet (2022).

This study aims to provide a comprehensive review of recent FSA research. The goal is to bridge technical and applied perspectives, enhancing the adoption of advanced sentiment analysis techniques in market forecasting and investment decision-making. Through an extensive examination of FSA studies, this dissertation will shed light on the dynamic interplay between sentiment analysis techniques and their practical applications in financial markets. Furthermore, it will review recent learning approaches, pre-trained language models, and evaluation methods, aiming to pave the way for more promising results in FSA research and applications.

1.1 Research Motivation

The motivation behind this research stems from the recognition of the critical role sentiment analysis plays in understanding market dynamics and predicting asset price movements. With the exponential growth of digital data and the proliferation of social media and news platforms, there is an abundance of unstructured textual data that contains valuable insights into market sentiment. Leveraging advanced sentiment analysis techniques can unlock this wealth of information, providing investors with a competitive edge in making timely and well-informed investment decisions.

1.2 Research Aims

The primary aim of this research is to explore how advanced sentiment analysis techniques, driven by ML and NLP, can enhance market forecasting accuracy and empower investment decisions. By analyzing the role of sentiment analysis in financial markets and evaluating traditional and state-of-the-art approaches, this research aims to provide insights into the practical applications and limitations of sentiment analysis for market forecasting.

1.3 Research Objectives

- Investigate the significance of sentiment analysis in informing investment decisions.
- Review traditional approaches to sentiment analysis and assess their effectiveness.
- Explore state-of-the-art techniques in sentiment analysis, including deep learning and contextual embeddings.
- Examine how ML techniques, including advanced preprocessing techniques and the uses of Bag-of-Words (BOW) and TF-IDF, can be applied to enhance market forecasting accuracy.
- Investigate the role of NLP in extracting actionable insights from textual data for market forecasting.
- Explore the use of SHAP and LIME to gain meaningful insights from sentiment analysis results
- Identify and analyze the challenges and limitations associated with sentiment analysis for market forecasting.

1.4 Research Questions

- What is the role of sentiment analysis in shaping investment decisions in financial markets?
- How effective are traditional approaches to sentiment analysis in capturing market sentiment?

- What are the latest advancements in sentiment analysis techniques, and how do they improve market forecasting accuracy?
- How can ML techniques be leveraged to enhance market forecasting using sentiment analysis?
- What are the key applications of NLP in extracting market sentiment from textual data?
- What are the primary challenges and limitations in utilizing sentiment analysis for market forecasting?

1.5 Research Problems

This research addresses several key problems, including:

- The need to accurately capture and interpret market sentiment from vast amounts of unstructured textual data.
- The challenge of integrating sentiment analysis techniques with existing market forecasting models.
- The limitations of traditional sentiment analysis approaches in capturing nuanced market sentiment.
- The need to overcome data quality, scalability, and interpretability issues in sentiment analysis for market forecasting.

1.6 Research Contribution

This research aims to contribute to the existing body of knowledge by:

Investigating the significance of sentiment analysis in informing investment decisions: This research provides insights into the practical applications of sentiment analysis in market forecasting and investment decision-making, enhancing understanding of its importance in the financial domain.

Reviewing traditional approaches to sentiment analysis: By assessing the effectiveness of traditional sentiment analysis approaches, this research contributes to the existing knowledge base and identifies areas for enhancement.

Exploring state-of-the-art techniques in sentiment analysis: By examining cutting-edge techniques, this research advances the field and guides future research directions.

Investigating the application of Machine Learning (ML) techniques: This research explores how ML techniques can enhance sentiment analysis, paving the way for more sophisticated models.

Examining the role of Natural Language Processing (NLP): By exploring NLP's role in sentiment analysis, this research sheds light on leveraging textual data for market forecasting.

Utilizing SHAP and LIME for meaningful insights: This research employs SHAP and LIME to interpret complex ML models, promoting transparency and comprehension.

Identifying and analyzing challenges and limitations in sentiment analysis: By addressing challenges and limitations, this research offers practical solutions to improve sentiment analysis accuracy in financial markets.

1.7 Structure

- Section 2 provides research background .

- Section 3 provides the methodology employed in this study.
- Section 4 delves into implementation of the Models.
- Section 5 shows the results and discussions.
- Section 6 Conclusion, Recommendations and future research.

2 BACKGROUND

This chapter provides an overview of key concepts essential for understanding the role of sentiment analysis in market forecasting and investment decisions.

2.1 Overview of Market Forecasting

Market forecasting is a critical aspect of financial planning and investment decision-making. It involves predicting future trends, movements, and behaviors of financial markets, including stocks, commodities, currencies, and indices. The goal of market forecasting is to anticipate changes in market conditions, identify potential opportunities, and mitigate risks Petropoulos et al. (2022).

The process of market forecasting leverages historical data, such as price trends, volume, volatility, and other market indicators, to predict future price movements. These predictions are often facilitated by forecasting models, which can range from simple statistical models, like moving averages and linear regression, to more complex machine learning models.

Traditional approaches to market forecasting often rely on fundamental analysis, technical analysis, and economic indicators to make predictions. Fundamental analysis involves evaluating a company's financial statements, management, competitive advantages, and its market and competitors. Technical analysis, on the other hand, focuses on patterns in market data to identify trends and make forecasts. Economic indicators such as GDP, employment levels, and political stability are also considered in market forecasting Petropoulos et al. (2022).

However, these traditional approaches may not always capture the complexities and nuances of market dynamics, leading to suboptimal decisions. They often fail to account for the sentiment of investors, which can significantly influence market trends. This highlights the need for more advanced and nuanced approaches to market forecasting, such as sentiment analysis techniques powered by machine learning and natural language processing.

2.2 Sentiment Analysis in the Financial Domain

Sentiment analysis, a key component of Natural Language Processing (NLP), involves examining documents such as news articles, message board postings, or product reviews to determine their sentiment towards a certain topic Pang and Lee (2008). In the financial context, sentiment analysis is used to understand the opinions, expectations, or beliefs of market participants towards specific companies or financial instruments Brown and Cliff (2004).

There are two broad strategies for performing sentiment analysis: supervised and unsupervised approaches Zhou and Chaovalit (2008). Supervised approaches require a manually labeled dataset to train a classifier, which can then be used to determine the sentiment of further documents or sentences. Unsupervised approaches, on the other hand, rely on external knowledge such as predefined dictionaries to determine a sentiment measure Thelwall et al. (2011)

Several studies have applied these approaches in the financial domain. For instance, Antweiler and Frank (2004) used a supervised approach to analyze messages posted on finance message boards. They found that the sentiment expressed in these messages could predict stock returns. Similarly, Tetlock (2007) used an unsupervised approach to analyze the sentiment of a daily Wall Street Journal column and found that high pessimism led to a decline in market prices.

However, the influence of sentiment on stock returns is not strong enough to serve as the sole source for forecasting future stock returns Antweiler and Frank (2004). Therefore, combining sentiment analysis with an intraday text mining approach could be a promising strategy for improving the accuracy of market forecasts.

2.3 Traditional Approaches to Sentiment Analysis

Traditional approaches to sentiment analysis encompass a variety of techniques, including rule-based methods, lexicon-based methods, and manual annotation.

Rule-based methods rely on predefined rules and patterns to classify text sentiment. These rules can be based on grammatical structures, specific keywords, or combinations of words. While rule-based methods can be effective for simple and structured texts, they often struggle with more complex and unstructured texts Du, Tsai and Wang (2019).

Lexicon-based methods use sentiment lexicons or dictionaries to assign sentiment scores to words or phrases. These lexicons typically contain lists of words along with their associated sentiment scores. The sentiment of a text is then determined by aggregating the sentiment scores of the words it contains. However, lexicon-based methods can be limited by the quality and coverage of the lexicon, and they often fail to capture the context in which words are used Lengkeek, van der Knaap and Frasincar (2023).

Manual annotation involves human annotators labeling text data with sentiment labels based on their subjective judgment. While this approach can produce high-quality sentiment labels, it is time-consuming, costly, and lacks scalability.

2.4 State-of-the-Art Techniques in Sentiment Analysis

Recent strides in Machine Learning (ML) and Natural Language Processing (NLP) have propelled the development of cutting-edge techniques in sentiment analysis. Among these advancements are deep learning models, transfer learning, and transformer-based architectures.

Deep learning models, including Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have demonstrated prowess in automatically learning hierarchical representations of text data. These models excel at capturing intricate patterns and relationships within textual information, thereby enhancing sentiment analysis accuracy.

Transfer learning has emerged as a potent technique in sentiment analysis, enabling models pretrained on extensive text corpora to be fine-tuned for specific tasks. By leveraging pretrained models and adapting them to domain-specific contexts, transfer learning enhances performance and promotes generalization in sentiment analysis tasks.

Transformer-based architectures, exemplified by models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have garnered significant attention for their remarkable achievements across various NLP tasks, including sentiment analysis. These

architectures leverage self-attention mechanisms and large-scale pretraining to effectively capture contextual information and semantic nuances in textual data, thereby elevating the accuracy and efficacy of sentiment analysis processes Jim et al. (2024).

2.5 Application of Machine Learning in Market Forecasting

Machine Learning (ML) techniques have been extensively applied in market forecasting, leveraging a variety of models including decision trees, support vector machines, neural networks, and more recently, deep learning models. These models are capable of learning complex patterns in data and making predictions based on these patterns. Deep learning models, in particular, have shown promising results due to their ability to model high-level abstractions in the data.

The application of ML in market forecasting can be categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning algorithms, such as regression and classification, are used to learn predictive models from historical market data, which are then used to make future predictions. Unsupervised learning techniques, such as clustering and dimensionality reduction, are used to uncover hidden patterns and structures in market data. These techniques aid in market segmentation and anomaly detection, providing valuable insights into market trends and behaviors Gutiérrez-Fandiño et al. (2021).

Reinforcement learning algorithms, inspired by behavioral psychology, are used to learn optimal trading strategies. These algorithms interact with the market environment and receive feedback on their actions, allowing them to continuously improve and adapt their strategies over time.

In recent years, the advent of deep learning models, including decision trees, support vector machines, and neural networks, has further propelled advancements in market forecasting. These models excel at capturing intricate patterns and high-level abstractions in the data, yielding promising results in predicting market trends and behaviors.

2.6 Application of Natural Language Processing in Market Forecasting

Natural Language Processing (NLP) techniques, including sentiment analysis, text summarization, and named entity recognition, are indispensable tools for market forecasting and investment decisions. These techniques extract valuable information from textual data such as news articles, research reports, and social media posts.

Sentiment analysis allows investors to gauge both market and investor sentiment in real-time. This real-time analysis enables swift reactions to changing market conditions. Similarly, text summarization techniques provide concise summaries and actionable recommendations by extracting key information and insights from large volumes of data.

Named entity recognition algorithms further enhance the process by identifying and classifying entities mentioned in the text data. These entities can include companies, organizations, and financial instruments, which facilitates efficient information retrieval and analysis.

2.7 Related Works

Li et al. (2021) aim to generate a Chinese financial sentiment lexicon (CFDSL) for financial distress prediction (FDP). Their methodology involves constructing CFDSL using a deep learning-based framework, utilizing annual reports from Chinese listed companies. Results indicate deep learning effectively generates CFDSL, covering sentiment related to capital and stock markets, company conditions, and politics. Sentiment features calculated four years prior to the benchmark year yield optimal performance in FDP. CFDSL-based sentiment features show advantages over other lexicons, expanding methods of constructing financial sentiment lexicons and offering insights into Chinese listed companies' financial risk.

Lengkeek, van der Knaap and Frasincar (2023) aim to predict financial aspect classes and polarities within sentences using the hierarchical structure of Financial Question & Answer (FiQA) challenge data. Their methodology incorporates a two-step model to improve aspect class prediction. Results show a 7.6% F1 score improvement over direct aspect classification, enhancing state-of-the-art performance from 0.46 to 0.70. Hierarchical models, particularly RoBERTa, outperform BERT models. They recommend RoBERTa for financial data and explore text normalization techniques. Their study contributes to aspect-based sentiment analysis by proposing hierarchical models and evaluating their effectiveness.

Fatouros et al. (2023) investigate ChatGPT's potential in financial sentiment analysis, focusing on the forex market. They employ a zero-shot prompting approach, examining multiple ChatGPT prompts on a forex-related news headlines dataset. Results show ChatGPT outperforms FinBERT by approximately 35% in sentiment classification and correlates 36% higher with market returns. Prompt engineering in zero-shot contexts is crucial, enhancing sentiment analysis in financial applications. The study concludes by discussing implications, suggesting areas for further investigation, and highlighting large language models' promising outlook in financial sentiment analysis.

Jim et al. (2024) provide a comprehensive review of sentiment analysis, covering algorithms, applications, performance, and challenges. Employing the Systematic Literature Review (SLR) approach, they explore various sentiment analysis domains, pre-processing techniques, datasets, and evaluation metrics. The review delves into Machine Learning, Deep Learning, Large Language Models, and Pre-trained models, discussing their advantages. They also analyze recent experimental results and limitations, proposing future research directions to advance sentiment analysis. The review aims to serve as an interdisciplinary resource, providing insights for investors and researchers to promote innovation and decision-making.

Shams et al. (2024) leverage text-mining algorithms for sentiment analysis on financial datasets, introducing a hybrid feature selection model. Using the Financial Question & Answer (FiQA) dataset, they apply preprocessing techniques, feature extraction, and machine learning classifiers. The ANOVA-PSO hybrid model for LSTM classification achieves 75% accuracy, surpassing other models. Preprocessing and feature selection are vital in financial text classification, with the proposed hybrid method showing promising results. The study suggests further exploration of selected features and different embedding paradigms for financial sentiment analysis advancement.

Sy et al. (2023) analyze financial narratives using argument mining and a BERT-based ensemble learning approach. They segment arguments in earnings conference call data, employing two subtasks: Argument Unit Classification and Argument Relation Detection and Classification. Using the NTCIR-17 FinArg-1 Shared Task dataset, they introduce preprocessing and voting strategies to enhance prediction performance. The study advances text analysis in financial technology, demonstrating ensemble

techniques' potency and proposing a flexible framework for language models. Future work includes exploring newer language models and refining balancing techniques for stakeholder insights.

Gutiérrez-Fandiño et al. (2021) propose FinEAS, a financial sentiment analysis model based on supervised fine-tuned sentence embeddings from BERT. Observing similarities between financial and general domains, they utilize a general-domain model and focus on sentence embeddings. FinEAS outperforms baselines like vanilla BERT and FinBERT, exhibiting consistent performance and robust out-of-sample results. Despite using a single dataset, its large and diverse nature ensures reliability. The authors make their code and model weights publicly available, suggesting further exploration of Transformers in finance for sentence and/or document-level tasks.

Issam et al. (2022) develop a system to predict finance and non-finance tweets using CNN and LDA algorithms. Their model consists of data filtering, topic modeling by LDA, and prediction by CNN. They collect and filter Twitter data from June 2019 to June 2020, resulting in a dataset of 1,000,000 tweets. CNN achieves 99% accuracy, outperforming other classifiers. The study highlights CNN's effectiveness in sentiment analysis and its potential application in predicting stock market trends. Future work includes incorporating finance-related articles and reports and developing applications for stock price prediction.

Ahmad and Umar (2023) analyze financial textual data using machine learning and deep learning models. They compare MNB and LR as machine learning models with RNN, LSTM, and GRU as deep learning models. Preprocessing steps filter inappropriate texts, resulting in a dataset of 70,000 records categorized into positive, negative, and neutral sentiments. LSTM and GRU achieve higher accuracy than MNB and LR, highlighting advanced algorithms' effectiveness in sentiment analysis. The study underscores preprocessing steps' importance in improving model accuracy and suggests further optimization using optimization algorithms like Swarm or Bat for future work.

Ivanenko (2023) introduces an advanced Aspect-Based Financial Sentiment Analysis (ABFSA) framework utilizing contrastive learning. DeBERTa v3 and 2CL contrast learning are integrated into the methodology to enhance sentiment analysis granularity, specifically tailored for financial news within the SEntFiN dataset. Results reveal significant improvements in sentiment classification granularity and accuracy. The study underscores the potential of specialized ABFSA methodologies augmented with advanced NLP techniques for nuanced sentiment representation in financial narratives, offering actionable insights for stakeholders.Future directions include multilingual extensions for global markets and real-time trading validation. Expanding training data diversity and ethical model deployment guidelines are recommended, shaping future research in interdisciplinary financial sentiment analysis.

Esichaikul and Phumdontree (2018) propose SentiFine, integrating deep learning with fine-grained sentiment analysis for Thai financial news. The framework includes data acquisition, preprocessing, feature transformation, and model inference, leading to the development of a web-based system. Experiments with 17 deep neural network models identify United CNN Bidirectional GRU (UCBGRU) as optimal. The SentiFine system enables sentiment indicator viewing and daily news export. Future work includes real user system evaluation, ensuring practical FinTech application. SentiFine presents a novel approach for Thai financial sentiment analysis, leveraging deep learning for enhanced accuracy.

Sharma et al. (2023) introduce a supervised machine learning approach for ontology-based stock market sentiment prediction. Combining Convolutional Neural Networks (CNN) with the Artificial Bee Colony (ABC) algorithm addresses existing gaps, yielding significant performance improvements. Using 15,000

Stocktwits samples, sentiment classification achieves notable accuracy. The study advances sentiment analysis in stock markets, suggesting broader applications in financial forecasting. Future research might explore country-specific factors and further enhance prediction accuracy through time-specific determinants.

Zhang et al. (2023) propose a retrieval-augmented Large Language Models (LLMs) framework for financial sentiment analysis. Aligning LLMs with financial sentiment tasks and retrieving additional context from reliable sources enhance predictive accuracy. Training on a dataset amalgamated from Twitter Financial News and FiQA datasets, the framework outperforms traditional models and LLMs like ChatGPT and LLaMA. Future work could integrate economic data for comprehensive analysis, ensuring robustness in financial sentiment assessment.

Du et al. (2024) provide a comprehensive overview of Financial Sentiment Analysis (FSA), highlighting trends in techniques and applications. Insights into sentiment analysis scopes, trends, and challenges inform future research directions. Challenges include dataset availability and lexical resource limitations, with future directions emphasizing knowledge incorporation and multimodal data integration for enhanced FSA. The study sets a roadmap for advancing FSA, crucial for informed financial decision-making.

Jabeen et al. (2021) propose an LSTM-based model integrating sentiment analysis of coronavirus events for stock market forecasting. The framework demonstrates improved prediction accuracy, emphasizing the influence of social media sentiment on financial markets. Future research directions include multi-event sentiment analysis and model complexity enhancement for real-time applications.

Kansal and Kumar (2019) present a hybrid approach integrating artificial intelligence and cuckoo search optimization for financial sentiment analysis. The methodology involves ANN-based sentiment classification optimized by the cuckoo search algorithm. The study underscores the significance of machine learning in NLP domains and suggests further optimization variants for enhanced accuracy. Future research could explore challenges such as sarcasm and irony in textual data analysis. Kirchner explores sentiment analysis in financial domains, emphasizing lexicon-based approaches. The study investigates sentiment analysis methodologies, proposing a model to enhance granularity and effectiveness in financial sentiment analysis. The study underscores sentiment analysis' potential in finance and identifies avenues for refining sentiment scoring methods and analyzing sarcasm impact.

Memi¸s et al. (2024) conducted a study on sentiment analysis of financial tweets, focusing on Turkish language data. They collected tweets related to finance and manually labeled them as positive, negative, or neutral. Employing deep learning algorithms such as Neural Network, CNN, LSTM, GRU, and GRU-CNN, along with word embedding techniques, they achieved promising results. The study emphasizes the importance of sentiment analysis in financial contexts and suggests avenues for future research, including model enhancement and dataset expansion.

Kim et al. (2023) investigated the integration of sentiment analysis into stock price prediction models using a mathematical framework. They utilized FinBERT, a transformer model tailored for financial language, to analyze sentiment in news summaries from The New York Times. By incorporating sentiment scores into an LSTM model for stock price prediction, they observed improved accuracy compared to models without sentiment analysis. Their study highlights the significance of sentiment in financial

markets and underscores the potential of mathematical-based sentiment analysis to enhance stock price forecasting.

Du, Tsai and Wang (2019) presented a paper titled "Beyond Word-Level to Sentence-Level Sentiment Analysis for Financial Reports." Their research aimed to delve into sentence-level sentiment analysis within financial reports, focusing on identifying financial risk. They proposed a methodology for generating financial sentiment phrases (senti-phrases) and employed various sentence embedding models to enhance representation learning of financial risk sentences. Utilizing the 10-K corpus, they constructed a sentence-level risk-labeled dataset from annual SEC-mandated financial reports, specifically the "management's discussion and analysis of financial conditions and results of operations" (MD&A) section. The study underscored the importance of sentence-level sentiment analysis for financial risk assessment, offering insights into constructing financial sentiment lexicons and advancing sentiment analysis in finance.

Bressanelli (2022) thesis project explored sentiment analysis of financial news during the Covid-19 pandemic across traditional media and Twitter. Through sentiment analysis models like VADER and FinBERT, the study examined correlations between news sentiment and S&P 500 movements. While both models showed correlations, FinBERT exhibited stronger associations, particularly with financial tweets. Despite data quality challenges and tweet complexity, the study shed light on sentiment's impact on the stock market during crises, offering insights into sentiment's role in financial news and social media.

Yıldırım et al. (2019) compared deep learning (DL) and traditional machine learning (ML) models for sentiment analysis on StockTwits microblog data. DL models, especially LSTM variants, outperformed ML classifiers like Naive Bayes and Random Forest. LSTM's sequential data understanding capabilities contributed to its superior performance. Although dropout mechanisms didn't significantly enhance DL model success rates, they helped reduce biases. The study underscores DL's effectiveness in analyzing financial microblog data, offering insights into stock market forecasting using sentiment analysis.

Yekrangi and Nikolov (2023) addressed domain-specific sentiment analysis in financial markets through an optimized deep learning approach. Their research compared embedding techniques and classification algorithms for sentiment analysis, showing that fine-tuned embedding layers, especially with LSTM models, outperformed pretrained embeddings. The study emphasized the importance of domain-specific embeddings and proposed a human-in-the-loop approach for generating labeled datasets. While focusing on sentiment analysis in finance, the methodology's adaptability to other domains suggests broader applications, contributing to sentiment analysis advancements.

Atak (2023) explored sentiment analysis in financial disclosures within the Borsa Istanbul Stock Exchange. Leveraging deep learning and NLP techniques, the study analyzed sentiment from annual reports to understand its impact on firm value perceptions. By bridging sentiment indices with market dynamics, the research contributed to theoretical knowledge and practical applications in corporate communication practices. Insights gleaned from the study offer valuable guidance for stakeholders, policymakers, and regulators in capital markets, emphasizing the importance of effective information disclosure for informed decision-making.

Syeda (2022) focused on sentiment analysis of financial news using supervised learning techniques. By comparing traditional classifiers like Naive Bayes with transfer learning models like BERT and FinBERT, the study highlighted FinBERT's superior performance. The discussion delved into model performance

differences, correlating sentiment with stock prices and suggesting avenues for future work. The research advances understanding of sentiment analysis in financial markets and its implications for stock price prediction.

Hasselgren et al. (2022) explored utilizing sentiment data from social media (SM) for investment decisions. Their prototype integrated SM metrics into sentiment scores, demonstrating potential utility in investment decisions. While acknowledging limitations, the study laid a foundation for future research in utilizing SM sentiment for investment decision-making. Recommendations for integrating stock market data into sentiment charts and updating the system for multilingual support offer avenues for further exploration and improvement.

Karanikola et al. (2023) conducted a comprehensive study comparing classic machine learning (ML) algorithms with contemporary deep learning pre-trained models for financial sentiment analysis. They evaluated the performance of BERT, RoBERTa, and three variants of FinBERT against various ML algorithms using a merged dataset comprising the FiQA dataset and the Financial PhraseBank. Notably, pre-trained models like RoBERTa and BERT demonstrated superior performance over classic ML methods. Surprisingly, the FinBERT variants did not outperform these pre-trained models, indicating the need for further investigation. The study also highlighted the effectiveness of using TF-IDF representations in conjunction with the SMOTE data augmentation technique. Future research avenues may include exploring ensemble models that combine classic and contemporary pre-trained models to enhance sentiment analysis in financial contexts.

Alanazi et al. (2023) embarked on a study to monitor public mental health through sentiment analysis of financial text using machine learning techniques. Their research focused on analyzing finance-related content from The Guardian newspaper and employed support vector machine (SVM), AdaBoost, and single-layer convolutional neural network (SLCNN) techniques. Among these methods, the SLCNN approach achieved the highest classification accuracy, outperforming SVM and AdaBoost. The study provides valuable insights into the relationship between financial news sentiment and public mental health, with potential implications for policymaking and mental health interventions. However, the authors acknowledged limitations such as the dataset's size and timeframe, suggesting avenues for future research to address these constraints and further explore the impact of financial news on public well-being.

Patel (2021) delved into justification mining, aiming to develop a machine learning method to identify representative sentences and summarize sentiment in financial text, particularly 10-K filings. This approach addressed the challenges of AI explainability and model interpretability, crucial in finance. The study evaluated transfer learning methods and introduced justification mining using transformer models and clustering algorithms. Results indicated successful transfer learning and the superiority of transformer models, highlighting potential applications in understanding machine learning models. By offering easily interpretable information, this research contributed to AI explainability and model interpretability in finance, aiding in making more informed decisions.

Vicari and Gaspari (2021) investigated trading on news sentiment using deep learning (DL) models, focusing on LSTM networks. They analyzed daily news headlines from 2008 to 2016, extended to 2020, aiming to predict Dow Jones industrial average (DJIA) movements. Preprocessing involved merging and vectorizing news strings. However, the DL model did not outperform random chance, cautioning against sole reliance on DL methods for financial forecasting. The study acknowledged limitations and suggested

cautious handling of such models. Despite the outcomes, it provided insights for future research and underscored the complexity of financial sentiment analysis.

Sohangir et al. (2018) explored Deep Learning models for sentiment analysis on StockTwits, comparing doc2vec, LSTM, and CNN. Doc2vec showed limited improvement, while LSTM outperformed but fell short of desired accuracy. CNN exhibited superior sentiment analysis performance, with accuracy exceeding 90% after 10,000 steps. The study highlighted CNN's efficacy in extracting sentiment from financial texts, offering insights into improving sentiment analysis tasks in finance.

Kohsasih et al. (2022) proposed a sentiment analysis model for financial news using RNN-LSTM architecture. Their study emphasized the significance of Financial Sentiment Analysis (FSA) in investment decisions and aimed to improve existing models. The RNN-LSTM architecture demonstrated significantly higher performance metrics compared to other models, offering insights into sentiment trends and aiding in making informed investment decisions.

Methmal (2020) focused on sentiment analysis for financial market prediction using deep learning techniques. By analyzing microblog messages from social media platforms, sentiment analysis models were compared, showing significant improvements over baseline models. Despite challenges like working with a small dataset, the research achieved superior results with deep learning models compared to traditional machine learning approaches. This study contributed to financial market prediction by demonstrating the effectiveness of sentiment analysis using deep learning techniques, offering valuable insights for making informed investment decisions.

2.7.1 Research Findings:

Deep Learning Models: Deep learning models, particularly CNN and LSTM, have shown promising results in sentiment analysis on financial data. Sohangir et al. (2018) found that CNN exhibited superior sentiment analysis performance, with accuracy exceeding 90% after 10,000 steps. Kohsasih et al. (2022) found that the RNN-LSTM architecture demonstrates significantly higher performance metrics compared to other models.

Pre-trained Models: Pre-trained models like BERT and RoBERTa have shown superior performance in financial sentiment analysis.

Advanced NLP and Machine Learning Techniques: These techniques, including deep learning models, have shown promising results in sentiment analysis, particularly in the financial domain.

Preprocessing Techniques: Preprocessing techniques tailored for financial texts, such as entity-sensitive sentiment analysis and fine-grained sentiment classification, enhance the granularity and accuracy of sentiment analysis models.

Hybrid Approaches: Hybrid approaches combining machine learning algorithms with optimization techniques, such as the cuckoo search algorithm, improve feature selection and model performance in sentiment analysis tasks.

Context Augmentation: Sentiment analysis models augmented with external context from reliable sources, such as retrieval-augmented large language models, demonstrate improved predictive accuracy in financial sentiment analysis.

Incorporation of Sentiment Analysis**:** Incorporating sentiment analysis into stock market forecasting models, particularly through the integration of sentiment scores with deep learning architectures like LSTM, enhances prediction accuracy and provides valuable insights for investors and analysts.

2.7.2 Research Gaps and Challenges:

Language Limitations: Many studies focus on sentiment analysis in English-language financial texts, leaving a gap in research regarding sentiment analysis in other languages, particularly in emerging markets.

Data Integration: Limited attention is given to the integration of sentiment analysis with other data sources such as economic indicators or market sentiment indices, presenting an opportunity for more comprehensive forecasting models.

Methodology Limitations: Existing methodologies often rely on pre-defined lexicons or datasets, which may not adequately capture the nuances of sentiment in financial texts, indicating a need for more adaptable and context-aware approaches.

Data Source Limitations: The majority of studies focus on sentiment analysis of textual data from news articles and social media, neglecting other valuable sources such as corporate reports, earnings calls, or analyst notes.

Interpretability and Explainability: While deep learning models show promising results, their interpretability and explainability remain challenging, hindering their adoption in real-world financial decision-making processes.

Accuracy Limitations: There is limited accuracy in predicting financial market prices using existing sentiment analysis techniques.

Linguistic Challenges: There are challenges in handling complex linguistic issues, particularly in microblog datasets like Twitter.

Dataset Size Limitations: Limited dataset sizes may hinder the performance of machine learning models.

Feature Engineering Limitations: There is a lack of comprehensive feature engineering methods tailored to financial domain-specific data.

Dynamic Nature of Financial Markets: There is difficulty in achieving high accuracy due to the dynamic nature of financial markets.

2.7.3 Summary and Future Directions:

Promising Techniques: Advanced NLP and machine learning techniques, including deep learning models, have shown promising results in sentiment analysis, particularly in the financial domain.

Preprocessing Techniques: Preprocessing techniques tailored for financial texts, such as entity-sensitive sentiment analysis and fine-grained sentiment classification, enhance the granularity and accuracy of sentiment analysis models.

Hybrid Approaches: Hybrid approaches combining machine learning algorithms with optimization techniques, such as the cuckoo search algorithm, improve feature selection and model performance in sentiment analysis tasks.

Context Augmentation: Sentiment analysis models augmented with external context from reliable sources, such as retrieval-augmented large language models, demonstrate improved predictive accuracy in financial sentiment analysis.

Incorporation of Sentiment Analysis: Incorporating sentiment analysis into stock market forecasting models, particularly through the integration of sentiment scores with deep learning architectures like LSTM, enhances prediction accuracy and provides valuable insights for investors and analysts.

Future Directions: There is a need for further investigation into why FinBERT variants did not outperform pre-trained models like BERT and RoBERTa. There is also a need to explore ensemble models that combine classic and contemporary pre-trained models to enhance sentiment analysis in financial contexts.

Overall, the studies highlight the importance of sentiment analysis in financial market prediction and investment decisions. While traditional machine learning models have been used, there is a growing interest in leveraging deep learning architectures for improved accuracy. Feature engineering, dataset preprocessing, and ensemble techniques play crucial roles in enhancing model performance. However, several challenges and gaps remain, presenting opportunities for future research in this field.

3 METHODOLOGY

This research employs a comprehensive methodology that builds upon the foundations established by prior studies in the field of Financial Sentiment Analysis (FSA). It adopts a systematic approach to tackle identified challenges and capitalize on opportunities highlighted in the existing body of literature.

The methodology integrates traditional machine learning techniques such as Logistic Regression (LR), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost). These techniques have proven their effectiveness in previous studies and continue to be valuable tools in sentiment analysis.

Recognizing the rapid advancements in the field, this research also acknowledges the need for further exploration of state-of-the-art models like Bidirectional Encoder Representations from Transformers (BERT) and its financial variant, FinBERT. These transformer-based models have shown promising results in various Natural Language Processing (NLP) tasks, including sentiment analysis, and their potential contribution to FSA is an exciting prospect worth investigating.

Through this advanced methodology, the research aims to not only contribute to the existing knowledge base but also pave the way for future studies in Financial Sentiment Analysis. By continuously integrating and evaluating new techniques and models, it seeks to stay at the forefront of this rapidly evolving field.

3.1 Rationale for Methodology

The rationale for the chosen methodology is deeply rooted in a thorough review of existing literature in the field of Financial Sentiment Analysis (FSA), with particular emphasis on the works of Karanikola et al. (2023) and Ahmad and Umar (2023). These studies underscore the need for an in-depth investigation into advanced transformer-based models such as BERT, FinBERT, and RoBERTa, and the importance of sophisticated preprocessing steps.

This research aims to address the challenges identified in previous studies and seize emerging opportunities by synthesizing insights from these foundational works. The selected methodology, characterized by a blend of traditional machine learning techniques such as Logistic Regression (LR),

Support Vector Machines (SVM), and XGBoost, is complemented by an exploration of advanced transformer-based models including BERT, FinBERT, and RoBERTa. This balanced approach is designed to enhance the accuracy and robustness of sentiment analysis in financial contexts.

Each component of the methodology is carefully chosen to align with the research objectives and contribute to the advancement of knowledge in the field. The rationale behind each component is elucidated, demonstrating how it builds upon the findings of Karanikola et al. (2023), and Ahmad and Umar (2023), and how it aims to fill the gaps they identified. This comprehensive approach not only ensures the robustness of the research but also contributes to the ongoing evolution of methodologies in the field of Financial Sentiment Analysis.

3.2 Proposed Model Design

The design of the proposed model involves several stages, each contributing to the overall performance and interpretability of the final model.

Data Collection and Preprocessing: The initial dataset was sourced from Kaggle. Preprocessing steps included cleaning the data, checking for missing values, removing duplicates, and tokenizing the text. Additional text processing techniques such as converting to lowercase, removing stopwords, and lemmatization were also applied to prepare the data for modeling.

Exploratory Data Analysis (EDA): EDA was conducted to gain insights into the data and inform the modeling process. This involved two methods: a traditional Machine Learning (ML) method and a Natural Language Processing (NLP) model method.

Feature Engineering and Selection: For the traditional ML method, feature engineering and selection were performed using Bag of Words (BOW) with SelectKBest and Term Frequency-Inverse Document Frequency (TF-IDF) with SelectKBest.

Model Training and Optimization: Several models, including Logistic Regression (LR), Support Vector Machines (SVM), and XGBoost, were tuned, optimized, and trained using both BOW and TF-IDF methods. These models were then combined using a Voting Classifier to create an ensemble model (lr_combined for LR, and similarly for the other two models).

Model Evaluation and Interpretation: The models were evaluated using various metrics, including confusion matrices, accuracy, precision, recall, F-Score, and Area Under the Curve (AUC). Model interpretation was performed using Local Interpretable Model-Agnostic Explanations (LIME).

NLP Model Method: In addition to the traditional ML method, advanced NLP models such as BERT, FinBERT, and RoBERTa were also tuned, optimized, and trained. These models were evaluated and interpreted using the same metrics and methods as the traditional ML models.

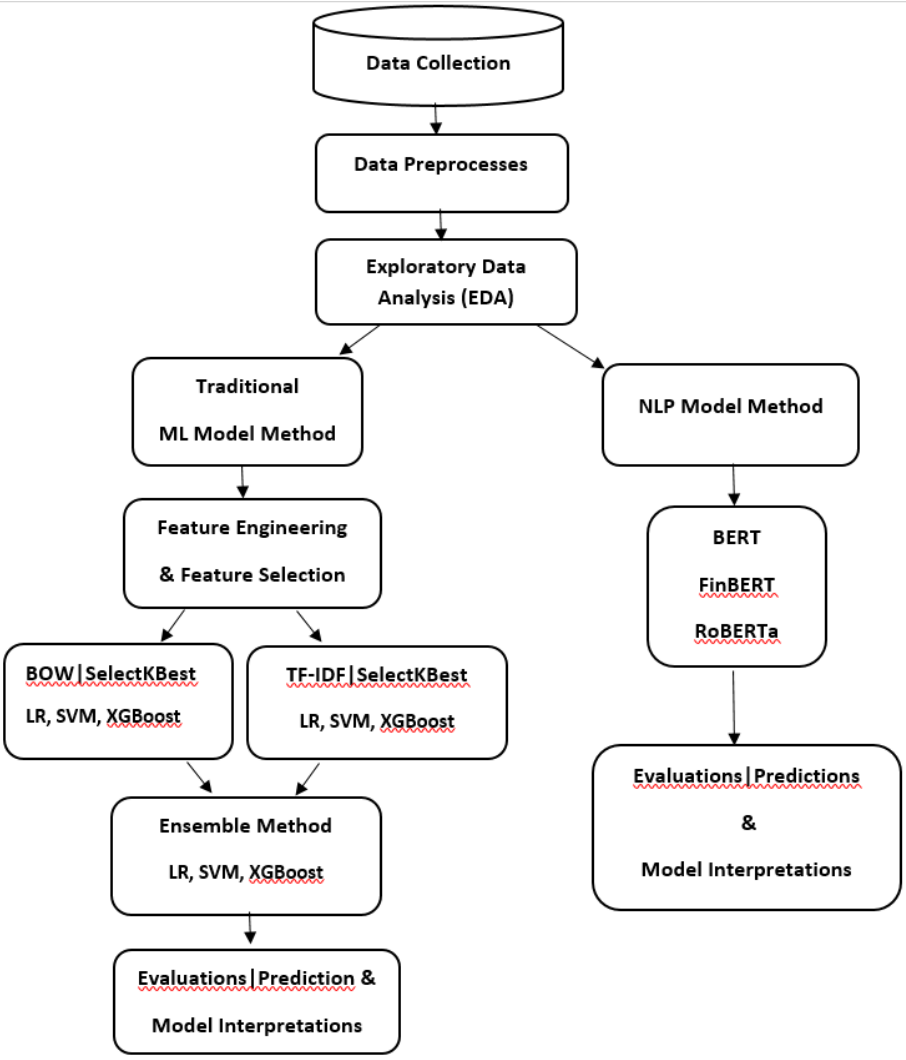


Fig. 2. Proposed Design Flowchart

4 IMPLEMENTATION OF THE MODELS

Implementing machine learning models for financial sentiment analysis involved a structured approach, as outlined in the proposed model design. This section elaborates on the steps taken to train and evaluate various models, providing insights into their performance across key metrics.

4.1 Data Description

The data used for the implementation of the models is derived from three distinct datasets and was sourced from Kaggle, a renowned platform for machine learning and data science competitions. The datasets utilized are as follows:

FiQA and Financial PhraseBank Dataset: This dataset, a combination of the FiQA (Financial Opinion Mining and Question Answering) and Financial PhraseBank datasets, consists of 5,842 entries. Each entry includes a sentence (text data) and a sentiment label (either 'neutral', 'positive', or 'negative'). The sentiment distribution is as follows: 3,130 neutral, 1,852 positive, and 860 negative sentiments. The FiQA dataset

was created in 2018 to improve aspect-based sentiment analysis Maia, Macedo & Handschuh et al. (2018), while the Financial PhraseBank dataset contains 4,840 sentences derived from the Lexis Nexis database, labeled by domain-expert annotators Malo, Pekka, et al. (2014).

SentFIN v1.1 Dataset: This dataset contains 10,753 news headlines, each annotated with sentiment labels for all financial entities appearing in the headlines. The dataset is balanced with 4,100 positive entities, 3,200 negative entities, and 4,500 neutral entities. It is particularly useful for Aspect-based Sentiment Analysis and can also be used for training models for extracting named entities.

Combined Dataset: The combined dataset merges the above datasets into a single DataFrame with 16,595 entries. Each entry includes a sentence (text data) and a sentiment label (either 'neutral', 'positive', or 'negative'). The sentiment distribution in this combined dataset is as follows: 7,665 neutral, 5,215 positive, and 3,715 negative sentiments.

These datasets provide a comprehensive and diverse collection of financial texts, offering a rich resource for training and evaluating the proposed models for financial sentiment analysis.

4.2 Data Preprocessing

Data preprocessing is a crucial step in preparing the financial sentiment dataset for analysis. The process involves cleaning and transforming the raw text data to make it suitable for model training and evaluation

4.2.1 Data Cleaning

The data cleaning process is a crucial step in preparing the dataset for model implementation. This process involved checking for missing values and duplicates in the dataset.

Checking for Missing Values: The dataset was examined for any missing values. The Python code snippet sentiment_df.isna().sum() was used to calculate the sum of all missing values in the dataset. The output confirmed that there were no missing values in the 'Text' and 'Sentiment' columns of the dataset.

Checking for Duplicates: The dataset was also checked for any duplicate entries using the Python code snippet sentiment_df.duplicated().sum(). The output revealed that there were 64 duplicate rows in the DataFrame, which constituted approximately 0.004% of the data. Given the negligible proportion of duplicates, the decision was made to remove these entries from the dataset.

4.2.2 Text Preprocessing

Text preprocessing is a crucial step in preparing the financial sentiment dataset for analysis. The following steps were undertaken to clean and preprocess the textual data:

Lowercasing: All text data was converted to lowercase to ensure consistency in the dataset. This helps in standardizing the text and avoids treating the same words differently due to case variations.

Punctuation Removal: Punctuation marks were removed from the text while retaining numbers. Punctuation marks do not usually contribute to sentiment analysis and are therefore removed to focus on meaningful words.

Tokenization: Tokenization was performed to split the text into individual words or tokens. This step breaks down the text into smaller units, allowing for further analysis at the word level.

Stopword Removal: Stopwords, such as "is," "and," "the," etc., were removed from the text. Stopwords are common words that do not carry significant meaning in sentiment analysis and can be safely discarded to reduce noise in the data.

Lemmatization: Lemmatization was applied to reduce words to their base or root form. This step helps in standardizing words by converting them to their dictionary form, thereby reducing the vocabulary size and improving the model's generalization capability.

Filtering Short Texts: Rows containing only one word in the 'Text' column were removed. Texts with very few words may not provide sufficient context for sentiment analysis and are therefore excluded from the dataset.

Encoding Sentiment Labels: The 'Sentiment' column was encoded into numerical labels using LabelEncoder. This step is necessary as machine learning models require numerical inputs for training.

Drop Original Sentiment Column**:** The original 'Sentiment' column was dropped from the dataset after encoding to avoid redundancy and streamline the data for model training.

```
# Initialize lemmatizer
lemmatizer = WordNetLemmatizer()

# Define a function for text preprocessing
def preprocess_text(text):
    # Convert text to lowercase
    text = text.lower()
    # Remove punctuation but keep numbers
    text = ''.join(ch for ch in text if ch not in string.punctuation)

    # Tokenization
    words = nltk.word_tokenize(text)

    # Stopword removal
    stop_words = set(stopwords.words('english'))
    words = [word for word in words if word not in stop_words]

    # Lemmatization
    words = [lemmatizer.lemmatize(word) for word in words]

    return ' '.join(words)

# Apply the function to the 'Text' column
sentiment_df['Text'] = sentiment_df['Text'].apply(preprocess_text)

# Remove rows where the 'Text' column has just one word
sentiment_df = sentiment_df[sentiment_df['Text'].apply(lambda x: len(x.split()) > 1)]
# Encode Sentiment Column
label_encoder = LabelEncoder()
sentiment_df['Sentiment_encoded'] = label_encoder.fit_transform(sentiment_df['Sentiment'])

# Remove the original Sentiment column
sentiment_df.drop(columns=['Sentiment'], inplace=True)
```

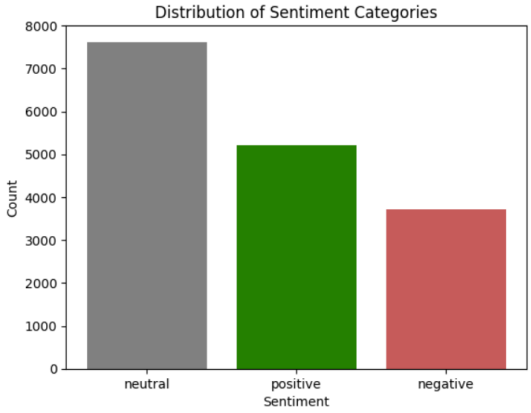
Each preprocessing step was carefully chosen to enhance the quality of the textual data and prepare it for subsequent analysis and model training.

4.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in the data science process as it allows us to understand the structure of our data and to extract meaningful insights. In this project, EDA was conducted on the financial sentiment dataset, focusing on the distribution and proportion of sentiment categories and the most common words associated with each sentiment.

Distribution of Sentiment Categories: A bar chart was used to visualize the number of sentences for each sentiment category. The chart revealed that the 'neutral' category had the highest count, followed by 'positive' and 'negative'.

Proportion of Sentiment Categories: A pie chart was used to display the proportion of each sentiment category in the dataset. The chart showed that 'neutral' sentiments made up the largest proportion, followed by 'positive' and 'negative' sentiments.



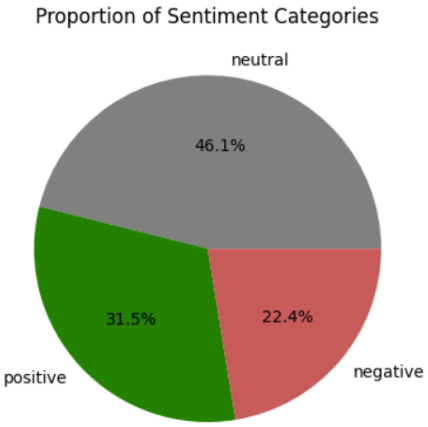


Fig. 2. Distribution of Sentiment Categories Fig. 3. Proportion of Sentiment Categories

Word Clouds for Each Sentiment Category: Word clouds were generated for each sentiment category to visualize the most common words associated with 'positive', 'negative', and 'neutral' sentiments. This provided an intuitive understanding of the words that are most representative of each sentiment.

Fig. 4. WordCloud of Sentiment Categories

4.4 Feature Engineering

Feature engineering is a pivotal step in the preprocessing pipeline, aiming to extract meaningful features from raw text data. In our analysis, we employed two popular techniques: Bag-of-Words (BoW) representation and TF-IDF representation.

For BoW representation, we utilized the CountVectorizer from scikit-learn to convert text documents into a matrix of token counts. This technique captures the frequency of occurrence of each word in the corpus, generating a sparse matrix representation. Following BoW conversion, we applied feature selection using SelectKBest with the chi-square test as the scoring function. The optimal value of k (number of selected features) was determined through experimentation, yielding an optimal score with k=2100. Finally, the dataset was split into training and testing subsets using train_test_split from scikit-learn.

Similarly, for TF-IDF representation, we employed the TfidfVectorizer from scikit-learn to transform text documents into a matrix of TF-IDF features. This method assigns weights to terms based on their frequency in the document and across the corpus, emphasizing terms that are unique to specific documents. Feature selection was again performed using SelectKBest with the chi-square test, and the dataset was split into training and testing sets.

4.5 Feature Selection

Feature selection is essential for enhancing model performance and reducing computational complexity by identifying the most relevant features for prediction. In our analysis, we employed the chi-square test-based feature selection method, implemented through SelectKBest from scikit-learn.

For both BoW and TF-IDF representations, we applied SelectKBest to select the top k features that exhibit the strongest association with the target sentiment labels. Through experimentation, we determined the optimal value of k to be 2100, achieving an optimal balance between predictive power and computational efficiency. This process helped streamline the feature space, retaining only the most informative features for subsequent model training and evaluation.

4.6 Data Splitting

Data splitting is a fundamental step in machine learning workflows, facilitating the assessment of model performance on unseen data. In our analysis, we partitioned the sentiment dataset into separate training and testing subsets using train_test_split from scikit-learn.

For both BoW and TF-IDF representations, the dataset was split into training and testing sets, with 80% of the data allocated for training and 20% for testing. This partitioning ensured that models were trained on a sufficient amount of data to learn meaningful patterns while also allowing for robust evaluation on unseen data. The random_state parameter was set to 42 to ensure reproducibility across different runs of the experiment.

4.7 Machine Learning Models

Machine learning models play a crucial role in sentiment analysis tasks, providing predictive capabilities to classify text data into sentiment categories. In this section, we explore various machine learning and natural language processing (NLP) models, each with its unique characteristics and performance.

4.7.1 Hyper-Parameter Tuning

Hyper-parameter optimization is a vital process in refining machine learning models. This process involves the careful selection of an optimal set of hyperparameters to boost the model's performance. Commonly employed techniques encompass grid search, random search, Bayesian optimization, and gradient-based optimization. The objective is to achieve a harmonious balance between model complexity and its capacity to generalize, ensuring the models perform at their peak in real-world scenarios and effectively interpret unseen data. The evaluation of hyper-parameter tuning typically hinges on performance metrics such as accuracy, precision, recall, and F1-score.

In our analysis, we employed hyper-parameter tuning to optimize the performance of each machine learning and NLP model. By fine-tuning the hyper-parameters, we aimed to enhance model performance, improve generalization, and mitigate issues such as overfitting or underfitting.

4.7.2 Logistic Regression (LR)

Logistic Regression, a statistical method predominantly used for binary classification tasks, establishes a relationship between a dependent variable and one or more independent variables. It does this by estimating probabilities using a logistic function, which ranges from 0 to 1. In cases with multiple outcomes, it employs multinomial logistic regression, while for ordered categories, it utilizes ordinal logistic regression (Sanjay, 2017).

This method is highly valued for its versatility and interpretability, making it a useful tool in evaluating the impact of features on outcomes. Despite its linear approach, its simplicity and effectiveness have led to its widespread adoption across various domains, including finance, healthcare, and marketing. The logistic function as;

$$P = (y = 1|X) = \frac{1}{1 + e^{-wa}}$$

where 'e' is the numerical constant Euler's number, 'w 'is a constant, and 'a' is an input we put into the function.

In this study, we trained Logistic Regression (LR) models on Bag-of-Words (BoW) and TF-IDF representations of sentiment data. By tuning hyperparameters like regularization strength (C), we optimized model performance. LR models provide interpretable coefficients for features, aiding in identifying influential words for sentiment prediction. We also utilized ensemble techniques, like a VotingClassifier, to combine LR model predictions for improved accuracy

4.7.3 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a robust supervised learning technique used for both regression and classification tasks Hearst et al. (1998). It includes the Support Vector Regression (SVR) for regression problems and the Support Vector Classification (SVC) for classification tasks. Specifically, SVC identifies the best hyperplane for binary and multi-class situations, forming a linear or non-linear boundary in the input space. Kernels, which are frequently linked with Support Vectors, are used to establish the separation function. Jim et al. (2024). The formulation of SVM is given below:

$$f(x) = \sum_{x_j \in S} y_j k(x_j, x) + b \quad (1)$$

In Eq. (1), x_j signifies training patterns, y_j represents class labels ($y_j \in \{+1, -1\}$), and $\mathcal{S}$ is set of Support Vectors. The dual formulation leads to a minimization problem expressed as:

$$\min_{0 \le i \le C} W = \frac{1}{2}\sum_{ij} Q_{ijj} - \sum_{i} + b \sum_{i} y_{ii} \tag{2}$$

In Eq. (2), $Q_{ij} = y_i y_j K(x_i, x_j)$ represents symmetric positive kernel matrix. C penalizes error values, and the conditions for dual are:

$$g_i = \frac{\partial W}{\partial_i} = \sum_{i} Q_{ijj} + y_i b - 1 = y_i f(x_i) - 1 \tag{3}$$

$$\frac{\partial W}{\partial b} = \sum_{i} y_j = 0 \tag{4}$$

This categorizes the training set into the Support Vector set ($0 < i < C$, $g_i = 0$), the well-classified set ($i = 0$, $g_i > 0$) and the error set ($i = C$, $g_i < 0$). Now, introducing a quadratic penalty factor C_0 for the error points results in a modified kernel function, which turns the problem into a linearly separable case:

$$K_0(x_i, x_j) = K(x_i, x_j) + \frac{1}{C_0}\delta_{ij} \tag{5}$$

SVM was employed for sentiment classification tasks, utilizing a range of kernels including linear, polynomial, and radial basis function (RBF) kernels. We conducted hyper-parameter tuning to optimize parameters like the regularization parameter (C) and kernel parameters (gamma for RBF kernel, degree for polynomial kernel). Additionally, we utilized ensemble techniques, such as a VotingClassifier, to combine SVM model predictions to improve the robustness and accuracy of our sentiment classification tasks.

4.7.4 XGBoost

XGBoost, also known as Extreme Gradient Boosting, is a machine learning algorithm that falls under the category of ensemble methods based on decision trees. It operates within the framework of Gradient Boosting, using decision trees effectively to improve predictive performance. The term “XGBoost” is an abbreviation for “Extreme Gradient Boosting” (Chen et al., 2015).

Ensemble learning is a technique that harnesses the predictive power of multiple learners. In the boosting approach, trees are constructed sequentially, with each subsequent tree focusing on reducing the errors made by its predecessor. Each tree learns from the ones before it and adjusts by addressing the remaining errors. As a result, the next tree in the sequence learns from the adjusted residuals Jim et al. (2024).

Typically, boosting uses weak learners as base models, which are known for their high bias. These weak learners collectively contribute valuable insights into predictions, enabling the boosting technique to construct a robust learner by effectively combining the abilities of these individual weak learners. Assume a training dataset is denoted as x_t and their corresponding labels y_t. In the XGBoost algorithm, a classifier generates the final prediction $\hat{y}_l^t$ representing the predicted value or label for a given input x_t

$$\hat{y}_l^t = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_l^{t-1} + f_t(x_i) \tag{6}$$

Where $\hat{y}_l^{t-1}$ represents the prior prediction and $f_t(x_i)$ is denotes the new prediction. The objective in XGBoost is to minimize the following objective function to attain a high-quality model.

$$L^t = \sum_{l=1}^{n} l(y_i, \hat{y}_l) + \Omega(f_t) \tag{7}$$

Here, objective function incorporates loss function $l(y_i, \hat{y}_l)$ and regularization term $\Omega(f_t)$. With the existence of Eq. (6). Now, rewrite the objective function as follows.

$$L^t = \sum_{l=1}^{n} \left(l\left(y_i, \hat{y}_l^{t-1}\right) + f_t(x_i)\right) + \Omega(f_t) \tag{8}$$

The loss function is a measure to evaluate how effectively the model aligns with the training data, while regularization evaluates the complexity of the decision trees. The optimization of the loss function seeks to generate predictive models with enhanced accuracy, whereas optimizing regularization promotes the development of simpler and more generalized models.

XGBoost plays a pivotal role in sentiment classification tasks, capitalizing on its robust implementation and adeptness in handling high-dimensional feature spaces. We conducted hyper-parameter tuning to optimize crucial parameters such as the learning rate, maximum depth of trees, and regularization parameters. Additionally, we leveraged ensemble techniques, like the VotingClassifier, to amalgamate predictions from multiple XGBoost models for enhanced accuracy.

4.7.5 Ensemble Method

This study employed ensemble methods to combine the predictions of multiple models for sentiment classification tasks. This approach was applied to Logistic Regression (LR), Support Vector Machine (SVM), and XGBoost models.

For the Logistic Regression models, we defined base models using the best estimators from Bag-of-Words (BoW) and TF-IDF representations. These models were combined using a VotingClassifier with 'soft' voting. The ensemble was evaluated using cross-validation, and the trained ensemble achieved an accuracy score on the test data.

For the Support Vector Machine models, we defined base models using the best parameters from the grid search. These models were also combined using a VotingClassifier with 'soft' voting. The ensemble was evaluated using cross-validation, and the trained ensemble achieved an accuracy score on the test data.

For the XGBoost models, we defined base models using the best estimators from Bag-of-Words (BoW) and TF-IDF representations. These models were combined using a VotingClassifier with 'soft' voting. The ensemble was evaluated using cross-validation, and the trained ensemble achieved an accuracy score on the test data.

The ensemble method allowed us to leverage the strengths of multiple models, improving the robustness and accuracy of our sentiment classification tasks. The cross-validation scores and test scores for each ensemble provide a comprehensive evaluation of their performance.

4.8 NLP Models

Natural Language Processing (NLP) models leverage advanced techniques to analyze and understand human language, making them well-suited for sentiment analysis tasks. In this section, we explore three state-of-the-art NLP models: BERT, FinBERT, and RoBERTa, each offering unique capabilities and performance.

4.8.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language representation model that leverages the Transformer architecture. It's a pre-trained model capable of capturing the contextual relationships and meanings of words in sentences, and can be fine-tuned for specific NLP tasks such as text classification Devlin et al. (2018).

BERT has two primary iterations: the Base model and the Large model, characterized by their layers, hidden states, attention mechanisms, and trainable parameters. The Large model, while more computationally demanding, is capable of capturing intricate linguistic relationships Vaswani et al. (2017).

BERT undergoes a two-stage training process: pre-training and fine-tuning. During pre-training, BERT learns general language patterns and structures. During fine-tuning, BERT is exposed to task-specific datasets containing labelled data, and task-specific layers are added on top of the pre-trained BERT model. This process enables the model to adjust its pre-trained, generalized language representations to the intricacies of the target task Devlin et al. (2018).

In our analysis, we utilized BERT for sentiment classification tasks. We first split the dataset into train, validation, and test sets. The input text was tokenized using the BERT tokenizer, and TensorFlow datasets were created from these encodings.

We defined the BERT model for sequence classification and compiled it with a custom loss function. The model was trained on the training dataset and evaluated on the test dataset. The results showed the test loss and accuracy of the model.

4.8.2 RoBERTa

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is an advanced language representation model that evolved from the BERT architecture. It includes strategic refinements such as extended training duration, exposure to an augmented training dataset, heightened batch size, and the incorporation of longer sequences. It also omits the Next Sentence Prediction (NSP) task in favor of dynamic masking. These adjustments enhance RoBERTa's classification capabilities, demonstrating superior performance compared to its predecessor, BERT, across key NLP benchmarks Liu et al. (2019).

In our analysis, we utilized RoBERTa for sentiment classification tasks. We first loaded the RoBERTa tokenizer and tokenized the input text. TensorFlow datasets were created from these encodings. We then defined the RoBERTa model for sequence classification and compiled it with a custom loss function. The model was trained on the training dataset and evaluated on the test dataset. The results showed the test loss and accuracy of the model.

4.8.3 FinBERT

Financial BERT models, derived from the BERT architecture, are specialized for financial applications, focusing on sentiment analysis and insights extraction from financial texts. They undergo pre-training on

substantial financial corpora, followed by fine-tuning for sentiment analysis tasks within the financial domain. Notably, several variations of FinBERT models exist, each with its unique pre-training data and architecture enhancements, reflecting the evolving landscape of research in financial sentiment analysis.

Model Variants:

Araci's FinBERT: Pre-trained on the financial sub-dataset of Reurers dataset TRC2, exhibiting superior performance in accuracy and F1 score compared to other models Araci (2019).

Desola et al.'s FinBERT Variants : Comprising FinBERT Prime, FinBERT Pre2K, and FinBERT Combo models, pre-trained on 10-K filings from SEC's EDGAR system. Offers enhanced performance in Next Sentence Prediction and Masked Language Modeling tasks Desola et al. (2019).

Liu et al.'s FinBERT Model: Extensively pre-trained on Financial Web, Yahoo! Finance, and RedditFinanceQA datasets, demonstrating notable improvements over conventional BERT models across various benchmarks Liu et al. (2021).

Yang et al.'s FinBERT Model: Offering both uncased and cased versions, pre-trained on corporate reports, earnings call transcripts, and analyst reports. Outperforms traditional ML models and predecessor BERT models across multiple metrics Yang et al. (2020) & Huang et al. (2022).

Hazourli's FinancialBERT: Unofficially introduced, pre-trained on a diverse set of financial datasets including TRC2, Bloomberg Financial News, Corporate Reports, and Earnings Call Transcripts. Reports remarkable performance, surpassing baseline BERT and FinBERT models on Financial PhraseBank dataset Hazouli (2022).

In our analysis, we fine-tuned Yang et al.'s FinBERT Model for sentiment analysis on financial text data. By leveraging FinBERT's domain-specific knowledge and contextual understanding of financial language, we aimed to improve sentiment classification accuracy and capture subtle sentiment signals specific to the financial domain.

4.9 Evaluation and Prediction Metrics

In financial sentiment analysis, accurately discerning between positive, negative, and neutral sentiments is paramount. The confusion matrix serves as a fundamental tool in evaluating the performance of classification models in this domain. It provides a detailed breakdown of predictions into four components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Additionally, in multiclass scenarios like financial sentiment analysis, we introduce True Neutral (TNe) and False Neutral (FNe) categories to capture the classification of neutral sentiments.

4.9.1 Confusion Matrix

The confusion matrix is indispensable for assessing the efficacy of classification models, particularly in financial sentiment analysis. It categorizes predictions into:

- True Positives (TP): Correctly classified positives
- True Negatives (TN): Correctly classified negatives
- True Neutral (TNe): Correctly classified neutrals
- False Positives (FP): Misclassified positives
- False Negatives (FN): Misclassified negatives
- False Neutral (FNe): Misclassified neutrals

These categories are pivotal for calculating key performance metrics such as precision, recall, accuracy, and F1-score, providing a nuanced understanding of the model's effectiveness.

Actual			
Negative	True Negative TN	False Neutral FNe1	False Positive FP2
Neutral	False Negative FN1	True Neutral TNe	False Positive FP1
Positive	False Negative FN2	False Neutral FNe2	True Positive TP
Predicted	**Negative**	**Neutral**	**Positive**

Fig. 5. Components of Confusion Matrix (Positive, Negative, Neutral)

4.9.2 Accuracy

Reflecting the overall effectiveness of the model in classifying sentiments, accuracy is the ratio of correct predictions (TP, TN, and TNe) to the total number of predictions. High accuracy indicates a model that performs well across all sentiment classes.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN} + \mathrm{TNe}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{TNe} + \mathrm{FP} + \mathrm{FN} + \mathrm{FNe})} \tag{9}$$

4.9.3 Precision

Indicates the reliability of positive predictions. It is the ratio of TP to all predicted positives (TP and FP). High precision signifies a low rate of false alarms, which is crucial in avoiding misclassification of legitimate websites as phishing.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

4.9.4 Recall

Measures the model's ability to identify all actual phishing instances. It is the ratio of TP to all actual positives (TP and FN). High recall is vital for ensuring that phishing websites are not missed.

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

4.9.5 F1-Score

The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance, especially in situations where there is an imbalance between classes, as it considers both false positives and false negatives.

$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (12)$$

4.10 Model Explainability and Interpretability

In the pursuit of understanding complex machine learning models, the concepts of explainability and interpretability play pivotal roles. These aspects enable practitioners to comprehend the decisions made by models, thereby fostering trust and aiding in debugging, compliance, and model improvement efforts.

4.10.1 SHAP

SHAP (SHapley Additive exPlanations) is a versatile technique rooted in game theory, illuminating the output of machine learning models by assigning importance values to features. Introduced by Lundberg and Lee (2017) and inspired by game theory concepts Lloyd (1952), SHAP offers a nuanced understanding of how individual features influence predictions. By computing Shapley values, SHAP provides insights into feature importance, irrespective of the underlying model Joseph (2019). Integrating SHAP into model interpretation workflows enhances understanding and facilitates informed decision-making across various applications.

4.10.2 LIME

Local Interpretable Model-agnostic Explanations (LIME) is a powerful technique designed to elucidate individual predictions of machine learning models, including those considered black boxes. By generating locally faithful explanations, LIME approximates the model's behavior around specific instances, offering actionable insights into the rationale behind predictions. Operating independently of the original classifier algorithm, LIME creates interpretable explanations for each individual prediction, fitting local models to sample data points similar to the observation being explained Ribeiro et al. (2016). By integrating LIME into model interpretation workflows, practitioners enhance understanding, trust, and informed decision-making across diverse applications.

4.11 Deployment Environments

Cloud-Based Platforms: Services like AWS, Google Cloud, and Microsoft Azure offer robust environments for deploying machine learning models. These platforms provide scalable resources to handle varying loads, making them ideal for large-scale applications.

On-Premises Servers: For organizations with strict data privacy regulations or specific performance requirements, deploying on local servers might be a better option. This allows for greater control over the data and resources.

Edge Devices: In some cases, it might be beneficial to deploy models directly on edge devices (like mobile phones or IoT devices). This can reduce latency and network bandwidth requirements.

Hybrid Environments: A combination of cloud, on-premises, and edge deployments might be used to balance the benefits of each approach.

Integrated within Financial Software**:** The models can also be deployed within existing financial software systems, providing sentiment analysis capabilities directly where financial decisions are being made.

Web Applications: Models can be deployed as part of web applications, providing users with real-time sentiment analysis of financial texts.

APIs: Models can be deployed as APIs, allowing other applications and services to use the sentiment analysis capabilities.

5 RESULTS AND DISCUSSIONS

The study analyzes the outcomes of our sentiment analysis models, exploring their performance and implications for market forecasting and investment decisions in the financial domain.

5.1 Logistic Regression (LR)

The study evaluated three logistic regression (LR) models for sentiment classification: BOW LR, TF-IDF LR, and Ensemble LR.

5.1.1 BOW LR

The bow_lr model achieves a 74% accuracy, indicating effective sentiment classification. Precision values for negative, neutral, and positive classes are 0.73, 0.72, and 0.73 respectively, suggesting a moderate false positive rate. Recall values for each class remain consistent at 0.73, demonstrating the model's ability to identify sentiments accurately. While the F1-score is not provided, the overall performance appears satisfactory.

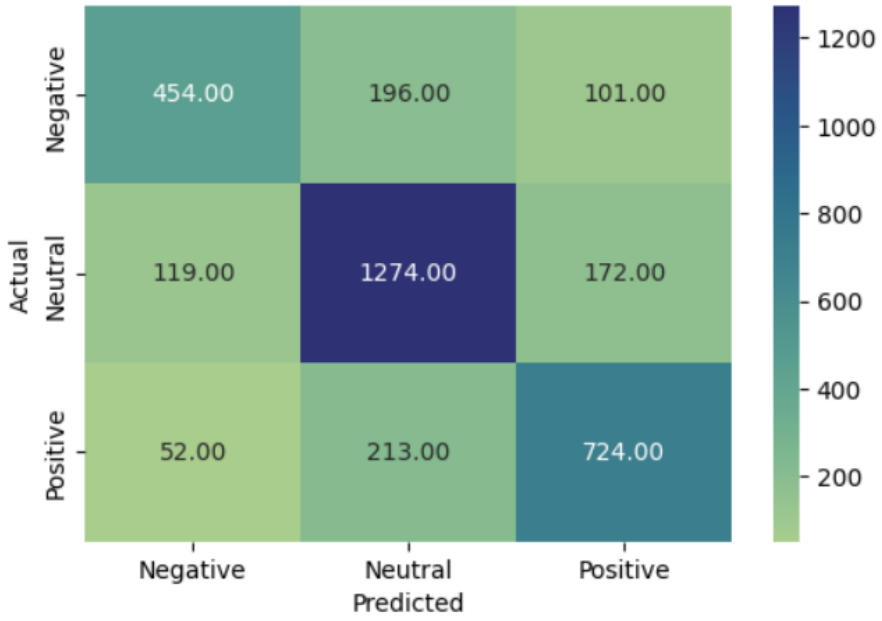


Fig. 6: Confusion matrix for the BOW LR Model

The confusion matrix reveals the bow_lr model's performance in sentiment classification. It correctly identifies 454 negative cases (True Negatives), 1274 neutral cases (True Neutrals), and 724 positive cases (True Positives). However, it misclassifies a significant number of cases into other categories:

- Negative predictions: 196 cases are misclassified as Neutral (False Neutrals) and 101 cases as Positive (False Positives).

- Neutral predictions: 119 cases are misclassified as Negative (False Negatives) and 172 cases as Positive (False Positives).
- Positive predictions: 52 cases are misclassified as Negative (False Negatives) and 213 cases as Neutral (False Neutrals).

5.1.2 TF-IDF LR

The tf-idf_lr model performs satisfactorily with an accuracy of 75%. Precision values for negative, neutral, and positive classifications stand at 0.75, 0.75, and 0.74 respectively, indicating a moderate false positive rate across all classes. Recall values for each class are 0.61, 0.83, and 0.72 respectively, reflecting the model's ability to effectively identify each class. Additionally, F1-scores, representing a balanced trade-off between precision and recall, are 0.67, 0.79, and 0.72 for negative, neutral, and positive classifications respectively.

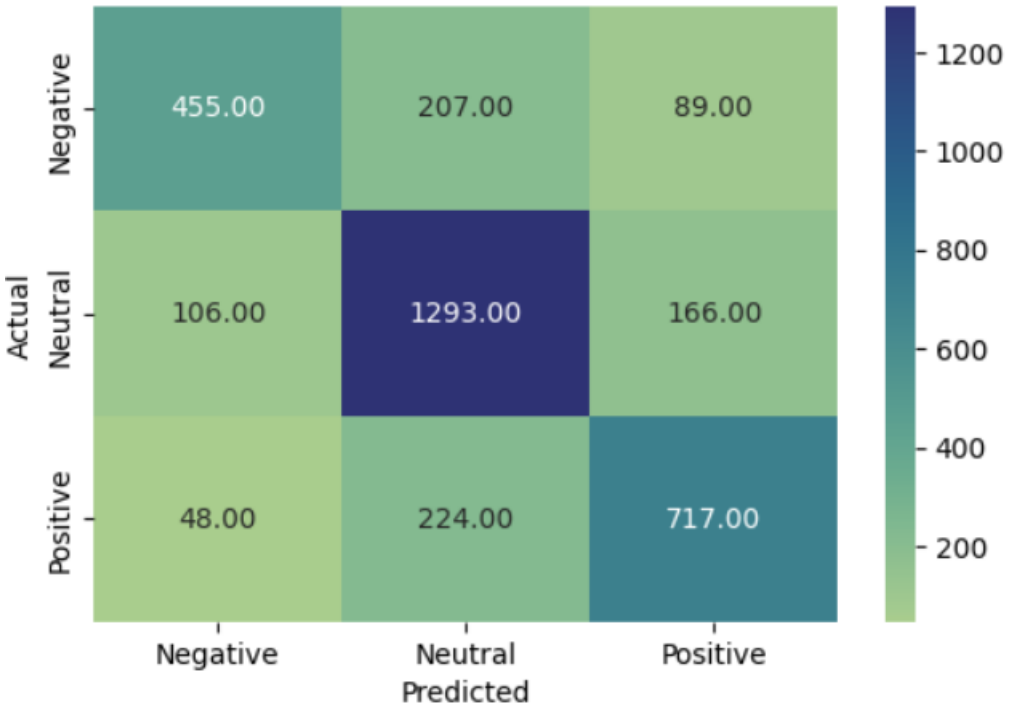


Fig. 7: Confusion matrix for the TF-IDF LR Model

The confusion matrix illustrates the performance of the tf-idf_lr model in sentiment classification. It accurately identifies 455 negative cases (True Negatives), 1293 neutral cases (True Neutrals), and 717 positive cases (True Positives). However, significant misclassifications occur:

- Negative predictions: 207 cases are misclassified as Neutral (False Neutrals) and 89 cases as Positive (False Positives).
- Neutral predictions: 106 cases are misclassified as Negative (False Negatives) and 166 cases as Positive (False Positives).
- Positive predictions: 48 cases are misclassified as Negative (False Negatives) and 224 cases as Neutral (False Neutrals).

5.1.3 Ensemble LR

The Ensemble_lr model demonstrates a satisfactory performance, achieving an accuracy of 75%. The precision values for negative, neutral, and positive classifications are 0.76, 0.74, and 0.77 respectively, indicating a moderate false positive rate across all classes. The recall values for each class are 0.57, 0.85, and 0.72 respectively, reflecting the model's ability to effectively identify each class. The F1-score, which represents a well-balanced trade-off between precision and recall, is 0.66, 0.79, and 0.74 for negative, neutral, and positive classifications respectively.

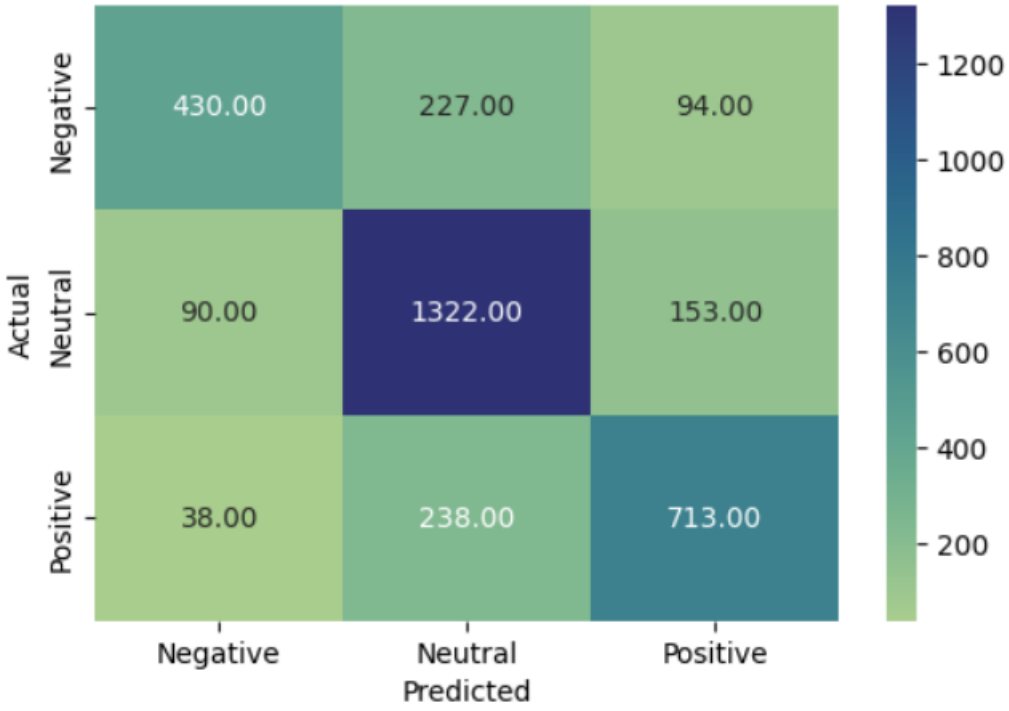


Fig. 8: Confusion matrix for the Ensemble_lr Model

The confusion matrix for the Ensemble_lr model reveals its performance in sentiment classification. It correctly identifies 430 negative cases (True Negatives), 1322 neutral cases (True Neutrals), and 713 positive cases (True Positives). However, significant misclassifications occur:

- Negative predictions: 227 cases are misclassified as Neutral (False Neutrals) and 94 cases as Positive (False Positives).
- Neutral predictions: 90 cases are misclassified as Negative (False Negatives) and 153 cases as Positive (False Positives).
- Positive predictions: 38 cases are misclassified as Negative (False Negatives) and 238 cases as Neutral (False Neutrals).

The TF-IDF LR and Ensemble LR models perform marginally better than the BOW LR model, particularly in identifying neutral sentiments. However, all models exhibit a moderate false positive rate and could benefit from further optimization to reduce misclassifications. The Ensemble LR model stands out with the highest recall value for neutral cases, indicating a potential advantage in this area.

5.2 Support Vector Machine (SVM)

The study assessed three Support Vector Machine (SVM) models for sentiment classification: BOW SVM, TF-IDF SVM, and Ensemble SVM

5.2.1 BOW SVM

The bow_svm model exhibits satisfactory performance, achieving a 74% accuracy rate. Precision values for negative, neutral, and positive classifications are 0.70, 0.75, and 0.74 respectively, suggesting a moderate false positive rate across all classes. Recall values for each class are 0.63, 0.80, and 0.71 respectively, indicating the model's effectiveness in identifying each class. The F1-scores, representing a balanced trade-off between precision and recall, are 0.66, 0.78, and 0.72 for negative, neutral, and positive classifications respectively.

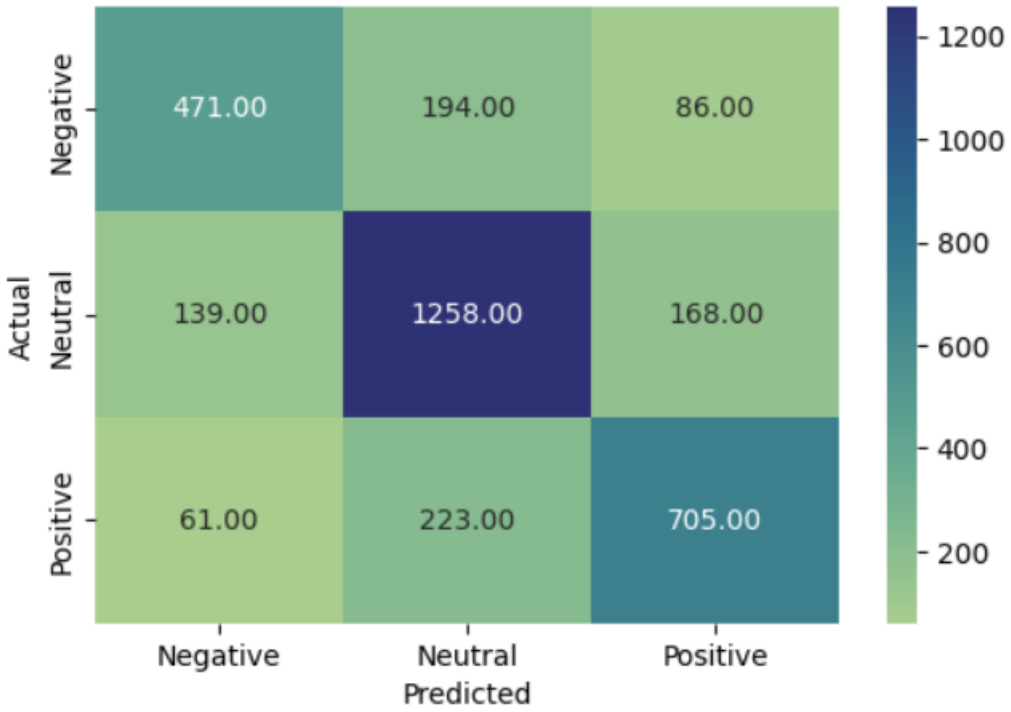


Fig. 9: Confusion matrix for the bow_svm Model

The confusion matrix illustrates the performance of the bow_svm model in sentiment classification. It accurately identifies 471 negative cases (True Negatives), 1258 neutral cases (True Neutrals), and 705 positive cases (True Positives). However, significant misclassifications occur:

- Negative predictions: It misclassified 194 as Neutral (False Neutrals) and misclassified 86 as Positive (False Positives).
- Neutral predictions: It misclassified 139 as Negative (False Negatives) and misclassified 168 as Positive (False Positives).

- Positive predictions: It misclassified 61 as Negative (False Negatives) and misclassified 223 as Neutral (False Neutrals).

5.2.2 TF-IDF SVM

The tf-idf_svm model performs satisfactorily with an accuracy of 75%. Precision values for negative, neutral, and positive classifications stand at 0.76, 0.75, and 0.74 respectively, indicating a moderate false positive rate across all classes. Similarly, recall values for each class are 0.63, 0.83, and 0.71 respectively, underscoring the model's effectiveness in class identification. The F1-scores further affirm this, with values of 0.67, 0.79, and 0.72 for negative, neutral, and positive classifications respectively.

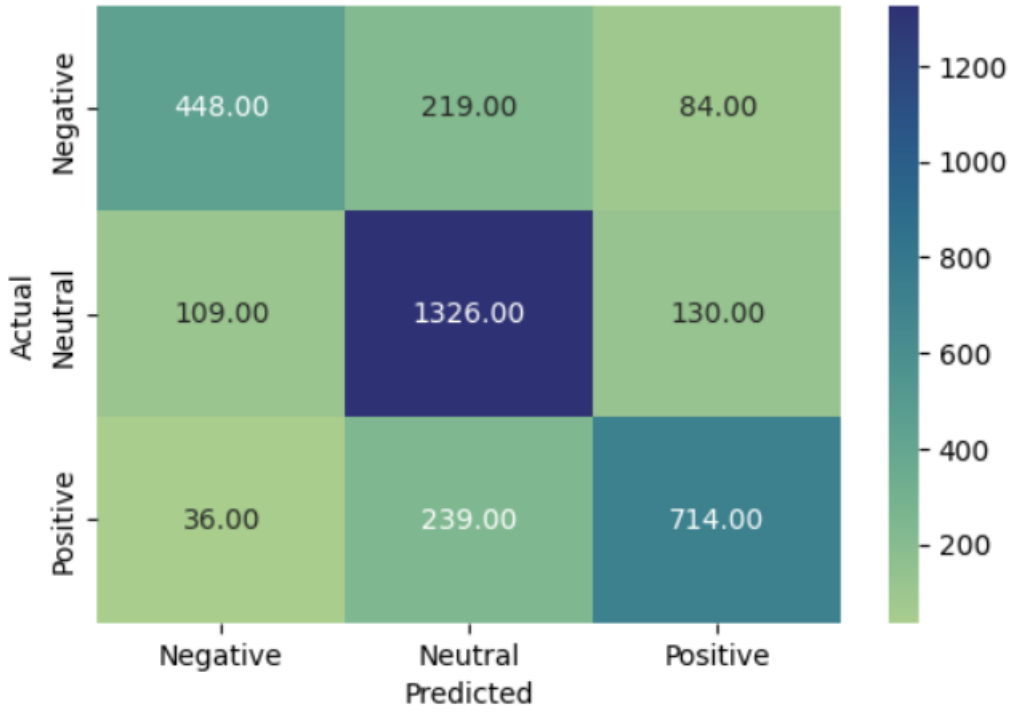


Fig. 10: Confusion matrix for the tf-idf_svm Model

The confusion matrix depicts the performance of the tf-idf_svm model in sentiment classification. It accurately identifies 448 negative cases (True Negatives), 1326 neutral cases (True Neutrals), and 714 positive cases (True Positives). However, significant misclassifications occurred:

- Negative predictions: 219 instances were misclassified as Neutral (False Neutrals) and 84 instances were misclassified as Positive (False Positives).
- Neutral predictions: 109 instances were misclassified as Negative (False Negatives) and 130 instances were misclassified as Positive (False Positives).
- Positive predictions: 36 instances were misclassified as Negative (False Negatives) and 239 instances were misclassified as Neutral (False Neutrals).

5.2.3 Ensemble SVM

The Ensemble_svm model demonstrates solid performance, achieving an accuracy of 75%. Precision values for negative, neutral, and positive classifications are 0.72, 0.76, and 0.76 respectively, indicating a moderate false positive rate across all classes. Correspondingly, recall values for each class are 0.63, 0.83, and 0.73 respectively, highlighting the model's ability to effectively identify each class. The F1-scores further confirm this, with values of 0.67, 0.79, and 0.75 for negative, neutral, and positive classifications respectively.

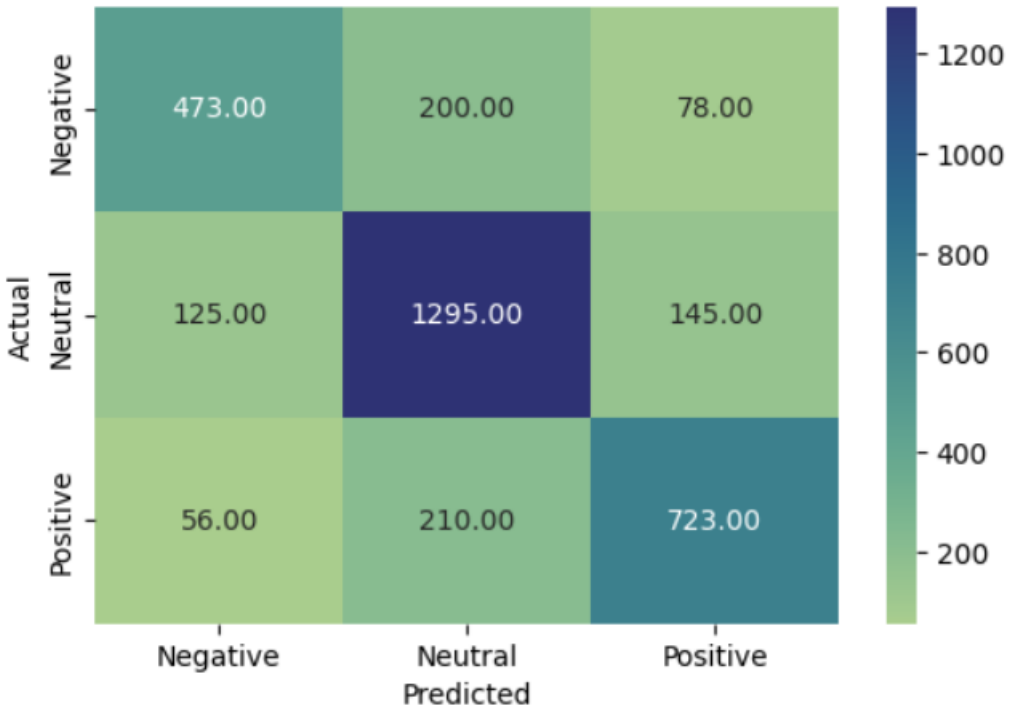


Fig. 11: Confusion matrix for the Ensemble SVM

The confusion matrix indicates the performance of the Ensemble_xgboost model. It accurately identifies 473 negative cases as true negatives, 1295 neutral cases as true neutrals, and 723 positive cases as true positives. However, significant misclassifications occurred:

- Negative predictions: It misclassified 200 as Neutral (False Neutrals) and misclassified 78 as Positive (False Positives).
- Neutral predictions: It misclassified 125 as Negative (False Negatives) and misclassified 145 as Positive (False Positives).
- Positive predictions: It misclassified 56 as Negative (False Negatives) and misclassified 210 as Neutral (False Neutrals)

Each model has a confusion matrix detailing true positives, false positives, and false negatives for sentiment classification. The Ensemble SVM model demonstrates the best performance among the three.

5.3 XGBoost

The study evaluated three XGBoost models for sentiment classification: BOW XGBoost, TF-IDF XGBoost, and Ensemble XGBoost.

5.3.1 BOW XGBoost

The bow_xgboost model exhibits a performance that meets expectations, achieving an accuracy of 71%. Precision values for negative, neutral, and positive classifications stand at 0.76, 0.69, and 0.73 respectively, indicating a moderate false positive rate across all classes. Similarly, recall values for each class are 0.54, 0.86, and 0.61 respectively, highlighting the model's ability to effectively identify each class. The F1-scores further underscore this, with values of 0.66, 0.77, and 0.67 for negative, neutral, and positive classifications respectively.

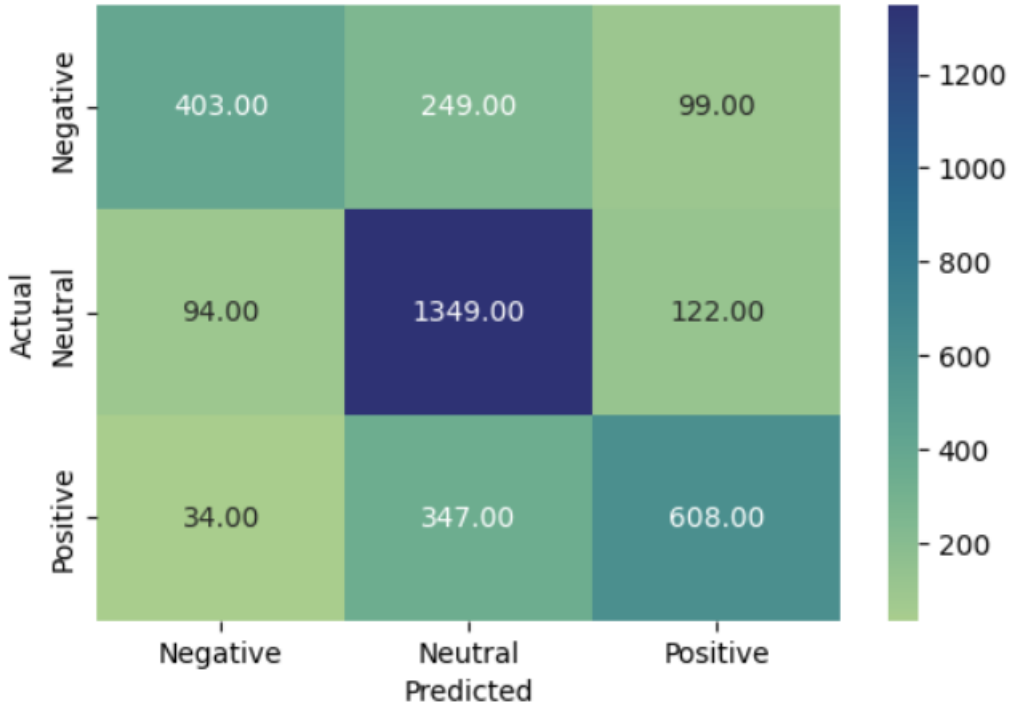


Fig. 12: Confusion matrix for the bow_xgboost Model

The confusion matrix reveals that the bow_xgboost model correctly identified 403 negative cases as such (True Negatives), 1349 neutral cases accurately (True Neutrals), and 608 positive cases as positive (True Positives). However, it misclassified significant numbers into other categories:

- Negative predictions: It misclassified 249 as Neutral (False Neutrals) and misclassified 99 as Positive (False Positives).
- Neutral predictions: It misclassified 94 as Negative (False Negatives) and misclassified 122 as Positive (False Positives).
- Positive predictions: It misclassified 34 as Negative (False Negatives) and misclassified 347 as Neutral (False Neutrals).

5.3.2 TF-IDF XGBoost

The tf-idf_xgboost model demonstrates a satisfactory performance, achieving an accuracy of 71%. The precision values for negative, neutral, and positive classifications are 0.76, 0.69, and 0.73 respectively, indicating a moderate false positive rate across all classes. The recall values for each class are 0.52, 0.86, and 0.61 respectively, reflecting the model's ability to effectively identify each class. The F1-score, which represents a well-balanced trade-off between precision and recall, is 0.62, 0.76, and 0.66 for negative, neutral, and positive classifications respectively.

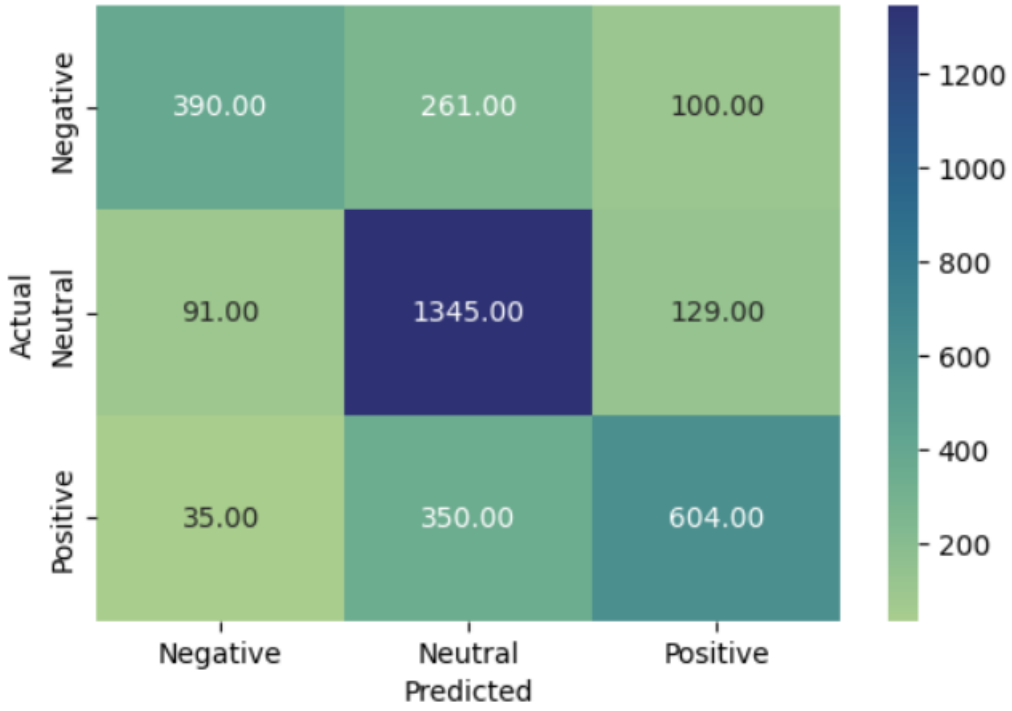


Fig. 13: Confusion matrix for the tf-idf_xgboost Model

The confusion matrix illustrates the performance of the tf-idf_xgboost model. It accurately identifies 390 negative cases as true negatives, 1326 neutral cases as true neutrals, and 604 positive cases as true positives. However, notable misclassifications occurred:

- Negative predictions: It misclassified 261 as Neutral (False Neutrals) and misclassified 100 as Positive (False Positives).
- Neutral predictions: It misclassified 91 as Negative (False Negatives) and misclassified 129 as Positive (False Positives).
- Positive predictions: It misclassified 35 as Negative (False Negatives) and misclassified 350 as Neutral (False Neutrals).

5.3.3 Ensemble XGBoost

The Ensemble_xgboost model demonstrates a satisfactory performance, achieving an accuracy of 71%. The precision values for negative, neutral, and positive classifications are 0.76, 0.69, and 0.73 respectively, indicating a moderate false positive rate across all classes. The recall values for each class are 0.52, 0.86, and 0.61 respectively, reflecting the model's ability to effectively identify each class. The

F1-score, which represents a well-balanced trade-off between precision and recall, is 0.62, 0.76, and 0.66 for negative, neutral, and positive classifications respectively.

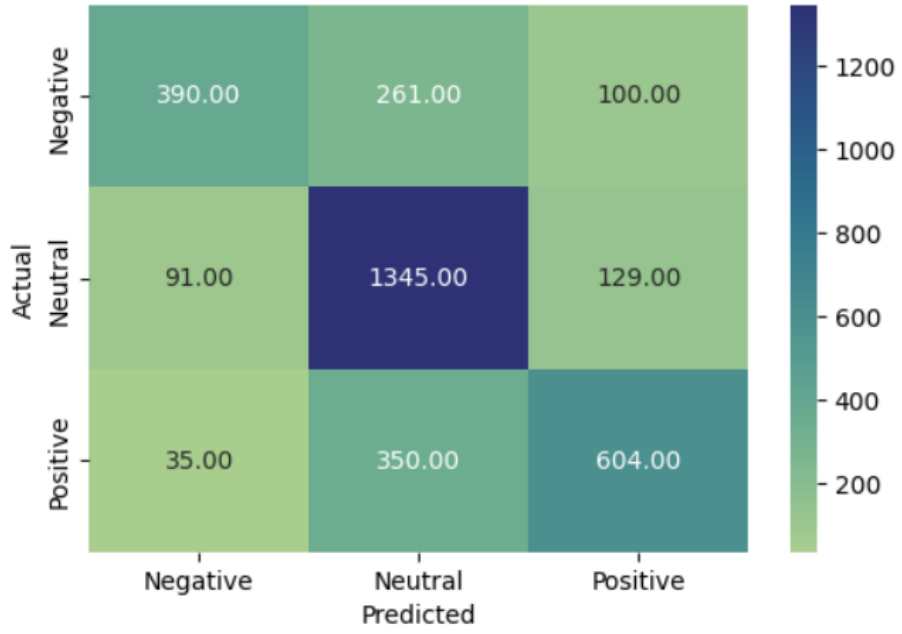


Fig. 14. Confusion matrix for the Ensemble_xgboost Model

The confusion matrix reveals that the Ensemble_xgboost model correctly identified 473 negative cases as such (True Negatives), 1295 neutral cases accurately (True Neutrals), and 604 positive cases as positive (True Positives). However, it misclassified significant numbers into other categories:

- Negative predictions: It misclassified 200 as Neutral (False Neutrals) and misclassified 78 as Positive (False Positives).
- Neutral predictions: It misclassified 125 as Negative (False Negatives) and misclassified 145 as Positive (False Positives).
- Positive predictions: It misclassified 56 as Negative (False Negatives) and misclassified 210 as Neutral (False Neutrals).

All three XGBoost models perform similarly, with room for improvement in reducing misclassifications. The Ensemble XGBoost model, with the highest recall for neutral cases, may offer unique advantages for certain applications.

5.4 BERT

The BERT model demonstrates a commendable performance, achieving an accuracy of 73%. The precision for negative, neutral, and positive classes are 0.62, 0.80, and 0.73 respectively, indicating a moderate false positive rate across the classes. The recall for these classes stands at 0.76, 0.73, and 0.71 respectively, reflecting the model's ability to effectively identify different sentiments. The F1-score for these classes are 0.68, 0.77, and 0.72 respectively, representing a balanced trade-off between precision and recall, highlighting the model's effectiveness in sentiment analysis.

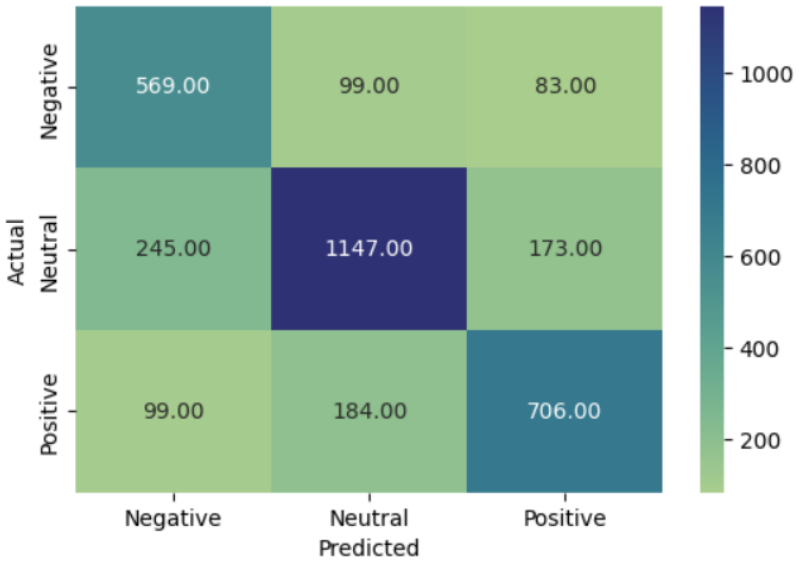


Fig. 15: Confusion matrix for the BERT Model

The confusion matrix reveals that BERT correctly identified 569 negative sentiments as such (True Negatives), 1147 neutral sentiments accurately (True Positives for Neutral), and 706 positive sentiments as accurate (True Positives). However, it misclassified:

- Negative predictions: It misclassified 99 as Neutral (False Neutrals) and misclassified 83 as Positive (False Positives).
- Neutral predictions: It misclassified 245 as Negative (False Negatives) and misclassified 173 as Positive (False Positives).
- Positive predictions: It misclassified 99 as Negative (False Negatives) and misclassified 184 as Neutral (False Neutrals).

5.5 RoBERTa

The RoBERTa model demonstrates a strong performance, achieving an accuracy of 76%. The precision for negative, neutral, and positive classes are 0.76, 0.78, and 0.73 respectively, indicating a moderate false positive rate across the classes. The recall for these classes stands at 0.66, 0.80, and 0.78 respectively, reflecting the model's ability to effectively identify different sentiments. The F1-score for these classes are 0.71, 0.79, and 0.76 respectively, representing a balanced trade-off between precision and recall, highlighting the model's effectiveness in sentiment analysis.

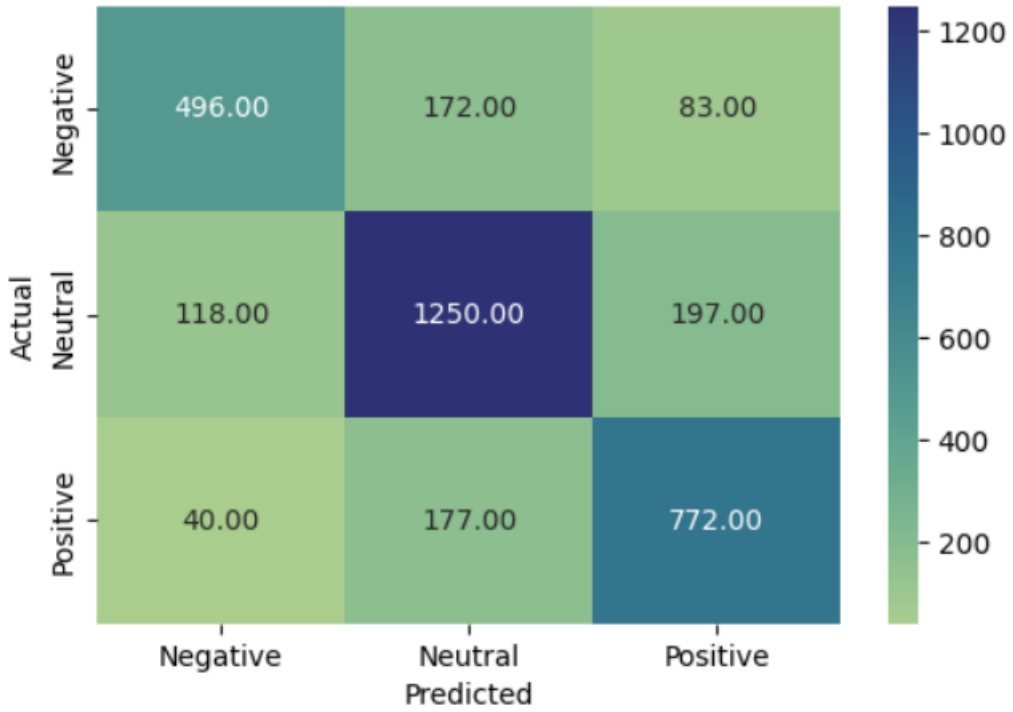


Fig. 16: Confusion matrix for the RoBERTa Model

The confusion matrix reveals that RoBERTa correctly identified 496 negative sentiments as such (True Negatives), 1250 neutral sentiments accurately (True Positives for Neutral), and 772 positive sentiments as accurate (True Positives). However, it misclassified:

- Negative predictions: RoBERTa misclassified 172 as Neutral (False Neutrals) and misclassified 83 as Positive (False Positives).
- Neutral predictions: It misclassified 118 as Negative (False Negatives) and misclassified 197 as Positive (False Positives).
- Positive predictions: It misclassified 40 as Negative (False Negatives) and misclassified 177 as Neutral (False Neutrals).

5.6 FinBERT

The FinBERT model demonstrates a strong performance, achieving an accuracy of 72%. The precision for negative, neutral, and positive classes are 0.62, 0.81, and 0.69 respectively, indicating a moderate false positive rate across the classes. The recall for these classes stands at 0.72, 0.68, and 0.78 respectively, reflecting the model's ability to effectively identify different sentiments. The F1-score for these classes are 0.67, 0.74, and 0.73 respectively, representing a balanced trade-off between precision and recall, highlighting the model's effectiveness in sentiment analysis.

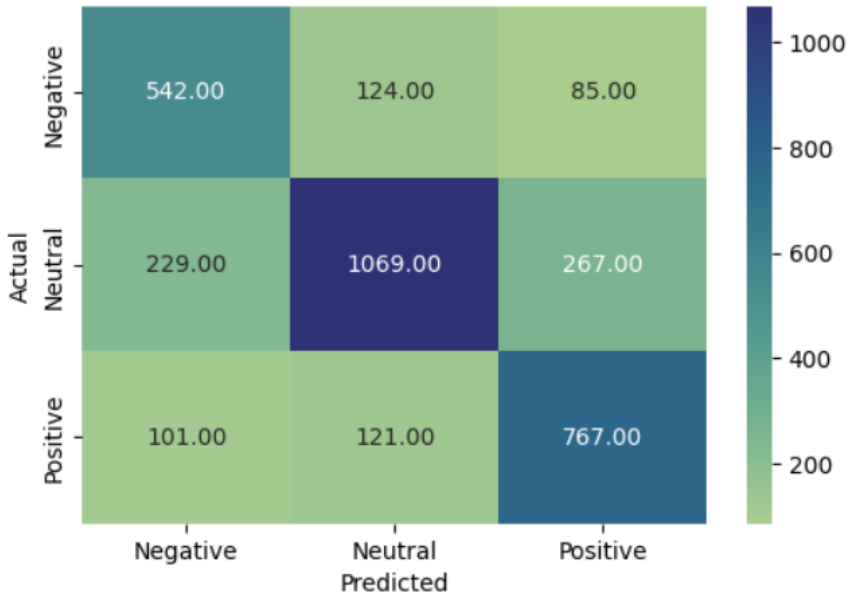


Fig. 17: Confusion matrix for the FinBERT Model

The confusion matrix reveals that FinBERT correctly identified 542 negative sentiments as such (True Negatives), 1250 neutral sentiments accurately (True Positives for Neutral), and 772 positive sentiments as accurate (True Positives). However, it misclassified:

- Negative predictions: FinBERT misclassified 172 as Neutral (False Neutrals) and misclassified 83 as Positive (False Positives).
- Neutral predictions: It misclassified 118 as Negative (False Negatives) and misclassified 197 as Positive (False Positives).
- Positive predictions: It misclassified 40 as Negative (False Negatives) and misclassified 177 as Neutral (False Neutrals).

5.7 Model Comparison

In this study, we evaluated a total of six models for sentiment classification: three traditional machine learning (ML) models, namely Logistic Regression (LR), Support Vector Machine (SVM), and XGBoost, and three Natural Language Processing (NLP) models, including BERT, RoBERTa, and FinBERT.

For the traditional ML models, each underwent three different training approaches: one with Bag-of-Words (BOW), another with Term Frequency-Inverse Document Frequency (TF-IDF), and finally, an ensemble method combining both models.

5.7.1 Traditional ML Models

Here, we analyze the performance of LR, SVM, and XGBoost across different training methods to determine their effectiveness in sentiment classification.

	Model	Accuracy	Precision	Recall	F1-Score
0	Bow_LogReg	0.741906	0.740815	0.741906	0.739598
1	Bow_SVM	0.736460	0.735127	0.736460	0.734840
2	Bow_XGBoost	0.714070	0.720350	0.714070	0.706996
3	Tf-Idf_LogReg	0.745840	0.745653	0.745840	0.743182
4	Tf-Idf_SVM	0.752799	0.753865	0.752799	0.749372
5	Tf-Idf_XGB	0.707716	0.714332	0.707716	0.700058
6	Ensemble_LogReg	0.745840	0.747665	0.745840	0.741741
7	Ensemble_SVM	0.753707	0.752703	0.753707	0.751673
8	Ensemble_XGB	0.707716	0.714332	0.707716	0.700058

Fig. 18. ML Models and the metrics

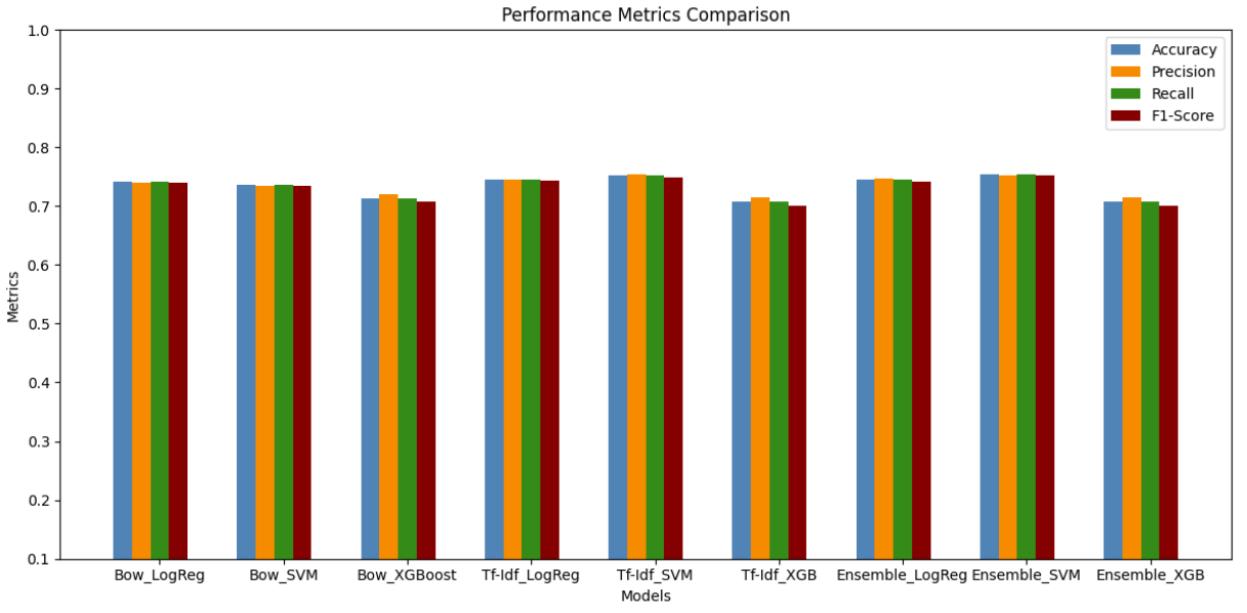


Fig. 19. ML Models and Graph Representation

5.7.1.1 Findings:

- TF-IDF outperformed BoW in SVM and LR models: This suggests that TF-IDF may capture more informative features for sentiment classification compared to BoW.
- Ensemble SVM outperformed individual SVM models: Indicates the effectiveness of ensemble techniques in improving performance.
- Consistent performance across XGBoost models: Despite variations in feature extraction techniques, XGBoost models showed similar performance, suggesting robustness in handling different types of features.
- Best Performing Model: The Ensemble SVM model performs the best with an accuracy of 75.37%. It also has the highest F1-score of 0.751673, indicating a good balance between precision and recall.
- Precision Analysis: The TF-IDF SVM model has the highest precision of 0.753865, suggesting it has a lower rate of false positives.
- Recall Analysis: The Ensemble LogReg and TF-IDF LogReg models have the highest recall of 0.745840, indicating these models are better at identifying true positives.
- F1-Score Analysis: The Ensemble SVM model has the highest F1-score, suggesting it provides the best balance between precision and recall.
- Special Findings: The Ensemble models generally perform better than the non-ensemble models, suggesting that combining the predictions of multiple models leads to improved performance.
- Room for Improvement: The Bow_XGBoost and TF-IDF XGB models have the lowest accuracy and F1-score, indicating these models might need further tuning or more training data.

5.7.2 NLP Models

The NLP models include BERT, RoBERTa, and FinBERT. These models are pre-trained on a large corpus of text and then fine-tuned for specific tasks, such as sentiment analysis in this case. They are capable of understanding the context of words and sentences, which makes them powerful tools for many NLP tasks.

In the following sections, we will delve into the performance of each model, comparing their strengths and weaknesses in the task of sentiment analysis. We will also discuss the implications of these findings for future research and applications.

	Model	Accuracy	Precision	Recall	F1-Score
0	BERT	0.732829	0.741040	0.732829	0.734668
1	RoBERT	0.761876	0.762104	0.761876	0.760963
2	FinBERT	0.732829	0.741040	0.732829	0.734668

Fig. 20. ML Models and the metrics

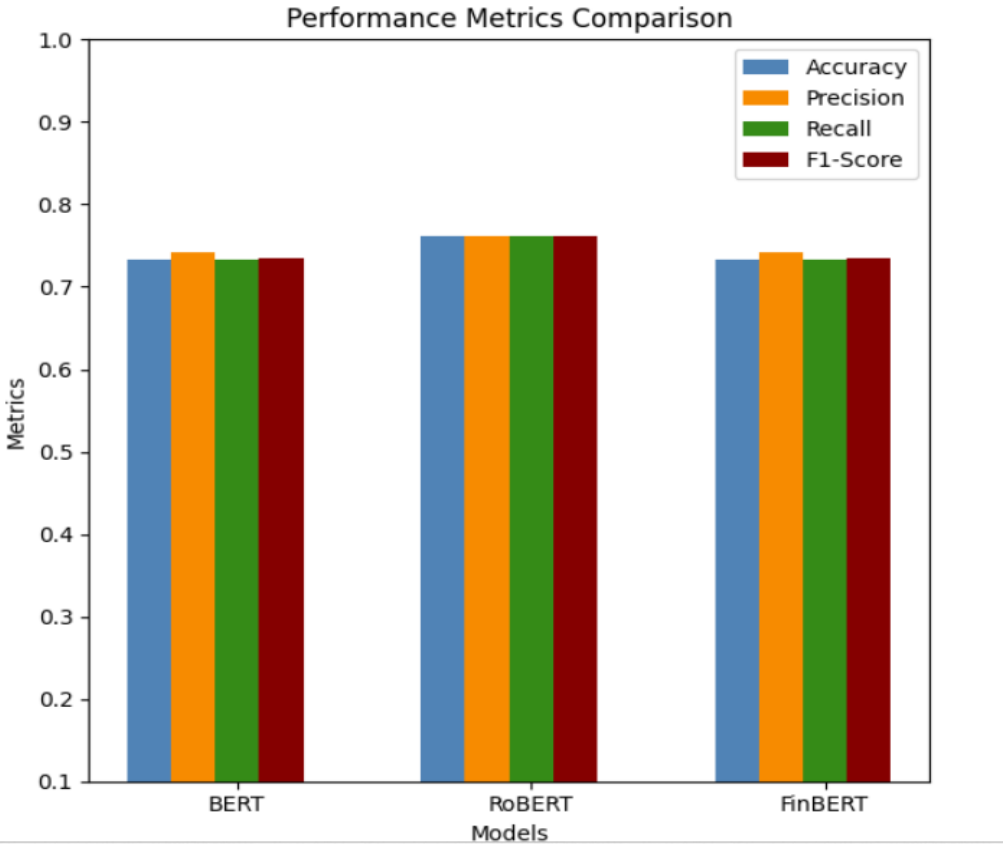


Fig. 21. NLP Models and Graph Representation

5.7.2.1 Findings:

- RoBERTa outperformed other models: With the highest accuracy of 76.19% and consistently high precision, recall, and F1-score, RoBERTa emerged as the top-performing model among BERT, RoBERTa, and FinBERT.
- Similar performance between BERT and FinBERT: Despite being trained on financial-specific data, FinBERT showed similar performance to BERT, suggesting that domain-specific embeddings may not always lead to substantial improvements in sentiment classification accuracy.
- Best Performing Model: RoBERTa demonstrated superior performance compared to BERT and FinBERT, with the highest accuracy and balanced precision, recall, and F1-score.
- Domain-specific embeddings: While FinBERT is tailored for financial sentiment analysis, its performance was comparable to generic NLP models like BERT, indicating potential limitations in leveraging domain-specific embeddings for sentiment classification in this context.

5.7.3 Summary:

RoBERTa**:** Emerged as the best-performing model overall, surpassing both traditional ML and other NLP models in accuracy, precision, recall, and F1-score.

TF-IDF SVM: Among traditional ML models, TF-IDF SVM achieved the highest accuracy and F1-score, demonstrating its effectiveness in sentiment classification tasks.

Consistency: While NLP models generally outperformed traditional ML models, the differences in performance were not substantial, suggesting that traditional ML methods can still be competitive in sentiment analysis tasks.

While NLP models, particularly RoBERTa, showcased superior performance in sentiment classification compared to traditional ML models, the differences were relatively marginal. Both approaches have their strengths and weaknesses, and the choice between them depends on factors such as the nature of the data, computational resources, and specific requirements of the task at hand. Further exploration and experimentation with various models and techniques may lead to better insights and improvements in sentiment analysis accuracy and efficiency

5.8 Model Explainability and Interpretability with SHAP and LIME

Here, we delve into the techniques employed to understand and interpret the predictions made by our sentiment analysis models, offering insights crucial for informed decision-making in financial markets

5.8.1 SHAP on Traditional ML Models

SHAP values provide a way to understand the contribution of each feature to the prediction of each instance for different classes

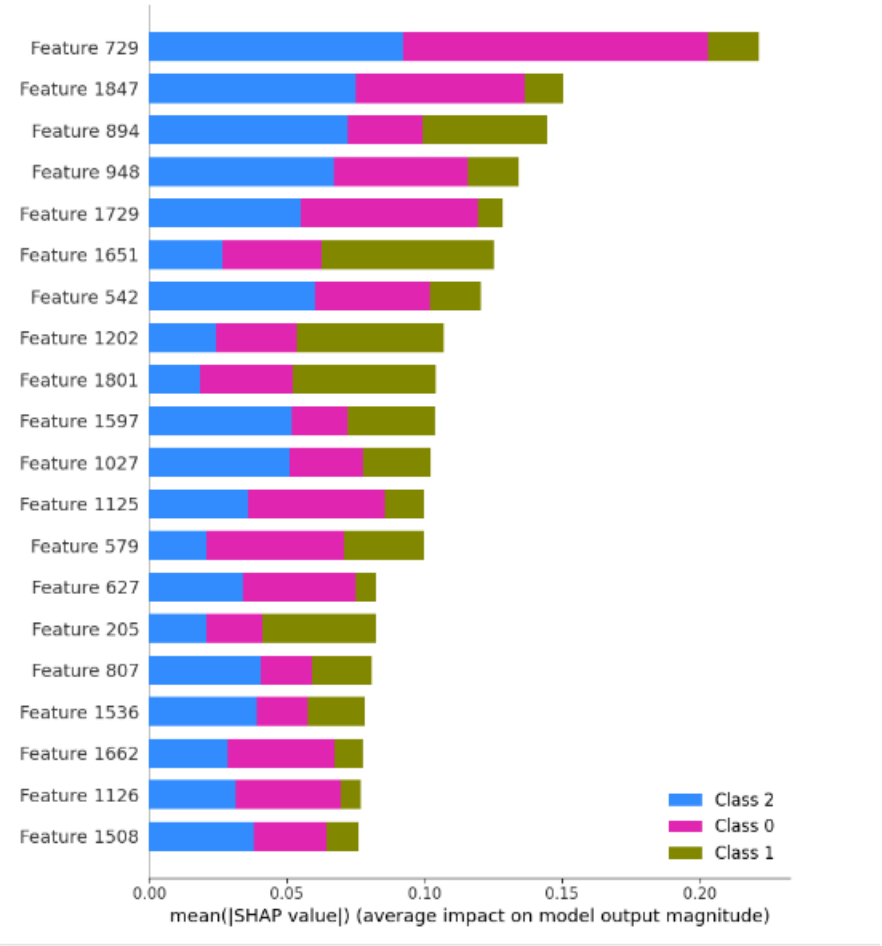


Fig. 22. ML Models and Graph Representation

The “Feature 729” has a significant impact across all three classes, while “Feature 1508” has a lesser impact. This suggests that “Feature 729” is more important for the model’s predictions than “Feature 1508”.

5.8.2 LIME on NLP Models

LIME is another widely used technique for model interpretability, particularly useful for understanding individual predictions in complex ML models.

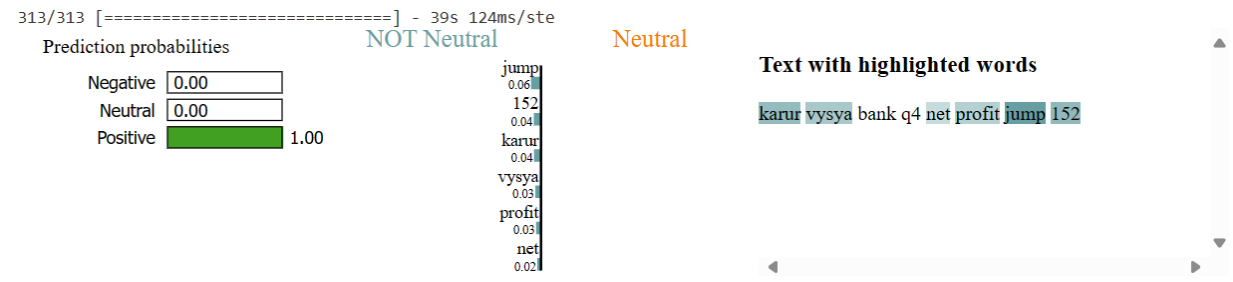


Fig. 23. Prediction with BERT Model for Sentence[5]

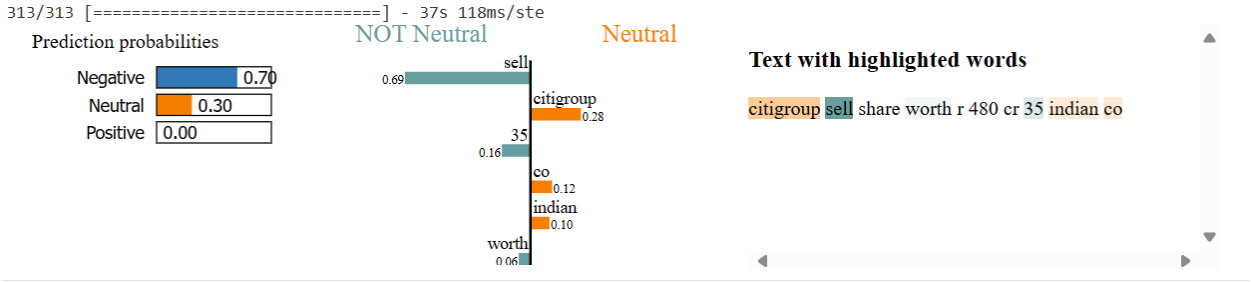


Fig. 24. Prediction with BERT Model for Sentence[10]

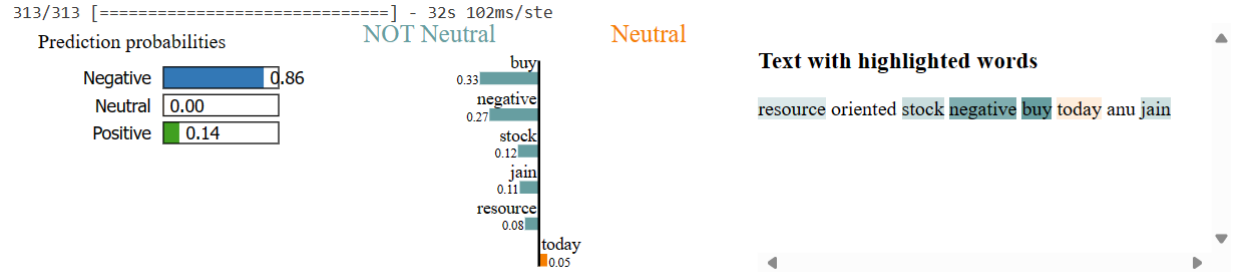


Fig. 25. Prediction with BERT Model for Sentence[100]

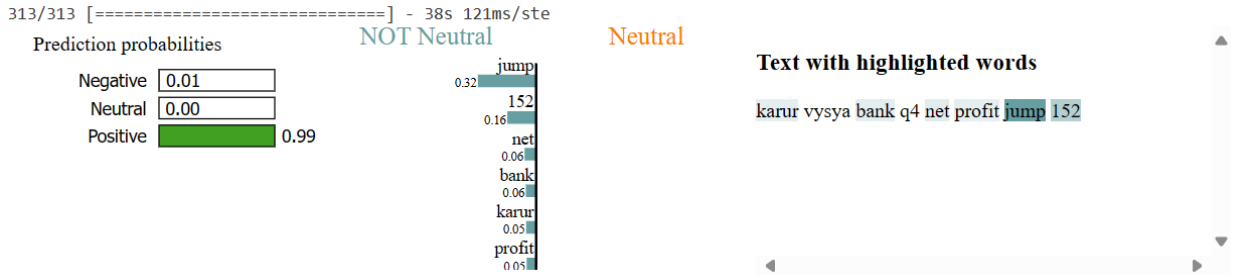


Fig. 26. Prediction with RoBERTa Model for Sentence[5]

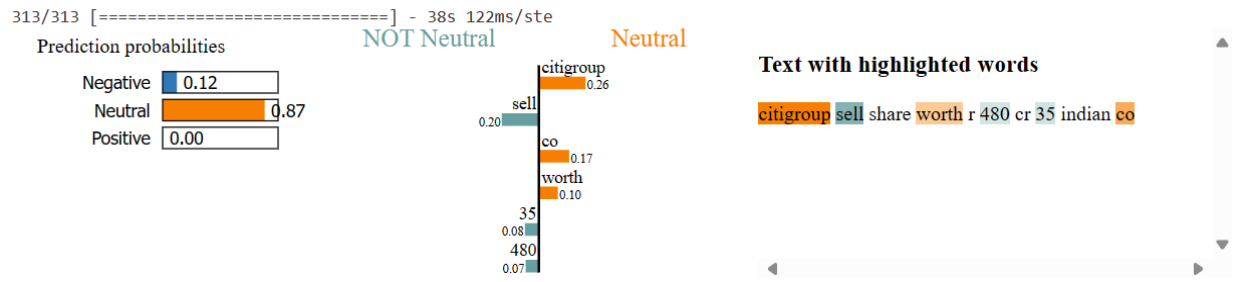


Fig. 27. Prediction with RoBERTa Model for Sentence[10]

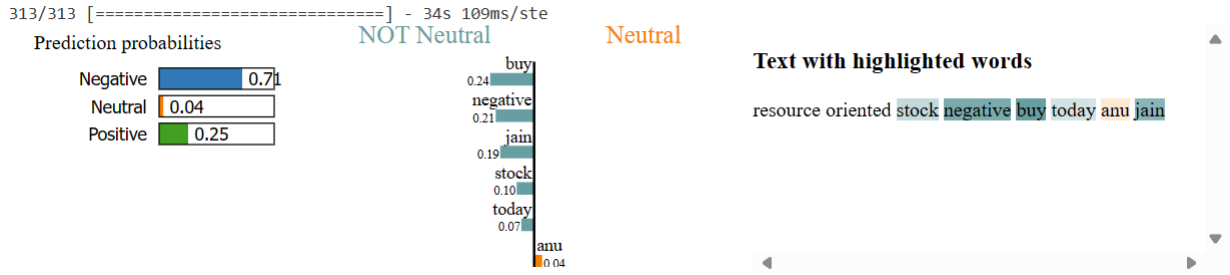


Fig. 28. Prediction with RoBERTa Model for Sentence[100]

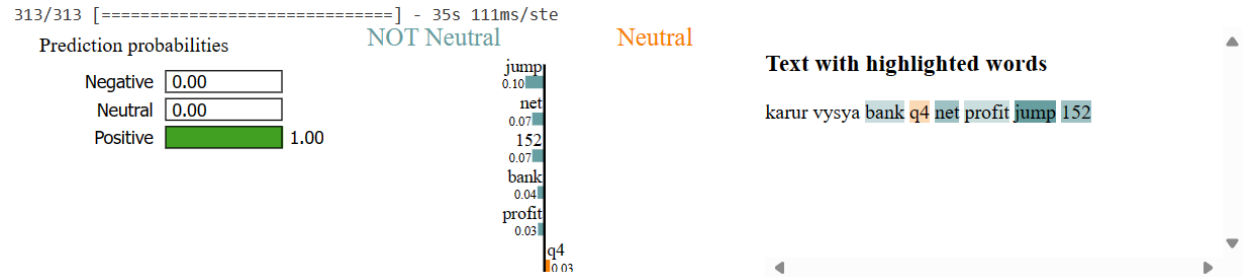


Fig. 29. Prediction with FinBERT Model for Sentence[5]

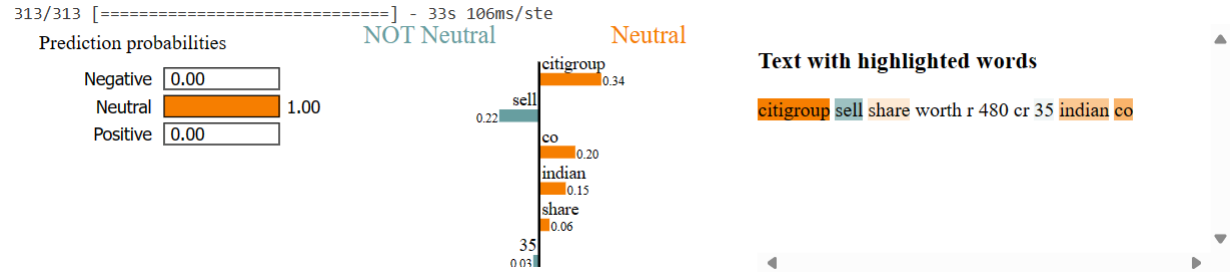


Fig. 30. Prediction with FinBERT Model for Sentence[10]

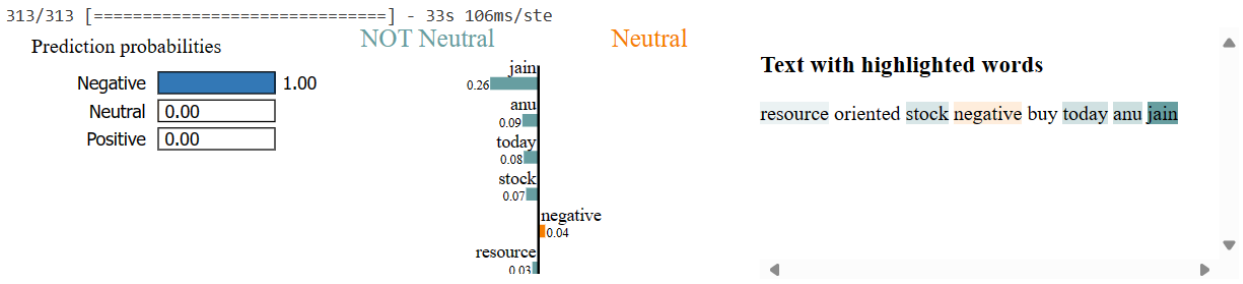


Fig. 31. Prediction with FinBERT Model for Sentence[100]

We can visualize the predictions and interpretations of the models on different sentences and their impacts.

5.9 LIMITATIONS

Despite the comprehensive analysis conducted in this dissertation, there are several limitations that need to be acknowledged:

Limited Scope of Models: While a wide range of machine learning (ML) and natural language processing (NLP) models were explored, there are numerous other models and techniques available in the field. The study may not have covered all possible approaches to sentiment analysis, which could limit the generalizability of the findings.

Data Quality and Quantity: The performance of the models heavily relies on the quality and quantity of the dataset used for training and evaluation. The dataset sourced from Kaggle may not fully capture the diversity and complexity of real-world financial sentiment analysis tasks.

Model Interpretability: While efforts were made to interpret the models' predictions using techniques like Local Interpretable Model-Agnostic Explanations (LIME), achieving full interpretability remains challenging, especially with complex models like deep learning architectures.

Misclassification Rates: Despite achieving high accuracy, all models had notable misclassifications. This indicates room for improvement in precision and recall.

Data Imbalance: The performance of the models could be affected by imbalanced class distribution in the training data. Techniques to handle class imbalance were not explored in this study.

Feature Representation: The reliance on BOW and TF-IDF methods for feature representation in traditional ML models may limit the models' ability to capture semantic nuances in the text.

Hyperparameter Tuning: Due to computational constraints, the hyperparameter tuning process may not have been exhaustive. Further exploration of hyperparameter spaces could lead to better-performing models.

Domain-Specific Considerations: Financial sentiment analysis poses unique challenges compared to sentiment analysis in other domains. The models' performance may vary when applied to different financial markets or types of financial data.

6 CONCLUSION

In conclusion, this dissertation investigated various ML and NLP techniques for empowering market forecasting and investment decisions through advanced sentiment analysis. Deep learning models, pre-trained models, and hybrid approaches were explored, with promising results observed across different methodologies.

The findings suggest that while advanced NLP models like RoBERTa demonstrated superior performance, traditional ML models such as TF-IDF SVM also performed competitively. Ensemble techniques proved to be effective in improving model performance, indicating the importance of combining multiple models for robust predictions.

Overall, the study contributes to the growing body of research in financial sentiment analysis by providing insights into the efficacy of different models and methodologies. However, there are still limitations and opportunities for further research in this domain.

6.1 FUTURE WORK

Data Augmentation**:** Increasing the diversity and size of the dataset through data augmentation techniques could enhance model performance and generalization.

Interpretability: Developing more robust methods for interpreting complex models like deep learning architectures would facilitate better understanding and trust in model predictions.

Semantic Analysis: Incorporating semantic analysis techniques, such as word embeddings or context-aware models, to enhance feature representation and capture more nuanced sentiment information.

Domain-Specific Models: Further exploration of domain-specific embeddings and pre-trained models tailored for financial sentiment analysis could lead to improved performance in this domain.

Dynamic Modeling: Incorporating temporal dynamics and market-related factors into sentiment analysis models could enhance their predictive capabilities for market forecasting tasks.

6.2 RECOMMENDATIONS

Model Diversity: Researchers and practitioners should consider exploring a diverse range of models and techniques, including both traditional ML and advanced NLP approaches, to identify the most effective solutions for specific financial sentiment analysis tasks.

Ensemble Methods: Ensemble techniques, particularly those combining different types of models, should be further investigated and utilized to harness the strengths of individual models and improve overall predictive performance.

Model Tuning: Further tuning of model hyperparameters and architectures could help reduce misclassification rates and improve overall performance.

Continuous Evaluation: Continuous evaluation and benchmarking of sentiment analysis models against evolving datasets and evaluation metrics are essential to ensure their relevance and effectiveness in real-world applications.

Domain Knowledge Integration**:** Integrating domain knowledge and expertise from financial analysts and investors into model development and interpretation processes can lead to more informed and actionable insights for investment decisions.

Open Research Collaboration: Collaborative efforts between academia, industry, and regulatory bodies can facilitate the development of standardized benchmarks, datasets, and evaluation protocols for advancing research in financial sentiment analysis.

7 REFERENCES

Abdullah, T., Ahmet, A. (2022). Deep learning in sentiment analysis: Recent architectures. ACM Comput. Surv. 55 (8), 1–37.

Ahmad, H.O. and Umar, S.U. (2023). Sentiment analysis of financial textual data using machine learning and deep learning models. *Informatica*, *47*(5).

Alanazi, S.A., Khaliq, A., Ahmad, F., Alshammari, N., Hussain, I., Zia, M.A., Alruwaili, M., Rayan, A., Alsayat, A. and Afsar, S. (2022). Public's mental health monitoring via sentimental analysis of financial text using machine learning techniques. *International Journal of Environmental Research and Public Health*, *19*(15), p.9695.

Antweiler, W. and Frank, M.Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of finance*, *59*(3), pp.1259-1294.

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. ArXiv. abs/1908. 10063. Available from: https//api.semanticscholar.org/CorpusID:201646244.

Atak, A. (2023). Exploring the sentiment in Borsa Istanbul with deep learning. *Borsa Istanbul Review*, *23*, pp.S84-S95.

Bressanelli, G. (2022). Sentiment Analysis of Financial News released during Covid-19 pandemic.

Brown, G.W. and Cliff, M.T., 2004. Investor sentiment and the near-term stock market. *Journal of empirical finance*, *11*(1), pp.1-27

Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A. (2017). Affective computing and sentiment analysis. In: A Practical Guide to Sentiment Analysis. Springer, pp. 1–10.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: Extreme gradient boosting. pp. 1–4, R package version 0.4-2, 1, 4.

DeSola, V., Hanna, K., Nonis, P. (2019). Finbert: pre-trained model on sec filings for financial natural language tasks. University of California.

Devlin, J., Chang, M.W., Lee K., Toutanova K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv181004805.

Du, C.H., Tsai, M.F. and Wang, C.J. (2019, May). Beyond word-level to sentence-level sentiment analysis for financial reports. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1562-1566). IEEE.

Du, K., Xing, F., Mao, R. and Cambria, E. (2024). Financial Sentiment Analysis: Techniques and Applications. *ACM Computing Surveys*.

Esichaikul, V. and Phumdontree, C. (2018, December.) Sentiment analysis of Thai financial news. In *Proceedings of the 2018 2nd International Conference on Software and e-Business* (pp. 39-43).

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G. and Kyriazis, D. (2023). Transforming sentiment analysis in the financial domain with ChatGPT. *Machine Learning with Applications*, *14*, p.100508.

Gutiérrez-Fandiño, A., Kolm, P. and Armengol-Estapé, J. (2021). Fineas: Financial embedding analysis of sentiment. *arXiv preprint arXiv:2111.00526*.

Hasselgren, B., Chrysoulas, C., Pitropakis, N. and Buchanan, W.J. (2022). Using social media & sentiment analysis to make investment decisions. *Future Internet*, *15*(1), p.5.

Hazourli, A. (2022). Financialbert-a pretrained language model for financial text mining. Technical Report.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B. (1998). Support vector machines. IEEE Intell. Syst. Appl. 13 (4), 18–28.

Huang, A., Wang, H., Yang, Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text. Contemporary Accounting Research. Available from: https//api.semanticscholar.org/CorpusID:252666016.

Jabeen, A., Afzal, S., Maqsood, M., Mehmood, I., Yasmin, S., Niaz, M.T. and Nam, Y. (2021). An LSTM based forecasting for major stock sectors using COVID sentiment. *Computers, Materials and Continua*, *67*(1), pp.1-21.

Joseph, A. (2019). Shapley Regressions: A Framework for Statistical Inference on Machine Learning Models. Available at: https://www.kcl.ac.uk/business/assets/pdf/dafm-working-papers/2019-papers/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models.pdf.

Issam, A., Mounir, A.K., Saida, E.M. and Fatna, E.M. (2022). Financial sentiment analysis of tweets based on deep learning approach. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*(3), pp.1759-1770.

Ivanenko, V. (2023). Enhancing aspect-based financial sentiment analysis through contrastive learning. *Innovative Technologies and Scientific Solutions for Industries*, (3 (25)), pp.138-147.

Jim, J.R., Talukder, M.A.R., Malakar, P., Kabir, M.M., Nur, K. and Mridha, M.F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, p.100059.

Kansal, V. and Kumar, R. (2019, March). A hybrid approach for financial sentiment analysis using artificial intelligence and cuckoo search. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)* (pp. 523-528). IEEE.

Karanikola, A., Davrazos, G., Liapis, C.M. and Kotsiantis, S. (2023). Financial sentiment analysis: Classic methods vs. deep learning models. *Intelligent Decision Technologies*, *17*(4), pp.893-915.

Kearney, C., and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis, 33, 171-185.

Kim, J., Kim, H.S. and Choi, S.Y. (2023). Forecasting the S&P 500 index using mathematical-based sentiment analysis and deep learning models: a FinBERT transformer model and LSTM. *Axioms*, *12*(9), p.835.

Kohsasih, K.L., Hayadi, B.H., Juliandy, C. and Pribadi, O. (2022, October). Sentiment Analysis for Financial News Using RNN-LSTM Network. In *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1-6). IEEE.

Lengkeek, M., van der Knaap, F. and Frasincar, F. (2023). Leveraging hierarchical language models for aspect-based sentiment analysis on financial data. *Information Processing & Management*, *60*(5), p.103435.

Li, S., Shi, W., Wang, J. and Zhou, H. (2021). A deep learning-based approach to constructing a domain sentiment lexicon: a case study in financial distress prediction. *Information Processing & Management*, *58*(5), p.102673.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M. and Chen, D. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv. 2019; abs/1907.11692. Available from: https//api.semanticscholar.org/CorpusID:198953378.

Liu, B. (2022). Sentiment Analysis and Opinion Mining. Springer Nature.

Liu, Z., Huang, D., Huang, K., Li, Z., Zhao, J. (2021). Finbert: A pre-trained financial language representation model for financial text mining. In: Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence; pp. 4513-9.

Lloyd, S. (1952). N-Person Games. Defense Tech. Inf. Cent., 295–314. 10.7249/p0295

Lundberg, S. M., Lee S. I. (2017). A Unified Approach to Interpreting Model Predictions. Adv. Neural Inf. Process. Syst. 30, 4765–4774. Available at: https://arxiv.org/abs/1705.07874

Maia, Macedo & Handschuh, Siegfried & Freitas, Andre & Davis, Brian & McDermott, Ross & Zarrouk, Manel & Balahur, Alexandra. (2018). WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. WWW '18: Companion Proceedings of the The Web Conference 2018. 1941-1942. 10.1145/3184558.3192301.

Malo, Pekka, et al. (2014). "Good debt or bad debt: Detecting semantic orientations in economic texts." Journal of the Association for Information Science and Technology 65.4 (2014): 782-796.

Memiş, E., Akarkamçı, H., Yeniad, M., Rahebi, J. and Lopez-Guede, J.M. (2024). Comparative Study for Sentiment Analysis of Financial Tweets with Deep Learning Methods. *Applied Sciences*, *14*(2), p.588.

Methmal, T.H.H. (2020). *Sentiment analysis for financial market prediction* (Doctoral dissertation).

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1–2), pp.1-135.

Patel, K. (2021). *Justification Mining: Developing a novel machine learning method for identifying representative sentences and summarising sentiment in financial text* (Doctoral dissertation, University of Oxford).

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M.Z., Barrow, D.K., Taieb, S.B., Bergmeir, C., Bessa, R.J., Bijak, J., Boylan, J.E. and Browell, J. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, *38*(3), pp.705-871.

Revathy, G., Alghamdi, S.A., Alahmari, S.M., Yonbawi, S.R., Kumar, A., Haq, M.A. (2022). Sentiment analysis using machine learning: Progress in the machine intelligence for data science. Sustain. Energy Technol. Assess. 53, 102557.

Ribeiro M. T., Singh S., Guestrin C. (2016). "Why Should I Trust You?". Knowledge Discov. databases. 16, 1135–1144. 10.1145/2939672.2939778

Sanjay, K. S. (2017).” Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms” International Journal of Engineering And Computer Science ISSN:2319-7242Volume6Issue.

Shams, R., Khosravian, J. and Samimi, P. (2024). Enhancing Financial Sentiment Analysis with a Hybrid Feature Selection Approach.

Sharma, N., Soni, M., Kumar, S., Kumar, R., Deb, N. and Shrivastava, A., 2023. Supervised Machine Learning Method for Ontology-based Financial Decisions in the Stock Market. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(5), pp.1-24.

Sohangir, S., Wang, D., Pomeranets, A. and Khoshgoftaar, T.M. (2018). Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, *5*(1), pp.1-25.

Syeda, F.S. (2022). Sentiment Analysis of Financial News with Supervised Learning.

Sy, E., Peng, T.C., Huang, S.H., Lin, H.Y. and Chang, Y.C. (2023, October). Fine-grained argument understanding with bert ensemble techniques: A deep dive into financial sentiment analysis. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)* (pp. 242-249).

Thelwall, M., Buckley, K. and Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), pp.406-418.

Tetlock, P.C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), pp.1139-1168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. and Gomez, A.N. (2017). Attention is all you need. Advances in Neural Information Processing Systems. 2017; 30.

Vicari, M. and Gaspari, M. (2021). Analysis of news sentiments using natural language processing and deep learning. *AI & society*, *36*(3), pp.931-937.

Yang, L., Li, Y., Wang, J., Sherratt, R.S. (2020). Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. IEEE Access. 82: 3522-30.

Yekrangi, M. and Nikolov, N.S. (2023). Domain-Specific Sentiment Analysis: An Optimized Deep Learning Approach for the Financial Markets. *IEEE Access*.

Yıldırım, S., Jothimani, D., Kavaklioğlu, C. and Başar, A. (2019, December). Deep learning approaches for sentiment analysis on financial microblog dataset. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 5581-5584). IEEE.

Zhang, B., Yang, H., Zhou, T., Ali Babar, M. and Liu, X.Y. (2023, November). Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance* (pp. 349-356).

Zhou, L. and Chaovalit, P. (2008). Ontology-supported polarity mining. *Journal of the American Society for Information Science and technology*, *59*(1), pp.98-110.