A COMPARATIVE STUDY ON MACHINE LEARNING BASED STOCK PREDICTION BY INCORPORATING SENTIMENT ANALYSIS USING FINBERT

A Thesis Submitted
In Partial Fulfillment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
Data Science
by

Vibhor Singh
(2K23/DSC/01)

Under the Supervision of
Dr. Rahul
Assistant Professor, Department of Software Engineering
Delhi Technological University

Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2025

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

ACKNOWLEDGEMENT

I take the opportunity to express my sincere gratitude to **Dr. Rahul**, Department of Software Engineering, Delhi Technological University, Delhi, for providing valuable guidance and constant encouragement throughout the work. His knowledge and expertise helped me at every stage of the preparation of this research plan.

I am extremely thankful to **Prof. Ruchika Malhotra**, Head of the Department, Department of Software Engineering, Delhi Technological University, Delhi & other faculty members of the Department of Software Engineering, Delhi Technological University, Delhi, for the motivation and inspiration. I am also thankful to all non-teaching staff at DTU, who have helped me directly or indirectly in the completion of this research plan.

Vibhor Singh
2K23/DSC/01

M.Tech Scholar (Full Time)

Department of Software Engineering

Delhi Technological University

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I Vibhor Singh hereby certify that the work which is being presented in the thesis entitled "A comparative study on machine learning based stock prediction by incorporating sentiment analysis using FinBERT" in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Data Science, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from 2023 to 2025 under the supervision of Dr. Rahul.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s) **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **Vibhor Singh** (2K23/DSC/01) has carried out his search work presented in this thesis entitled **"A comparative study on machine learning based stock prediction by incorporating sentiment analysis using FinBERT"** for the award of **Master of Technology** from Department of Software Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Rahul Supervisor
Assistant Professor
Department of Software Engineering
Delhi Technological University

Place: New Delhi
Date: 20/05/2025

ABSTRACT

Swift evolution of ML techniques, the financial forecasting scenario have been transformed tremendously. The domain of stock market prediction is one of the prominent applications of ML considering its inherent complexity and economic implication. Time series and statistical model was conventional cornerstone of market analysis, but they have limited potential for capturing intricate pattern. NLP sentiment analysis played a game changing role for more efficient stock prediction. NLP taps into huge pool of textual data of different sources, comprising of news and social media outlets. The collective mood and opinion of market participants can now be harness using power of NLP sentiment analysis which is to be fed to predictive model for better forecasting. Fusion of sentiment derived insights with ML algorithms presents a substantial leap which not only surges the predictive power of existing models but also provide nuanced understanding of the psychology of market movements driving factors. Consequently, financial industry witnessing a paradigm shift for the anticipation of stock prices fluctuations, with the support of AI driven sentiment analysis

This paper presents a machine learning-based stock prediction model that integrates sentiment analysis using FinBERT, it is a specialized model for financial sentiment analysis that uses BERT. This study focuses on enhancing financial stock forecasting by adding investors sentiment data with conventional stock price data. This study takes into consideration traditional time series model like SARIMA for stock price prediction and three FinBERT infused ML models namely SVR, RFR, GBR. Eventually all predictive models are compared through regression evaluation metrics like MAE, MSE, RMSE, R^2.

TABLE OF CONTENTS

Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
APPL	Apple Inc.
ARIMA	Autoregressive Integrated Moving Average
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
CSI300	China Securities Index 300
DL	Deep Learning
FinBERT	Financial Bidirectional Encoder Representations from Transformers
FTSE 100	Financial Times Stock Exchange 100 Index
GBR	Gradient Boosting Regression
GloVe	Global Vectors for Word Representation
KNN	K-Nearest Neighbors
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NASDAQ	National Association of Securities Dealers Automated Quotations
NLP	Natural Language Processing
R^2	Coefficient of Determination
RF	Random Forest
RFR	Random Forest Regression
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
S&P 500	Standard & Poor's 500
SARIMA	Seasonal Autoregressive Integrated Moving Average
SVR	Support Vector Regression
VADER	Valence Aware Dictionary and sEntiment Reasoner
Word2Vec	Word to Vector
XGBoost	Extreme Gradient Boosting

CHAPTER 1

INTRODUCTION

This chapter presents an introduction to stock market prediction by enabling the integration of sentiment analysis with financial forecasting. This chapter consists of background, problem statement, motivation, contribution, thesis organization and finally concluded with a summary.

1.1 OVERVIEW

This research conducts a comparative evaluation of various machine learning (ML) models enhanced with sentiment analysis techniques for forecasting stock prices, with a particular emphasis on FinBERT, a transformer model tailored for financial language. This study takes into consideration traditional time series model like SARIMA for stock price prediction and three FinBERT infused ML models namely SVR, RFR, GBR. These models are integrated with sentiment insights extracted from financial news using FinBERT. The analysis uses Apple Inc. (AAPL) stock data along with corresponding sentiment information spanning from February 3, 2023, to February 3, 2025. Eventually all predictive models are compared through regression evaluation metrics like MAE, MSE, RMSE, R^2. The findings indicate that ML models incorporating FinBERT-based sentiment analysis outperform others, with the FinBERT+GBR combination showing the highest predictive accuracy. This underscores the potential of merging sentiment-driven insights with modern ML algorithms for improved financial market prediction.

1.2 BACKGROUND

Predicting stock market movements remains a complex task due to the unpredictable and volatile nature of financial data. Traditional forecasting approaches, such as the SARIMA model, have been commonly applied to time-series data by utilizing historical trends and seasonal patterns. Despite their usefulness, these models often fall short when dealing with the non-linear behaviors and abrupt fluctuations characteristic of stock markets. In response, ML methods like SVR, RFR, and GBR have gained traction, as they are better equipped to handle intricate and high-dimensional data structures. At the same time, advancements in NLP have made it possible to derive meaningful insights from textual sources such as news articles and social media posts, which often capture public sentiment. FinBERT, it is a specialized model for financial sentiment analysis that uses BERT has proven highly effective in analyzing sentiment within financial contexts. Studies have increasingly demonstrated that combining sentiment analysis with ML algorithms improves forecasting performance, driving interest in hybrid modeling strategies for financial market prediction.

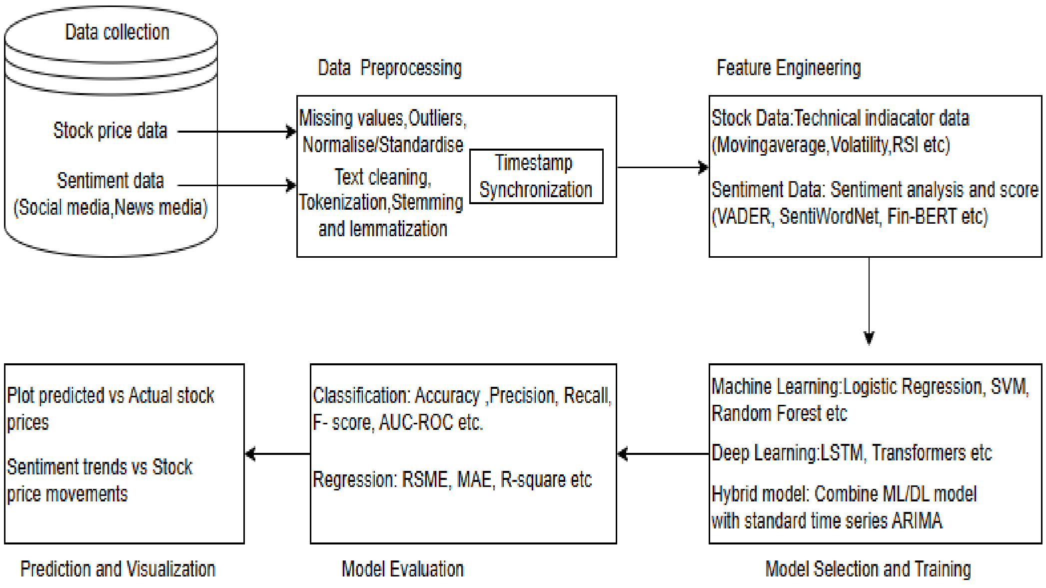


Fig. 1. 1 Generic steps for sentiment-based stock prediction

Initially stock data and sentiment data is collected from API like Twitter API and platform like StockTwits and from news portals like moneycontrol, Livemint financial etc. Different investigations have demonstrated that stock market data alone is sufficient to project stock valuations [1].

The next stage is to process the data which is essential since text data is usually disorderly and requires sorting and structuring. Some of techniques like tokenization, stop word removal, stemming, and lemmatization used for preparing the sentiment data. Text is represented in numbers using Bag of Word, term frequency inverse document frequency techniques [2]. Some advanced techniques also involve using model embeddings such as Word2Vec or GloVe embeddings to capture word semantics, thus providing richer features for sentiment analysis. Similarly, Wen jun Gu et al. integrate FinBERT and StockTwits to enhance the sentiment extraction process [3]. Advanced machine learning tools like web scraping (BeautifulSoup, Scrapy) and APIs streamline data collection, enabling effective integration of structured stock price data and unstructured sentiment data for predictive modelling.

Such research methods bestow correct sentiment scores from various text sources acquire phenomenal predictive analytics in financial markets. The prepared data will be fed into these algorithms for model selection and training after data processing. RF and XGBoost are just a couple of examples of some of these traditional techniques that are interpretable and generally efficient in modelling structured data. At the same time, other deep learning models include LSTM and attention-based architectures. They are however preferred in time series analysis as they understand the temporal dependencies and more complex patterns shown by financial market behavior. Non-linear patterns may be addressed through the hybrid models that integrate and use both deep and machine learning. The model can finally be assessed with classification metrics such as accuracy, F1-score, or regression metrics such as MAE, and RSME etc. The prediction and decision-making are all made from the best-performing model and deployed in different scenarios by the real-time feeding process based on sentiment data and market information to enable stock price forecasts or trading signals. Fig. 1.1 shows workflow of any generic predictive sentiment based stock prediction model.

1.3 PROBLEM STATEMENT

Traditional time-series models like SARIMA are limited in capturing the non-linear and volatile nature of stock price movements, resulting in suboptimal predictive performance. While ML models offer improved capabilities, their effectiveness can be further enhanced by incorporating contextual insights from market sentiment, which is often embedded in unstructured textual data. Existing studies have explored sentiment analysis in financial forecasting, but there is a lack of comprehensive comparative analyses evaluating the integration of transformer-based sentiment models, such as FinBERT, with diverse ML regression models. Additionally, the relative performance of these hybrid models against traditional approaches in specific market contexts, such as predicting AAPL stock prices, remains underexplored. This study of comparative model focuses on the gap by comparing SARIMA and FinBERT induced ML models to ascertain their effectiveness in financial forecasting.

1.4 MOTIVATION

This study provides motivation to improve stock price prediction in the ever-changing and dynamic market complexity and instability. Today AI- driven solutions provides financial industry a boon to anticipate stock price fluctuations effectively, which results in wider implication for investor, traders, and stakeholders decision making. This study integrates sentiment analysis with ML, which results in opening promising avenue to calculate both quantitative market data and qualitative sentiment of market providing overall understanding of price fluctuation drivers. FinBERT is capable of extracting domain related sentiment from texts related to finance, making it an effective tool for enhancing predictive models effectiveness. This comparative study compared traditional time series model and sentiment induced ML based models gives actionable insights into most effective methods for stock forecasting, which contribute hugely to the contemporary paradigm shift towards AI driven financial solutions.

1.5 CONTRIBUTION

This research significantly advances financial forecasting by thoroughly comparing a conventional time-series approach, SARIMA, with three ML models: SVR, RFR, and GBR, each enhanced by sentiment analysis derived from FinBERT. By integrating FinBERT to analyze sentiments from financial news, the study shows how this approach strengthens the predictive power of machine learning models for stock price forecasting. Using real-world data from Apple's stock and related news, the analysis demonstrates that the Gradient Boosting Regression model, combined with FinBERT, outperforms SARIMA and other machine learning methods, achieving a MAE of 2.5004 and an R^2 value of 0.9864. These results provide actionable insights for financial professionals looking to implement AI-based forecasting tools, emphasizing the value of blending sentiment analysis with sophisticated machine learning techniques. The study also suggests future research paths, such as incorporating diverse data sources, exploring deep learning frameworks, and refining model parameters to further improve stock prediction methods.

1.6 THESIS ORGANIZATION

This thesis included of seven chapters, each chapter covers different aspect of stock price prediction with fusion of incorporation of sentiment analysis. Chapters are arranged in a coherent and logical manner from introduction to conclusion.

1.6.1. Chapter 1: Introduction

This chapter introduces the thesis by discussing the history and necessity of sentiment centered stock prediction, the evolution of monetary forecasting models, and the relevance of this research. It includes the problem statement, motivation, and main contributions of the study. The chapter settles with an overview of the thesis structure.

1.6.2. Chapter 2: Literature Survey

Chapter 2 analyses prevailing literature related to financial forecasting models. It includes a summary table of 20 significant papers, highlighting their methodologies and findings, and identifies research gaps that this thesis aims to address.

1.6.3. Chapter 3: Research Objectives

Chapter 3 defines the research objectives and questions to support the research. It delivers a clear overview of the aims and a summary of the specific objective of the research.

1.6.4. Chapter 4: Methodology

Chapter 4 describe the research methodology which includes data collection, data preprocessing, model training, and model evaluation. A proposed high-level design was constructed for reader's better understanding. At last, the experimental setup and Python libraries/modules are also mentioned.

1.6.5. Chapter 5: Results and Discussion

Chapter 5 provide the experimental results which includes result from model training with a comparison with existing financial forecasting models. Lastly, it mentions a discussion of the results to point out the limitations and key points of model performance.

1.6.6. Chapter 6: Conclusion, Future Scope, and Social Impact

Chapter 6 summarizes all the research findings in this study, then it discusses the future scope of this study, and at last it states the social impact of financial forecasting in real world.

1.6.7. References

The section lists down all references cited in the thesis which were used for successful completion of experimental analysis and it also helps to support the credibility of this study.

1.7 SUMMARY

This thesis, organized into seven chapters, systematically investigates the incorporation of sentiment analysis with ML to enhance stock market prediction, offering a logical progression from introduction to conclusion. Chapter 1 introduces the significance of sentiment infused stock prediction, tracing the evolution of financial forecasting models, and outlines the problem statement, motivation, contributions, and thesis structure. Chapter 2 reviews literature on financial forecasting, summarizing 20 key papers in a table to highlight methodologies, findings, and research gaps addressed by this study. Chapter 3 defines the research objectives and questions, clarifying the aim to improve predictive accuracy through sentiment integration. Chapter 4 details the methodology, encompassing data collection, preprocessing, model training (e.g., SVR, RFR, GBR, SARIMA), and evaluation, supported by a high-level design and Python libraries like yfinance. Chapter 5 presents experimental results, comparing sentiment-enhanced models against traditional approaches using metrics like MAE and R^2, and discusses performance limitations and strengths. Chapter 6 settles with a summary of answers, suggests future research directions such as incorporating varied data sources, and highlights the social impact of improved financial forecasting for economic decision-making. Chapter 7 lists all references, ensuring the study’s credibility and grounding the experimental analysis in established literature.

CHAPTER 2

LITERATURE SURVEY

This chapter exhibits a literature survey the field of sentiment-driven stock prediction, an area of research that aims to predict market direction combining traditional financial analysis tasks and sentiment generated by investors and the community. It details some of the key steps that are inherent to a prediction pipeline (data acquisition, pre-processing, sentiment analysis, feature engineering and predictive modelling). Sentiment analysis enhances the stock direction prediction. This article summarizes various ways of sentiment-based stock prediction by reviewing various research articles and highlights increase in performance of predictive model which incorporate sentiment analysis in their algorithm.

2.1 OVERVIEW

The whole investment decision-making process can benefit greatly from the merging of investor emotions with traditional financial analysis. This innovative mix allows market participants to do well in capitalising on emerging trends in technology, going ahead to exploit the vast experience of leveraging sentimentally social media analytics. Various research exemplifies how an easier and probably faster methodology of sentiment-based stock prediction models gives rise to enhanced financial profits. For instance, N. J. Li et al. obtained predictions from LSTM leading to an astonishing prediction accuracy of 87.86% with respect to the CSI300 index [4]. X. Guo et al. have shown how the twitter sentiment score model was able to forecast trending for the FTSE 100 index by accuracy of 67.22%, outperforming traditional econometric models [5].

The combination of sentiment analysis, stock exchanges' historical data, and modern machine learning techniques has led to enhanced prediction accuracy. Among these hybrid models are CNNs for sentiment classification and LSTMs for technical analysis, which performed even better than models without such features and even single models in forecasting stock prices on the Shanghai Stock Exchange, according to work from Jing et al. (2021) [6]. This adds credence to the use of sentiment-based models to increase predictive ability for those who approve of predicting future trends in the stock market.

2.2 REVIEW OF RELATED WORKS

D. Shah et al. aim to determine how news sentiments affect stock market performance, especially in the pharmacological industry. It entails building a financial sentiment analysis dictionary and creating a dictionary-based sentiment analysis model. After employing this customised dictionary, the model achieves accuracy of 70.59% for short term prediction [7]. J. Huang et al. looks at how stock prices are determined by social media viewpoints, using text scrapping techniques, transforming unstructured data from social media captures the researchers' attention since there is strong association between sentiments and variation of stock prices. It is noted that higher sentiment scores lower the classification accuracy of a traditional LR model, thereafter response to this problem, they proposed an advanced model through which sentiment scores can be incorporated into the system. Indeed, this method rises the exactness of the stock price prediction which is beneficial to investors as it enables them to formulate better investment plans [8].

A. Kanavos et al. focuses on the relationship between Twitter sentiments and stock scores leveraging sentiment analysis. It scrapes and analyses Twitter posts daily using n-grams and classification techniques to improve precision. Processing data through Spark Streaming, the idea is to monitor, filter, and sample Twitter data from FinTwit [9]. M. Kesavan et al. proposed system combines sentiment analysis and stock-market forecast based on time-series data and DL. Their pipeline comprises sentiment extraction from social media and news forum such as Twitter, the task of sentiment polarity has been applied to the outcome which results increasing the prediction accuracy and helps investors to take wiser decision [10]. X. Li et. al. utilise the technical indicator from the stock price information and the news article sentiment data, then build a DL model to obtain the serial relationship of time series data. Eventually with the help of four different type of sentiment dictionary using five-year Hong Kong Stock Exchange data reveals that Loughra-McDonald financial dictionary perform better than remaining dictionaries [11].

I. K. Nti et. al. analysed stock data from the Ghana Stock Exchange (2010–2019) to predict future prices over various timeframes. Prediction accuracy improved when multiple data source: Google Trends, Twitter, forums, and financial news were combined, reaching up to 77.12%. The results also emphasize the existence of a solid association between social media and stock market activities which indicates that investors can mark use of online data to well predict the market and make informed investment choices [12]. Y. Huang et. al. called for the construction of a multichannel collective network for stock price forecasting using social media sentiment analysis and representing the results in candlesticks chart. Sentiment features were extracted from Twitter using NLP tools, while historical stock data was converted into candlestick charts to capture price movement patterns. The model integrates both data types using two CNN branches: one for sentiment classification and another for image-based pattern recognition. When tested on five major stocks: Google, Tesla, Apple, IBM, and Amazon, the approach outperformed single-data-source models, achieving a 75.38% accuracy for Apple. Longer prediction periods yielded better results than shorter ones [13]. W. J. Liu et. al. set forth a model that considers news across multiple sources alongside stock prices and predicts stock market indices. Combination of sentiment attention mechanism and TrellisNet (SA-TrellisNet) uses the news and stock sentiment index, generated with CNN and LSTM for news and stock tweets, to weight stock data and improve TrellisNet training, prediction accuracy, and forecasting. The architecture is assessed with seven major stock indices, including NASDAQ, and S&P 500, and it competes with existing prediction methods [14].

N. Das et. al. studied how public sentiment influenced stock market predictions of the Corona virus pandemic using LSTM architectures. The study employs VADER, LR, Lonughra-McDonald, Henry, and other seven sentiment analysis tools on scraped data from stock headline, tweets, financial news, Facebook comments etc. The research finds that Linear SVC-based sentiment scores derived from Facebook comments achieve 98.11% accuracy. When sentiment scores were integrated within the stock data, the prediction accuracy of 98.32% was achieved, marking the extreme effect emotions have on the actions of the stock market [15]. R. Chiong et. al. proposes an ensemble RNN model, LSTM, GRU, SimpleRNN fusion, for stock forecast. By using sentiment examination and the sliding window technique, they effectively extract key features, outperforming other models in comparison [16]. Fig. 2.1 shows publisher- wise distribution of research papers used in this study and Fig. 2.2 depicts year wise distribution of research paper used in study. Table 2.1 presents brief description of research papers that are used in this study for literature review.

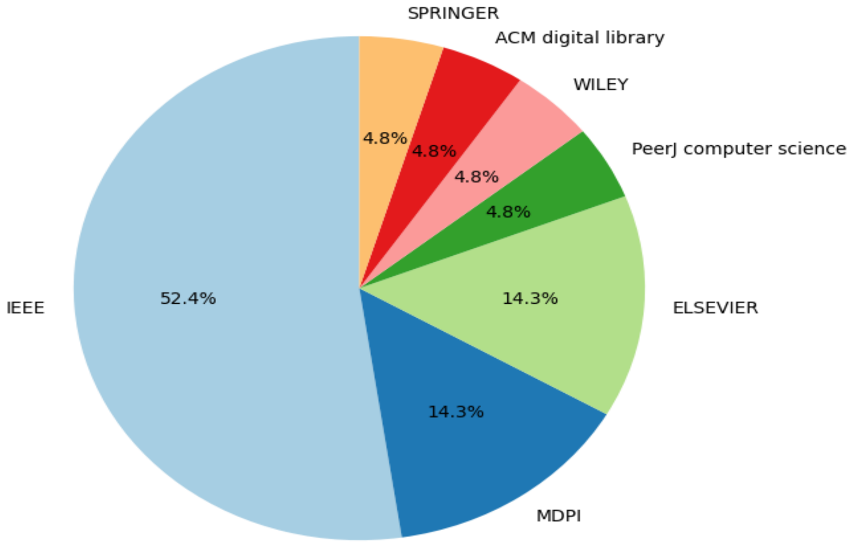


Fig. 2. 1 Publisher distribution.

Ref.	Stock Data	Sentiment Data	Time Period	Sentiment analysis	Predictive model	Model evaluation result
[17]	National Stock Exchange (NSE): Stocks of BOB, PNB, HDFC, ICIC	NDTV, Livemint, Financial Express, Moneycontrol and Business Today.	August 2018 and September 2018	NAIVE BAYES	KNN, SVM, NAIVE BAYES, NN	Accuracy, Precision, Recall, F1 Measure
[18]	Colombo Stock Exchange: ASPI, S&PSL20	Twitter API	Three time periods: Sep-Dec 2009, Jan-Mar 2015, and May-Jul 2020.	LR, NAIVE BAYES	SVM, NAIVE BAYES, KNN, LR, Decision tree	Accuracy
[19]	Historical stock data: SPDR S&P 500 Index ETF (SPY)	Stocktwits	April 1, 2020, to February 28, 2022	FinBERT	Ensemble SVM	Accuracy, Precision Recall, F1-score, Measure
[20]	Yahoo Finance: StockNet, NASDAQ stock exchange.	Twitter API	1 January 2014 to 1 January 2016	Neutro-sophic Logic (NL) integrated VADER	LSTM	Accuracy, Matthews's correlation coefficient (MCC)
[21]	Yahoo Finance, for Apple stock	StockTwits	January 2010 to March 2017	SVM	SVM	Accuracy, Precision, Recall
[22]	Yahoo Finance, stocks analyzed - Microsoft (MSFT)	Twitter, StockTwits	16 July 2020 to 31 October 2020	TextBlob, VADER	MLP, SVM, KNN, LR, DT, RF, NB,	AUC, F1-score
[23]	Yahoo Finance API: Microsoft (MSFT), Amazon (AMAZ), and Tesla (TSLA)	The New York Times	January 1, 2015, to August 13, 2020	VADER	LSTM	MSE, MAE
[24]	Yahoo Finance, Stocks: Microsoft (MSFT).	Twitter, StockTwits	16-07-2020 to 31-10-2020	TextBlob, VADER	SVM and LR	Accuracy, F-score, AUC
[25]	6 stocks from the Chinese stock market	East Money forum	July 1, 2016, to June 30, 2022	Customised financial dictionary	MS-SSA-LSTM	MAPE, MAE, and RMSE, R-squared
[26]	9 high volume stocks in BIST 100 (the Istanbul Stock Exchange)	Twitter, Mynet Finans, Public Disclosure Platform (KAP), Bigpara	September 1, 2018, to September 1, 2019	TextBlob	(Word2Vec, GloVe, FastText) + (CNN, RNN, LSTM)	Accuracy

Table. 2. 1 Description For Prominent Research Papers For Sentiment-Based Stock Prediction

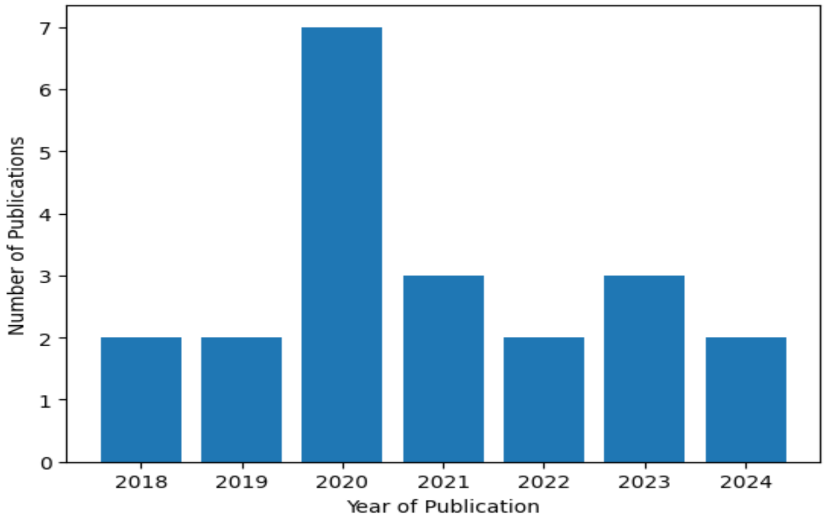


Fig. 2. 2 Year-wise publication.

2.3 RESEARCH GAPS

2.3.1. Research Gap 1 (Limited Use of Domain-Specific Sentiment Models)

Description: While sentiment analysis has been increasingly applied to stock market prediction, many studies rely on general-purpose sentiment models (e.g., VADER, TextBlob) that fail to capture the nuanced sentiment in financial texts. These models often miss domain-specific terminology and context, leading to suboptimal sentiment features for predictive modeling. There is a lack of studies leveraging specialized financial sentiment models, like FinBERT, to improve the correctness of stock price predictions.

Proposed Solution: This study addresses the gap by utilizing FinBERT, a model for financial sentiment analysis which uses BERT, to extract high-quality sentiment features from financial news. By integrating FinBERT-derived sentiment with conventional stock features, the proposed model (combining SARIMA, SVR, RFR, and GBR) captures domain-specific market sentiment, improving prediction correctness as demonstrated by lower MAE, MSE, RMSE, and greater R^2 compared to baseline models.

2.3.2 Research Gap 2 (Lack of Comprehensive Model Comparison with Sentiment Integration)

Description: While sentiment analysis is recognized as a game-changer in stock prediction, few studies systematically compare the performance of multiple ML models incorporating sentiment data against traditional time series models. This limits the understanding of which models best leverage sentiment features and under what conditions, hindering practical adoption in financial forecasting.

Proposed Solution: The proposed study fills this gap by steering a inclusive contrast of four predictive models: SARIMA (time series) and three ML fusion models (SVR, RFR, GBR) all incorporating FinBERT-derived sentiment from financial news. The use of standardized evaluation metrics (MAE, MSE, RMSE, R^2) provides a robust benchmark to recognize the furthermost effective model for sentiment augmented stock forecast, offering actionable insights for financial practitioners.

2.4 SUMMARY

The literature survey examines sentiment-driven stock prediction, which combines traditional financial analysis with investor and community sentiment to predict market trends. It details the prediction pipeline, encompassing data acquisition, preprocessing, sentiment analysis, feature engineering, and predictive modeling, with sentiment analysis significantly improving stock direction forecasting. The review synthesizes research demonstrating enhanced performance of sentiment-based models, such as N. J. Li et al.'s LSTM model achieving 87.86% accuracy for the CSI300 index [4] and X. Guo et al.'s Twitter sentiment model yielding 67.22% accuracy for the FTSE 100 [5]. Hybrid models integrating sentiment, historical stock data, and advanced machine learning techniques like CNNs and LSTMs, as shown by Jing et al., outperform traditional models on the Shanghai Stock Exchange. Notable studies include Shah et al.'s dictionary-based model for pharmaceuticals (70.59% accuracy) and Huang et al.'s multichannel CNN model for major stocks (75.38% accuracy for Apple), highlighting the influence of news and social media emotion. However, gaps remain, such as the limited adoption of domain-specific sentiment models like FinBERT and insufficient comparisons of sentiment-integrated machine learning models against time series models. The proposed study addresses these by employing FinBERT for financial sentiment analysis and comparing SARIMA with ML models (SVR, RFR, GBR) using MAE, MSE, RMSE, and R^2, offering improved accuracy and practical insights for financial forecasting.

CHAPTER 3

RESEARCH OBJECTIVES

This chapter states the main aim for conducting the research by solving the framed research questions. This chapter was created for solving the research gaps found in the relevant related works stated in above chapter. Additionally, the study queries will be replied in below chapters using an experimental analysis.

3.1 OVERVIEW

The advancement of ML and NLP has meaningfully reshaped the background of stock market anticipation, particularly by enabling the incorporation of investor sentiment alongside conventional financial indicators. However, existing research still presents notable limitations—chief among them being the underutilization of sentiment models customized to the financial domain and the scarcity of thorough evaluations comparing sentiment-augmented ML approaches with traditional time-series forecasting techniques. This study seeks to link these gaps by suggesting a stock prediction framework that fit in sentiment analysis using FinBERT, a model specifically optimized for interpreting monetary sentiment using BERT. The extracted sentiment signals from financial news are combined with standard stock market data to train and assess the performance of three ML models: SVR, RFR, and GBR in comparison with the classical SARIMA model. The objective is to measure the influence of domain-specific sentiment features to predictive accuracy and to classify the utmost active method for sentiment-informed stock forecast. Through this, the research aspires to offer valuable insights for both scholarly inquiry and real-world financial decision-making

3.2 RESEARCH QUESTIONS

To monitor the study and report the recognised research gaps, the following investigation questions are proposed:

3.2.1 How does the use of a domain-specific sentiment model like FinBERT improve the correctness of stock price forecasting related to general-purpose sentiment models?

This question explores the effectiveness of FinBERT in capturing nuanced financial sentiment from news data, addressing the gap in the limited adoption of domain-specific sentiment models.

3.2.2 To what extent do sentiment-augmented ML models (SVR, RFR, GBR) outperform a old-fashioned time series model (SARIMA) in stock price estimating?

This question investigates the comparative performance of ML fusion models versus SARIMA, addressing the gap in comprehensive model comparisons for sentiment-integrated stock prediction.

3.2.3 How can the integration of FinBERT-derived sentiment features with conventional stock features enhance the predictive power of ML models?

This question examines the role of feature engineering in combining sentiment and financial data, contributing to improved model performance and addressing the gap in sentiment-driven feature optimization.

3.2.4 What are the practical implications of using sentiment-driven ML models for financial forecasting, particularly in terms of model interpretability and adoption in the financial industry?

This question assesses the real-world usage of the proposed models, exploring their interpretability and potential barriers to adoption, thereby providing actionable insights for practitioners.

3.3 SUMMARY

This chapter presents research objectives and questions designed to fill the critical gaps in sentiment induced financial prediction. This study works to improve the prediction of stock price and also provide a complete understanding of sentiment’s part in stock forecast by employing FinBERT for finance domain specific sentiment and compare it with time series model and sentiment induced ML models. By comparing performance of different models, optimizing feature integration, evaluating practical application research question evaluates effectiveness of FinBERT. Results of this study provides valuable insights by providing a robust framework for sentiment induced stock prediction for financial investors and all stake holders .

CHAPTER 4

METHODOLOGY

Chapter 4 clarifies the procedure used for comparative study aimed at enhancing stock price forecast by assimilating emotion analysis with ML techniques. This chapter showcases a detail work of the proposed work such as collection of data, data pre-processing, model training, and model evaluation. The setup used during this work is also explained for detailed view of the work.

4.1 OVERVIEW

This chapter presents the methodology used in a comparative analysis designed to improve stock price forecast by combining emotion analysis with ML techniques. The approach utilizes FinBERT, a model custom-made for the financial domain which uses BERT to extract sentiment information from financial news. These sentiment features are integrated with traditional stock market data to enhance forecasting accuracy. The study evaluates the performance of a classical time-series model, SARIMA alongside three ML models: SVR, RFR, and GBR, all incorporating sentiment insights derived from FinBERT. The procedure includes steps such as data collection, preprocessing, model development, and performance evaluation. The analysis is based on Apple Inc. (AAPL) stock prices and relevant financial news spanning from February 3, 2023, to February 3, 2025. Model effectiveness is measured using widely accepted regression metrics: MAE, MSE, RMSE, and R^2. By accepting this systematic method, the study goals to report key research gaps, particularly the underuse of financial-domain sentiment models and the lack of in-depth comparisons between sentiment-enhanced ML and traditional forecasting methods, thereby offering a comprehensive framework for sentiment-driven market prediction.

4.2 PROPOSED WORK

The proposed methodology involves a systematic pipeline for stock price prediction, integrating sentiment analysis with machine learning. Fig. 4.1 in the below section illustrates the workflow, which includes data collection, preprocessing, sentiment analysis using FinBERT, model training, and performance evaluation. The following subsections detail each component of the proposed work.

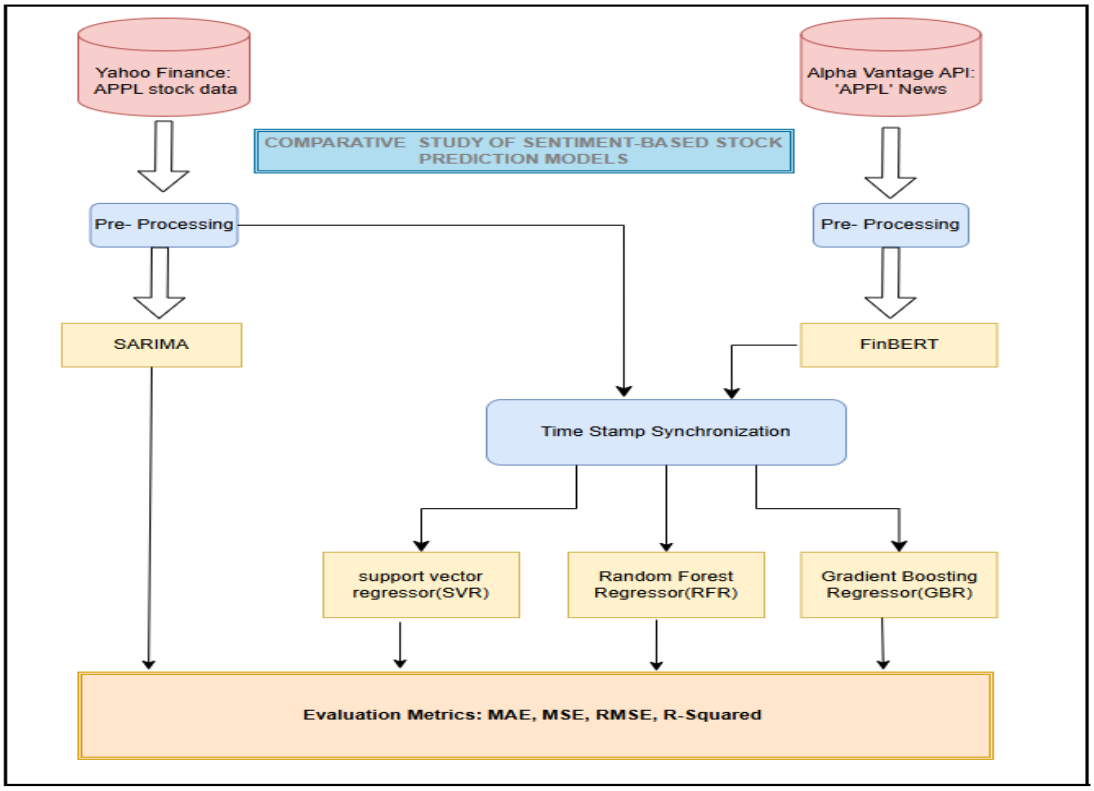


Fig. 4. 1 workflow of proposed comparative study

4.2.1 Data collection

This comparative study employs two primary datasets which are as follows:

a) **APPL Stock dataset**: This study employed Yahoo Finance Python library i.e. yfinance which retrieve historical stock data of Apple Inc. (APPL), traded on NASDAQ. The retrieve dataset obtains five important fields namely open, high, low, close, and volume of APPL stock. These fields provide a strong numerical base of APPL stock for predictive modelling.

b) **News Dataset**: News article related to APPL ticker have been retrieved from API of Alpha Vantage news which act as sentiment data for FinBERT analysis. This dataset majorly contains news article title and summary of article with timestamp, sentiment score generation have been done by feeding both title and summary of news article to the FinBERT. The time period of news article is aligned with stock two-year dataset to ensure temporal consistency.

4.2.2 Data pre-processing

Data processing is significant to check the value and compatibility of the stock and sentiment datasets for model training. The preprocessing steps are as follows::

a) **Stock Data Preprocessing**:

Missing values in stock dataset due to non-trading days are managed by forward filling method to maintain time series data continuity. Normalisation of numerical features have been done using Min-max scaling to standardise their ranges, while maintaining their compatibility with ML algorithms. Close price of next day has been selected as target variable, while remaining fields like open, low, high, volume taken as input features

b) **Sentiment Data Preprocessing**:

News article have been cleaned by eliminating elements that are not relevant like special characters, HTML tags, stop words, to emphasize on important content. Words of news articles have been divided into tokens by tokenization that aligns with the input requirements of FinBERT. Each news article title and summary text are fed to FinBERT, which in turn classifies each article sentiment score on timestamp order. Later sentiment score aggregation is done for each day to align temporal resolution of numerical stock data. Finally, each day sentiment score and numerical stock data features is integrated which will act as a unified dataset for every FinBERT induced ML models for comparison

4.2.3 Model Training

Four predictive models are trained to forecast AAPL stock prices, with FinBERT-derived sentiment scores integrated into the three machine learning models. The models are:

i. SARIMA

Overview

The SARIMA model is a statistical time-series forecasting method that covers the ARIMA model by incorporating seasonal patterns. It is designed to collect both non-seasonal and seasonal trends, cycles, and dependencies in time-series data, making it suitable for financial forecasting tasks like stock price prediction. In this study, SARIMA serves as the baseline model, trained solely on historical stock data without sentiment features, to provide a point of comparison for the sentiment-augmented machine learning models.

Theoretical Foundation

SARIMA contains three important components namely Autoregressive, Integrated and moving average in addition to seasonal fluctuation component to effectively predicts the time series data. The SARIMA is mathematically denoted by SARIMA(p,d,q) (P,D,Q,s) where p, d, q are the component of non-seasonal part representing autoregressive, differencing and moving average respectively and P, D, Q are the seasonal component part representing autoregressive, differencing and moving average respectively, s denotes the length of cycle.

SARIMA assumes linearity and stationarity (after differencing), which can limit its ability to capture complex, non-linear patterns in volatile stock price data.

Dataset:

SVR is trained on a combined dataset of AAPL stock features (‘open’, ‘high’, ‘low’, ‘close’, ‘volume’) from Yahoo Finance for 03-02-2023 to 03-02-2025.

ii. FinBERT + SVR

Overview

SVR is a ML method based on Support Vector Machines, adapted for regression tasks. In this study, SVR is augmented with FinBERT-derived sentiment scores to forecast AAPL stock prices, harnessing its ability to model non-linear relationships in high-dimensional data. The combination of FinBERT sentiment and stock features aims to capture both quantitative market dynamics and qualitative investor sentiment.

Implementation Details

- **Dataset**:

  SVR is trained on a combined dataset of AAPL stock features (‘open’, ‘high’, ‘low’, ‘close’, ‘volume’) from Yahoo Finance and FinBERT-derived sentiment scores from financial news (sourced via Alpha Vantage API) for 03-02-2023 to 03-02-2025.

- **Preprocessing**:

  - **Stock Features**: Missing values are handled via interpolation, and features are normalized using Min-Max scaling to ensure compatibility with SVR’s sensitivity to feature scales.
  - **Sentiment Features**: News articles are cleaned (removing HTML tags, stop words), tokenized, and processed by FinBERT to generate daily sentiment scores (positive, negative, neutral). These scores are aggregated and aligned with stock data.
  - **Feature Integration**: The dataset combines normalized stock features and sentiment scores into a single input vector for each day.

- **Training**:

  The model is trained on 80% of the data using the Scikit-learn library in Python. The ‘close’ price is the target variable, with input features including stock data and sentiment scores.

iii. FinBERT + RFR

Overview

RFR is an ensemble machine learning method that pools multiple decision trees to increase predictive correctness and reduce overfitting. In this study, RFR is trained on a combined dataset of AAPL stock features and FinBERT-derived sentiment scores, leveraging its robustness in modeling non-linear relationships and high-dimensional data.

Implementation Details

- **Dataset**:

  Identical to SVR, RFR uses a combined dataset of AAPL stock features (‘open’, ‘high’, ‘low’, ‘close’, ‘volume’) and FinBERT-derived sentiment scores for 03-02-2023 to 03-02-2025.

❖ **Preprocessing**:

Same as SVR, with stock features normalized and sentiment scores integrated into a unified input vector.

❖ **Training**:

The model is trained on 80% of the data using Scikit-learn's RandomForestRegressor. The 'close' price is the target, with stock and sentiment features as inputs.

iv. FinBERT + GBR

Overview

GBR is an ensemble machine learning technique that builds sequential decision trees to minimize residual errors, offering high predictive accuracy for regression tasks. In this study, GBR is trained on the combined dataset of AAPL stock features and FinBERT-derived sentiment scores, achieving the best performance among all models.

Implementation Details

❖ **Dataset**:

Same as SVR and RFR, using AAPL stock features and FinBERT-derived sentiment scores for 03-02-2023 to 03-02-2025.

❖ **Preprocessing**:

Identical to SVR and RFR, with normalized stock features and integrated sentiment scores.

❖ **Training**:

The model is trained on 80% of the data using Scikit-learn's Gradient Boosting Regressor. The 'close' price is the target variable.

4.2.4 Model Evaluation

The performance of the four models is evaluated using four regression metrics to assess their predictive accuracy and explanatory power:

- MAE: Calculate absolute difference average between predicted and actual price.
- MSE: Calculates the squared difference average between predicted and actual price.
- RMSE: It provide square root of MSE.
- R^2: It used to evaluate goodness of fit of the model.

These metrics are calculated on the test dataset, and the results are compared to determine the most effective model for sentiment-augmented stock prediction.

4.3 EXPERIMENTAL SETUP

The research was carried out in a Python-driven environment designed to support consistent and scalable outcomes. The hardware system used to meet intensive computational tasks like data preparation, sentiment score generation, model deployment was carried with an Intel core i7 CPU, 16GB RAM, and an NVIDIA GPU. The main programming language employed in this study is Python (version 3.8 or later), in association with specialized libraries like Numpy, Pandas, Scikit Learn etc. Stock data of APPL was retrieved from Yfinance library, while the APl stock related news article was fetched with the help of API provided by Alpha Vantage. Hugging Face transformer framework provide FinBERT which performed sentiment analysis for stock news data. Scikit learn library have been deployed for ML predictive models like SVR, RFR, and GBR, in addition with evaluation metrics. Statsmodels provide implementation of SARIMA time series model, in addition Pandas and Numpy provide effective data processing and handling. Results of visualization for comparing evaluation metric of models was provided by Matplotlib and Seaborn. Dataset tenure for both stock and sentiment have been take from 03-02-2023 to 03-02-2025 giving a comprehensive timeframe for analysis.

Training of all predictive models and their evaluation have been performed in identical conditions under this experiment setup, providing a just comparison ecosystem for sentiment induced stock prediction.

4.4 SUMMARY

This chapter presents the methodology for a comparative study on stock price prediction, assimilating FinBERT-based sentiment analysis with ML models. The proposed work involves collecting Apple stock and financial news data, preprocessing them for compatibility, training SARIMA and three FinBERT-augmented ML models (SVR, RFR, GBR), and evaluating their performance using MAE, MSE, RMSE, and R^2. The experimental setup leverages a Python-based environment with relevant libraries and hardware to ensure efficient computation and reproducibility. By addressing the research gaps in domain-specific sentiment analysis and model comparisons, this methodology provides a robust framework for enhancing stock market forecasting accuracy, with potential implications for financial decision-making.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter presents the experimental results and discussion for a comparative study on stock price prediction, integrating sentiment analysis using FinBERT with machine learning models

5.1 OVERVIEW

The study evaluates four models: SARIMA, FinBERT+SVR, FinBERT+ RFR, and FinBERT+ GBR using Apple Inc. (AAPL) stock data and financial news sentiment from 03-02-2023 to 03-02-2025. The performance of these models is assessed using four regression metrics: MAE, MSE, RMSE, and R^2. The results object to address the research gaps recognised in the literature, specifically the limited use of domain-specific sentiment models and the lack of comprehensive comparisons between sentiment-augmented machine learning models and traditional time-series models. The discussion analyzes the findings in the context of these gaps, the research questions, and their implications for financial forecasting.

5.2 EXPERIMENTAL RESULTS

The performance of the four predictive models—SARIMA, FinBERT+SVR, FinBERT+RFR, and FinBERT+GBR—was evaluated on the test dataset, comprising 20% of the AAPL stock data from 03-02-2023 to 03-02-2025. The models were assessed using MAE, MSE, RMSE, and R^2 to measure predictive accuracy and explanatory power. The results are summarized below in Table 5.1 highlighting the comparative performance across all metrics.

Model Name	MAE	MSE	RMSE	R^2
SARIMA	6.9316	83.0328	9.1122	0.1969
FinBERT+SVR	3.8762	23.5417	4.8520	0.9487
FinBERT+RFR	3.1527	14.5136	3.8097	0.9859
FinBERT+GBR	2.5004	10.4263	3.2290	0.9864

Table 5. 1 Regression Metrics of Predictive Models

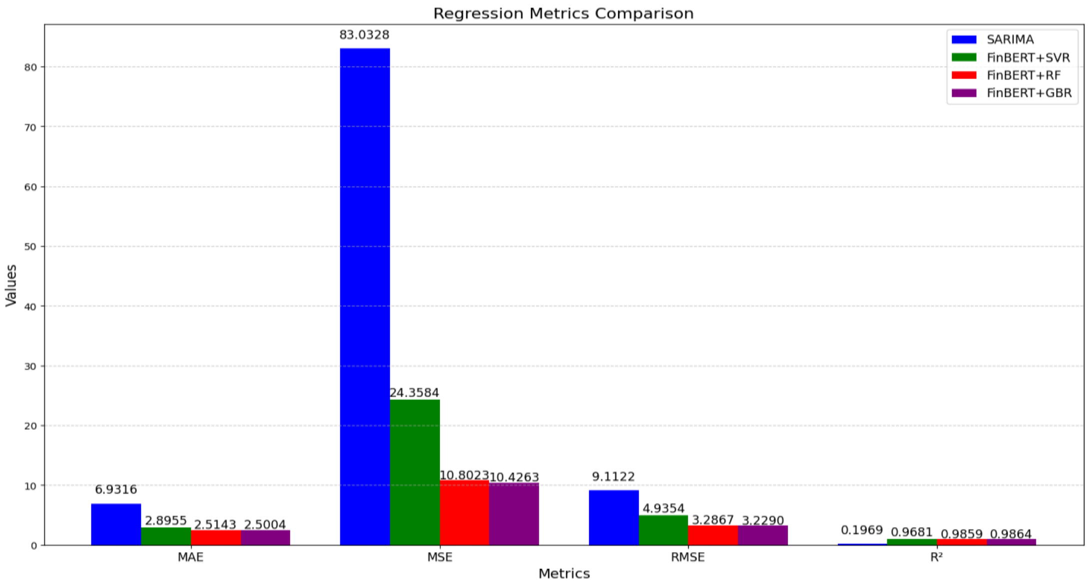


Fig. 5. 1 Bar chart of the performances of various comparative models

Fig. 5.1 illustrate the bar chart of comparative performance, with FinBERT+GBR consistently showing the lowest error metrics and highest R^2 score, followed closely by FinBERT+RFR, then FinBERT+SVR, and significantly outperforming SARIMA

5.2.1 Model 1: SARIMA

The SARIMA model, serving as the baseline, exhibited the poorest performance across all metrics. With an MAE of 6.9316, MSE of 83.0328, and RMSE of 9.1122, it indicates significant prediction errors, reflecting its limited ability to capture the non-linear and volatile patterns in AAPL stock prices. The R^2 score of 0.1969 suggests that only 19.69% of the variance in the stock prices is explained by the model, underscoring its inadequacy for accurate forecasting in this context as depicted in Fig 5.2.

5.2.2 Model 2: FinBERT+SVR

The FinBERT+SVR model, which integrates sentiment scores from financial news with stock features, showed substantial improvement over SARIMA. Reduction in error can be seen in the evaluation metrics of MAE of 3.8762, MSE of 23.5417, and RSME of 4.8520. Significant role of adding FinBERT induced sentiment can be seen through improved R2 score of 0.9487 that reflects significant 94.87% variance in prices of stock as shown in Fig 5.5. But its performance is surpassed by other FinBERT induced ML models.

5.2.3 Model 3: FinBERT+RFR

Regression metrics of FinBERT+RFR have significantly decreased with MAE of 3.1527. MSE of 14.5136, and RMSE of 3.8097. RFR is very robust in handling nonlinear relationships and high dimensional data as it can be seen through R2 of 0.9859 as it explained 98.59% variance of data as shown in Fig 5.4. Incorporation of FinBERT induced sentiments have increased the capability of model to effectively captures market dynamics, making it a strong competitor among other tested models..

5.2.4 Model 4: FinBERT+GBR

Regression metrics of FinBERT+GBR have best performed with lowest value across all metrics with MAE of 3.1527. MSE of 14.5136, and RMSE of 3.8097. GBR have outperformed all other models in handling nonlinear relationships and high dimensional data as it can be seen through R2 of 0.9859 as it explained 98.59% variance of data, explaining highest predictive capability as shown in Fig 5.3. The gradient boosting approach, with incorporation of FinBERT induced sentiments have increased the capability of model to effectively captures market dynamics, making it most competitor among other tested models in the study.

.

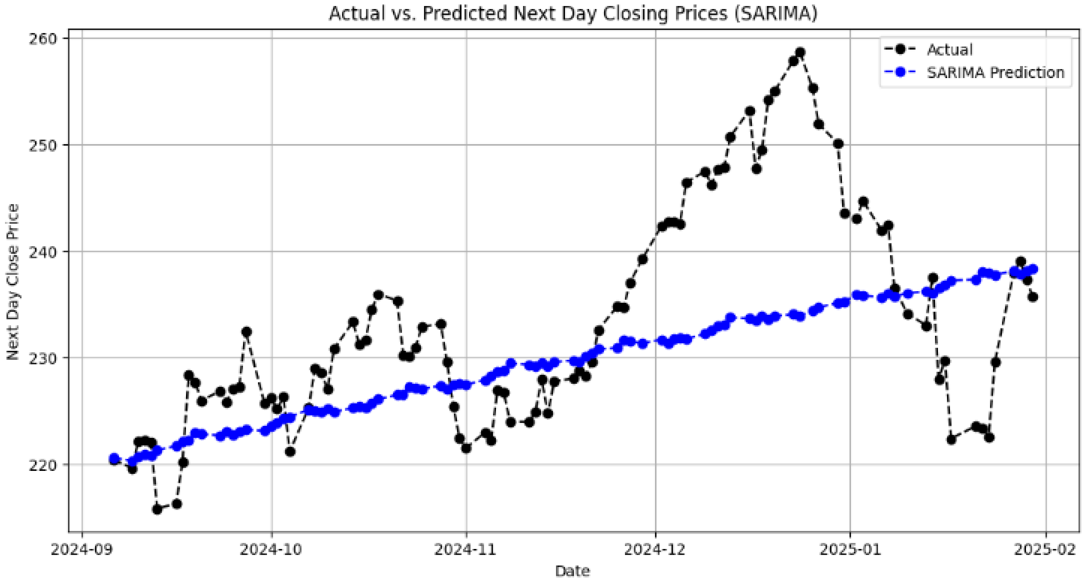


Fig. 5. 2 Linechart of nextday closing price test results(SARIMA)

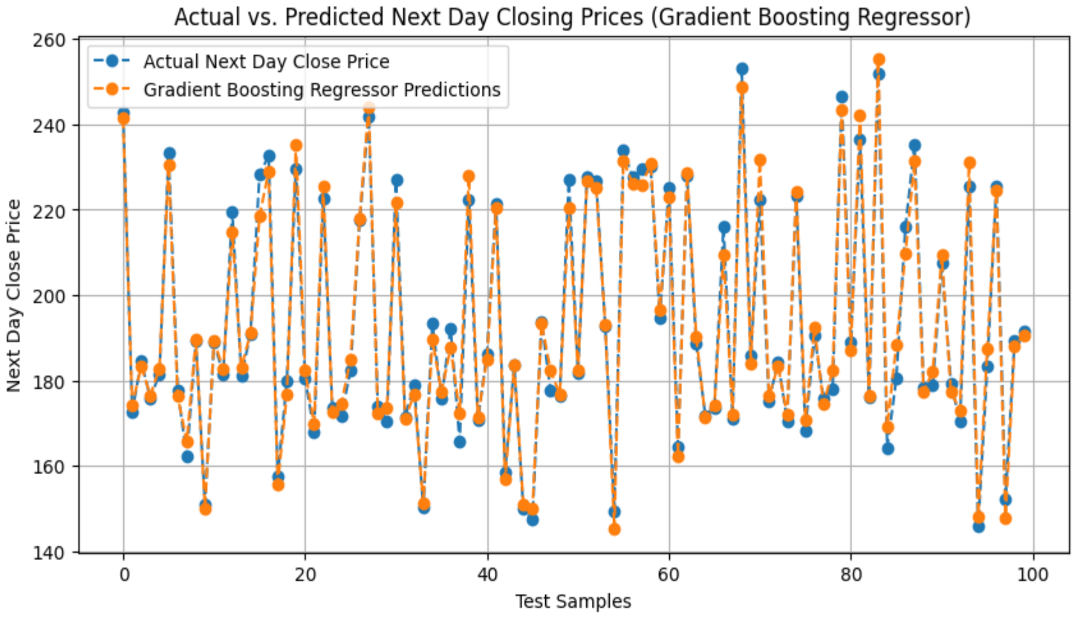


Fig. 5. 3 Linechart of nextday closing price test results(GBR)

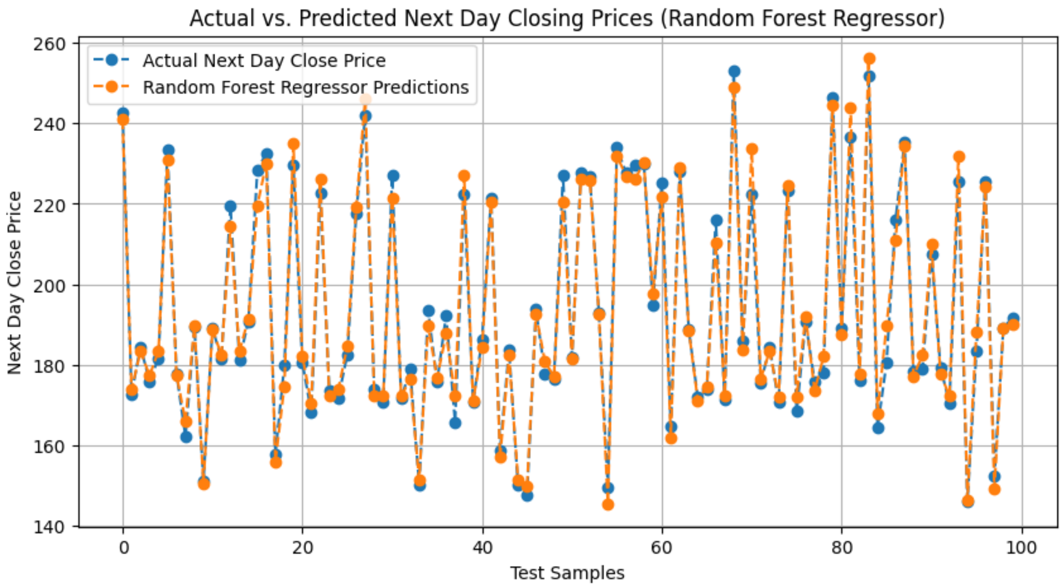


Fig. 5. 4 Linechart of nextday closing price test results(RFR)

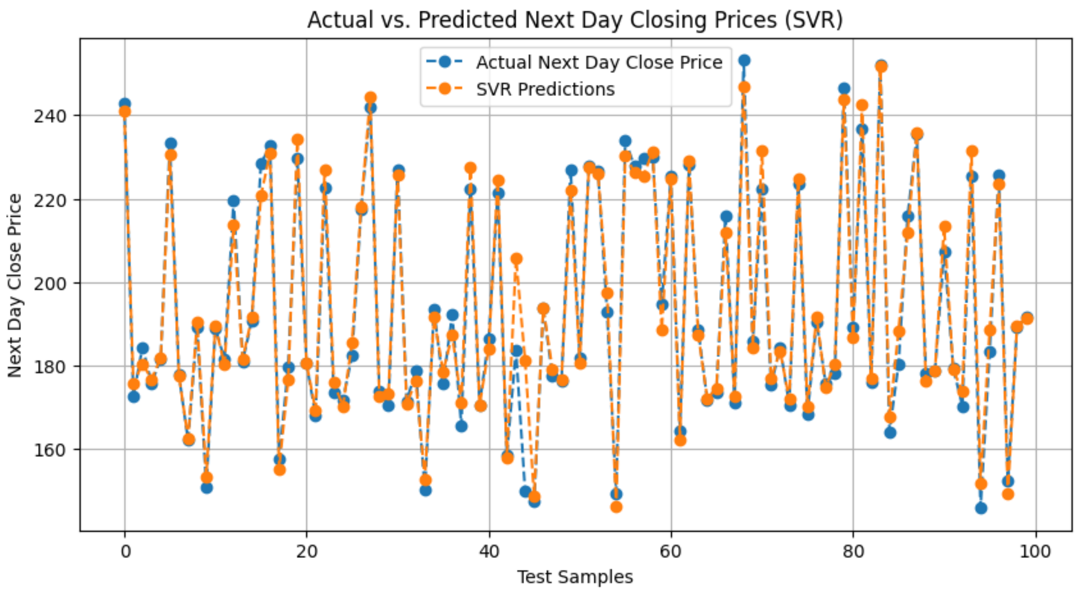


Fig. 5. 5 Linechart of nextday closing price test results(SVR)

5.3 DISCUSSION

The experimental results validate the hypothesis that integrating domain-specific sentiment analysis with machine learning models enhances stock price prediction accuracy, addressing the research gaps and questions outlined in Chapter 3. Below, the findings are discussed in the context of the research objectives and their implications.

Research Question 1: How does the use of a domain-specific sentiment model like FinBERT improve the accuracy of stock price predictions compared to general-purpose sentiment models?

The superior performance of FinBERT+SVR, FinBERT+RFR, and FinBERT+GBR over SARIMA highlights the effectiveness of FinBERT in capturing nuanced financial sentiment. Unlike general-purpose sentiment models (e.g., VADER, TextBlob), which often miss domain-specific terminology, FinBERT's pre-training on financial texts enables it to extract accurate sentiment scores from news articles. The low error metrics (e.g., MAE of 2.5004 for FinBERT+GBR) and high R^2 scores (up to 0.9864) demonstrate that FinBERT-derived sentiment features significantly enhance predictive accuracy, addressing the gap in the limited use of domain-specific sentiment models.

Research Question 2: To what extent do sentiment-augmented ML models (SVR, RFR, GBR) outperform a traditional time-series model (SARIMA) in stock price forecasting?

The results clearly show that FinBERT-augmented ML models outperform SARIMA across all metrics. SARIMA's high error rates (MAE: 6.9316, RMSE: 9.1122) and low R^2 (0.1969) reflect its inability to handle the non-linear and volatile nature of stock prices, consistent with literature findings (e.g., Zhang et al., 2019). In contrast, FinBERT+GBR's near-perfect R^2 (0.9864) and low errors underscore the advantage of ML models in leveraging sentiment and stock features, addressing the gap in comprehensive model comparisons.

Research Question 3: How can the integration of FinBERT-derived sentiment features with conventional stock features enhance the predictive power of ML models?

The integration of sentiment scores with stock features ('open', 'high', 'low', 'close', 'volume') enabled the ML models to capture both quantitative market dynamics and qualitative investor sentiment. The feature engineering pipeline, which aligned daily sentiment scores with stock data, allowed models like GBR to model complex interactions, resulting in minimal errors (e.g., MSE: 10.4263 for FinBERT+GBR). This confirms the importance of sentiment-driven feature optimization, as highlighted in the literature (e.g., Yang et al., 2023).

Research Question 4: What are the practical implications of using sentiment-driven ML models for financial forecasting, particularly in terms of model interpretability and adoption in the financial industry?

The high R^2 scores and low errors of FinBERT+GBR and FinBERT+RFR suggest strong potential for practical adoption in financial forecasting. However, GBR's superior performance comes with increased computational complexity, which may pose challenges for real-time applications. Interpretability remains a concern, as ensemble models like GBR and RFR are less transparent than SARIMA. Future work could incorporate explainable AI techniques (e.g., SHAP values) to enhance trust and regulatory compliance, facilitating adoption in the financial industry.

5.4 SUMMARY

This chapter presented the results and discussion of a comparative study on stock price prediction using SARIMA and three FinBERT-augmented ML models (SVR, RFR, GBR). The experimental results demonstrated that FinBERT+GBR achieved the best performance (MAE: 2.5004, MSE: 10.4263, RMSE: 3.2290, R^2: 0.9864), significantly outperforming SARIMA (MAE: 6.9316, R^2: 0.1969) and other ML models. The discussion highlighted the effectiveness of FinBERT in capturing financial sentiment, the superiority of ML models over traditional time-series approaches, and the importance of sentiment-driven feature integration. These findings address the research gaps in domain-specific sentiment analysis and model comparisons, offering valuable insights for financial forecasting. The results underscore the potential of sentiment-augmented ML models for practical applications, though challenges in interpretability and computational efficiency warrant further exploration.

CHAPTER 6

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

6.1 CONCLUSION

This study focuses on the effectiveness on sentiment integration with machine learning for stock price prediction by investigating comparative analysis of traditional time series model SARIMA with three FinBERT induced ML models i.e. SVR, RFR, GBR. This comparative study applied stock data of Apple Inc. stock (APPL) and sentiment data from 03-02-2023 to 03-02-2025 while evaluating the model performance through MAE, MSE, RSME, and R^2 metrics. This comparative study shows that FinBERT induced models have significantly beat SARIMA, moreover FinBERT+GBR gains the best performance amongst all FinBERT based models. The results of study proof the effectiveness of FinBERT for capturing financial domain specific investors sentiment and capability of ML models in tackling complex intricate patterns.

6.2 FUTURE SCOPE

This study will surely open several path for future advance research that want to explore sentiment based stock prediction. Firstly addition of multiple data sources can enhance sentiment feature set like social media platform (twitter, stocktwits), macroeconomic indicator (inflation, GDP, interest rates), and multimodal data (audio from earning calls. Market analysis video). Secondly, make use of advance deep learning models like LSTM, Transformers induced models, or hybrid models like CNN+LSTM which are capable of obtaining complex pattern and temporal dependencies in financial forecasting of stock data. Thirdly proposed models can be optimized using hyper parameters which can be done through automated techniques like grid search, Bayesian optimization, and genetic algorithm, which ultimately results in reducing computational overhead. Fourthly, proposed models can take input from diverse range of stocks across various sectors and different market conditions like emerging markets, recessionary market etc. Fifthly, proposed model should work on interpretability of model by incorporating explainable AI techniques like SHAP, LIME, and PDP etc. Lastly for fast processing of real-time sentiment analysis, low latency pipelines should be made to encompass the practical high frequency trading utility scenario.

6.3 SOCIAL IMPACT

All the sentiment induced stock prediction model have prominent social implications, especially in taking educated decision making, providing economic stability. These comparative models provide strength to financial stakeholders by utilizing the power of FinBERT and ML models to achieve higher prediction results; traditionally these advance analytical processing done through rigorous statistical calculations. It will help to mitigate financial equality through democratization and wealth creation by enabling border participation in stock market investments. Moreover, sentiment infused ML models enhances the financial forecasting of stocks by penalising speculative trading disseminated misinformation or emotional outrage as model captures intricate complexities. Transparency of market analysis and harvesting trust among investors can be achieved by making financial news data available for public. Additionally caution for over-reliance on automated predictive models have ethical considerations which should be addressed properly to not amplify market instability. Every necessary regulatory framework should be be complied in a responsible manner for model interpretability and deployment to harness the positive social impact while reducing the risks. Eventually this study helps to build a more aware and equitable financial ecosystem, which in long run benefit society by encouraging sound investment culture and economic resilience.

REFERENCES

[1] N. Rahul, S. Sarangi, P. Kedia, and N. Monika, “Analysis of various approaches for stock market prediction,” Journal of Statistics and Management Systems, vol. 23, no. 2, pp. 285–293, Feb. 2020.

[2] S. Kalra and J. S. Prasad, “Efficacy of news sentiment for stock market prediction,” 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Feb. 2019.

[3] Wen jun Gu, Yi hao Zhong, Shi zun Li, Chang song Wei, Li ting Dong, Zhuo yue Wang, and Chao Yan. “Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis”. 8th International Conference on Cloud and Big Data Computing (ICCBDC '24), 67–72. Nov. 2024.

[4] N. J. Li, N. H. Bu, and N. J. Wu, “Sentiment-aware stock market prediction: A deep learning method,” International Conference on Service Systems and Service Management, pp. 1–6, Jun. 2017.

[5] X. Guo and J. Li, "A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 472-477.

[6] N. Jing, Z. Wu, and H. Wang, “A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction,” Expert Systems with Applications, vol. 178, p. 115019, Apr. 2021.

[7] D. Shah, H. Isah, and F. Zulkernine, “Predicting the effects of news sentiments on the stock market,” 2021 IEEE International Conference on Big Data (Big Data), pp. 4705–4708, Dec. 2018.

[8] J. Huang and J. Liu, “Using social media mining technology to improve stock price forecast accuracy,” Journal of Forecasting, vol. 39, no. 1, pp. 104–116, Jun. 2019.

[9] A. Kanavos, G. Vonitsanos, A. Mohasseb and P. Mylonas, "An Entropy-based Evaluation for Sentiment Analysis of Stock Market Prices using Twitter Data," 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA, Zakynthos, Greece, 2020, pp. 1-7

[10] M. Kesavan, J. Karthiraman, R. T. Ebenezer, and S. Adhithyan, “Stock Market Prediction with Historical Time Series Data and Sentimental Analysis of Social Media Data,” 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 477–482, May 2020.

[11] X. Li, P. Wu, and W. Wang, “Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong,” Information Processing & Management, vol. 57, no. 5, p. 102212, Feb. 2020.

[12] I. K. Nti, A. F. Adekoya, and B. A. Weyori, “Predicting stock market price movement using sentiment analysis: Evidence from Ghana,” Applied Computer Systems, vol. 25, no. 1, pp. 33–42, May 2020.

[13] Y. Huang, and T.-T. Ho “Stock price movement prediction using sentiment analysis and CandleStick chart representation,” Sensors, vol. 21, no. 23, p. 7957, Nov. 2021.

[14] W.-J. Liu, Y.-B. Ge, and Y.-C. Gu, “News-driven stock market index prediction based on trellis network and sentiment attention mechanism,” Expert Systems with Applications, vol. 250, p. 123966, Apr. 2024.

[15] N. Das, B. Sadhukhan, T. Chatterjee, and S. Chakrabarti, “Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak,” Social Network Analysis and Mining, vol. 12, no. 1, Jul. 2022.

[16] R. Chiong, Z. Fan, Z. Hu, and S. Dhakal, “A novel ensemble learning approach for stock market prediction based on sentiment analysis and the sliding window method,” IEEE Transactions on Computational Social Systems, vol. 10, no. 5, pp. 2613–2623, Aug. 2022.

[17] S. Kalra and J. S. Prasad, “Efficacy of news sentiment for stock market prediction,” 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Feb. 2019.

[18] M. V. D. H. P. Malawana and R. M. K. T. Rathnayaka, "The Public Sentiment analysis within Big data Distributed system for Stock market prediction– A case study on Colombo Stock Exchange," 2020 5th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2020, pp. 1-6.

[19] J.-X. Liu, J.-S. Leu, and S. Holst, “Stock price movement prediction based on Stocktwits investor sentiment using FinBERT and ensemble SVM,” PeerJ Computer Science, vol. 9, p. e1403, Jun. 2023.

[20] B. A. Abdelfattah, S. M. Darwish, and S. M. Elkaffas, “Enhancing the prediction of stock market movement using Neutrosophic-Logic-Based sentiment analysis,” Journal of Theoretical and Applied Electronic Commerce Research, vol. 19, no. 1, pp. 116–134, Jan. 2024.

[21] R. Batra and S. M. Daudpota, “Integrating StockTwits with sentiment analysis for better prediction of stock price movement,” 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Mar. 2018.

[22] P. Koukaras, C. Nousi, and C. Tjortjis, “Stock market prediction using microblogging sentiment analysis and machine learning,” Telecom, vol. 3, no. 2, pp. 358–378, May 2022.

[23] Y. Guo, “Stock Price Prediction Based on LSTM Neural Network: The Effectiveness of News Sentiment Analysis,” 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME), pp. 1018–1024, Nov. 2020.

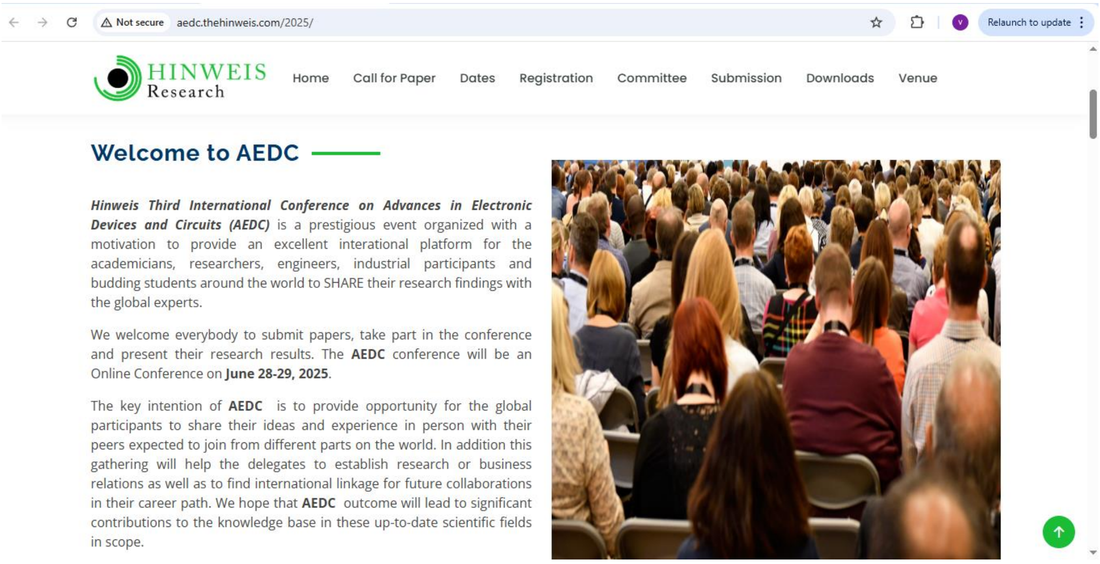
[24] C. Nousi and C. Tjortjis, "A Methodology for Stock Movement Prediction Using Sentiment Analysis on Twitter and StockTwits Data," 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Preveza, Greece, 2021, pp. 1-7.

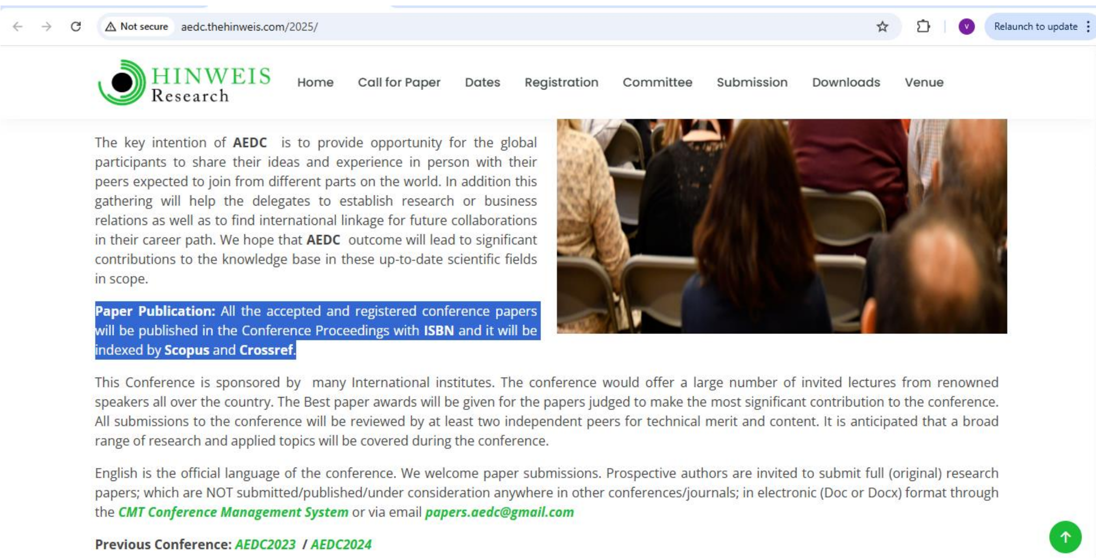
[25] G. Mu, N. Gao, Y. Wang, and L. Dai, “A stock price prediction model based on investor sentiment and optimized deep learning,” IEEE Access, vol. 11, pp. 51353–51367, Jan. 2023.

[26] Z. H. Kilimci and R. Duvar, “An efficient word embedding and deep learning based model to forecast the direction of stock exchange market using Twitter and financial news sites: a case of Istanbul Stock Exchange (BIST 100),” IEEE Access, vol. 8, pp. 188186–188198, Jan. 2020.

LIST OF PUBLICATIONS

[1] Vibhor Singh, Rahul, " A Review Of Sentiment-Driven Intelligent Systems For Analyzing The Influence Of News Media On Stock Market Prediction". The paper has been Accepted at the Hinweis Third International Conference on Advances in Electronic Devices and Circuits (AEDC-2025), 28-29 June 2025. Indexed by Scopus. Paper Id: AEDC-2025_35.





Gmail

vibhor singh <vibhor.official0219@gmail.com>

AEDC2025 :: Acceptance Confirmation and Registration Details

1 message

AEDC Conference <papers.aedc@gmail.com> Wed, May 28, 2025 at 7:41 PM
To: vibhor.official0219@gmail.com, rahul@dtu.ac.in

Dear AEDC-2025 Author,

Warm greetings from Hinweis Research!

We are thrilled to inform you that your submitted paper for **Hinweis Third International Conference on Advances in Electronic Devices and Circuits (AEDC)** has been accepted. Congratulations on this significant achievement! Your dedication to your research is highly commendable.

http://aedc.thehinweis.com/2025

Here are the important details regarding your acceptance:

Review Result and Acceptance Certificate:

The consolidated review result is attached along with this email which includes the paper ID. Kindly use the paper ID for any further communication. The review result itself is the acceptance certificate.

Publication and Indexing:

All accepted conference papers will be published in the **Conference Proceedings** with an **ISBN Number** and will be indexed by **Scopus and Crossref**, further enhancing the visibility of your work.

In all your future communications with us, please be sure to quote your **Paper ID and Category** for efficient processing.

Should you have any questions or require further assistance, please do not hesitate to reach out to us. We are here to support you throughout the conference preparation process.

Once again, congratulations on your paper's acceptance! We look forward to your valuable contribution at **AEDC-2025**.

--
Have more questions? Check our FAQs, then come to us! Click Here

--

AEDC Conference
Hinweis Research

+91 9074 609 778 thehinweis.com

2 attachments

AEDC-2025_35.pdf
355K

AEDC-2025_36.pdf
354K

AEDC Conference Jun 17, 2025, 12:05 PM (4 days ago)

to me, rahul@dtu.ac.in

Dear Author,

We are pleased to confirm the receipt of your registration form, payment proof and the Camera Ready Paper for the AEDC2025 International Conference. Any queries or clarifications needed will be promptly addressed by the respective department. Your active participation and dedication are greatly appreciated and we extend our sincere gratitude for your valuable contributions to this event, thank you...

--

Have more questions? Check our FAQs, then come to us! Click Here

--

AEDC Conference
Hinweis Research

+91 9074 609 778 thehinweis.com

[2] Vibhor Singh, Rahul, "A Comparative Study of Sentiment-Driven Intelligent Systems Using FinBERT". The paper has been Accepted at the Hinweis Third International Conference on Advances in Electronic Devices and Circuits (AEDC-2025), 28-29 June 2025. Indexed by Scopus. Paper Id: AEDC-2025_36.

Not secure aedc.thehinweis.com/2025/ Relaunch to update

HINWEIS Research Home Call for Paper Dates Registration Committee Submission Downloads Venue

Welcome to AEDC

Hinweis Third International Conference on Advances in Electronic Devices and Circuits (AEDC) is a prestigious event organized with a motivation to provide an excellent international platform for the academicians, researchers, engineers, industrial participants and budding students around the world to SHARE their research findings with the global experts.

We welcome everybody to submit papers, take part in the conference and present their research results. The **AEDC** conference will be an Online Conference on **June 28-29, 2025**.

The key intention of **AEDC** is to provide opportunity for the global participants to share their ideas and experience in person with their peers expected to join from different parts on the world. In addition this gathering will help the delegates to establish research or business relations as well as to find international linkage for future collaborations in their career path. We hope that **AEDC** outcome will lead to significant contributions to the knowledge base in these up-to-date scientific fields in scope.

Not secure aedc.thehinweis.com/2025/

Relaunch to update

HINWEIS Research

Home Call for Paper Dates Registration Committee Submission Downloads Venue

The key intention of **AEDC** is to provide opportunity for the global participants to share their ideas and experience in person with their peers expected to join from different parts on the world. In addition this gathering will help the delegates to establish research or business relations as well as to find international linkage for future collaborations in their career path. We hope that **AEDC** outcome will lead to significant contributions to the knowledge base in these up-to-date scientific fields in scope.

Paper Publication: All the accepted and registered conference papers will be published in the Conference Proceedings with **ISBN** and it will be indexed by **Scopus** and **Crossref**.

This Conference is sponsored by many International institutes. The conference would offer a large number of invited lectures from renowned speakers all over the country. The Best paper awards will be given for the papers judged to make the most significant contribution to the conference. All submissions to the conference will be reviewed by at least two independent peers for technical merit and content. It is anticipated that a broad range of research and applied topics will be covered during the conference.

English is the official language of the conference. We welcome paper submissions. Prospective authors are invited to submit full (original) research papers; which are NOT submitted/published/under consideration anywhere in other conferences/journals; in electronic (Doc or Docx) format through the *CMT Conference Management System* or via email *papers.aedc@gmail.com*

6/11/25, 12:32 PM Gmail - AEDC2025 :: Acceptance Confirmation and Registration Details

Gmail **vibhor singh <vibhor.official0219@gmail.com>**

AEDC2025 :: Acceptance Confirmation and Registration Details

1 message

AEDC Conference <papers.aedc@gmail.com> Wed, May 28, 2025 at 7:41 PM
To: vibhor.official0219@gmail.com, rahul@dtu.ac.in

Dear AEDC-2025 Author,

Warm greetings from Hinweis Research!

We are thrilled to inform you that your submitted paper for **Hinweis Third International Conference on Advances in Electronic Devices and Circuits (AEDC)** has been accepted. Congratulations on this significant achievement! Your dedication to your research is highly commendable.

http://aedc.thehinweis.com/2025

Here are the important details regarding your acceptance:

Review Result and Acceptance Certificate:

The consolidated review result is attached along with this email which includes the paper ID. Kindly use the paper ID for any further communication. The review result itself is the acceptance certificate.

Publication and Indexing:

All accepted conference papers will be published in the **Conference Proceedings** with an **ISBN Number** and will be indexed by **Scopus and Crossref**, further enhancing the visibility of your work.

In all your future communications with us, please be sure to quote your **Paper ID and Category** for efficient processing.

Should you have any questions or require further assistance, please do not hesitate to reach out to us. We are here to support you throughout the conference preparation process.

Once again, congratulations on your paper's acceptance! We look forward to your valuable contribution at **AEDC-2025**.

--

Have more questions? Check our FAQs, then come to us! Click Here

--

AEDC Conference
Hinweis Research

+91 9074 609 778 thehinweis.com

2 attachments

AEDC-2025_35.pdf
355K

AEDC-2025_36.pdf
354K

AEDC Conference
to me, rahul@dtu.ac.in

Jun 17, 2025, 12:05 PM (4 days ago)

Dear Author,

We are pleased to confirm the receipt of your registration form, payment proof and the Camera Ready Paper for the AEDC2025 International Conference. Any queries or clarifications needed will be promptly addressed by the respective department. Your active participation and dedication are greatly appreciated and we extend our sincere gratitude for your valuable contributions to this event, thank you...

--

Have more questions? Check our FAQs, then come to us! Click Here

--

AEDC Conference
Hinweis Research

+91 9074 609 778 thehinweis.com

ANNEXURE-IV

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis A comparative study on machine learning based Stock prediction by incorporating sentiment analysis using finBERT

Total Pages 34 Name of the Scholar VIBHOR SINGH

Supervisor (s)

(1) Dr. Rahul

(2)

(3)

Department of Software Engineering

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 8%, Total Word Count: 8348

Date: 20-05-2025

Candidate's Signature

Signature of Supervisor(s)

Vib_only_thesis16_05_25_17_46.docx

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:96254492

Submission Date

May 16, 2025, 6:39 PM GMT+5:30

Download Date

May 16, 2025, 6:42 PM GMT+5:30

File Name

Vib_only_thesis16_05_25_17_46.docx

File Size

766.8 KB

34 Pages

8,348 Words

50,357 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 8 words)

Match Groups

63 Not Cited or Quoted 8%
Matches with neither in-text citation nor quotation marks

1 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

4%	Internet sources
2%	Publications
6%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Vib_only_thesis16_05_25_17_46.docx

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:96254492

Submission Date

May 16, 2025, 6:39 PM GMT+5:30

Download Date

May 16, 2025, 6:43 PM GMT+5:30

File Name

Vib_only_thesis16_05_25_17_46.docx

File Size

766.8 KB

34 Pages

8,348 Words

50,357 Characters

30/5/25

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

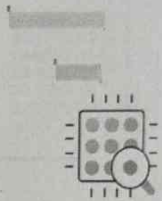
Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.



False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

DECLARATION

We/I hereby certify that the work which is presented in the Major Project-II/Research Work entitled A comparative study on ML based stock prediction by incorporating sentiment analysis in partial fulfilment of the requirement for the award of the Degree of Bachelor/Master of Technology in Mtech and submitted to the Department of SE, Delhi Technological University, Delhi is an authentic record of my/our own, carried out during a period from 2023-25, under the supervision of Dr Rahul.

The matter presented in this report/thesis has not been submitted by us/me for the award of any other degree of this or any other Institute/University. The work has been published/accepted/communicated in SCI/ SCI expanded/SSCI/Scopus indexed journal OR peer reviewed Scopus indexed conference with the following details:

Title of the Paper: A Review of Sentiment-Driven Intelligent Systems for analyzing the influence of News media on stock market Prediction
Author names (in sequence as per research paper): Vibhor Singh, Dr. Rahul
Name of Conference/Journal: AEDC - 2025
Conference Dates with venue (if applicable): 28-29 June 2025, virtual mode
Have you registered for the conference (Yes/No)?: Yes
Status of paper (Accepted/Published/Communicated): Accepted
Date of paper communication: 18/5/2025
Date of paper acceptance: 28/5/2025
Date of paper publication: yet to be done

Student(s) Roll No., Name and Signature
VIBHOR SINGH
23/DSC/01

SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I, further certify that the publication and indexing information given by the students is correct.

Place: New Delhi

Date: 30/5/2025

Supervisor Name and Signature Dr. Rahul 30/5/25

NOTE: PLEASE ENCLOSE RESEARCH PAPER ACCEPTANCE/ PUBLICATION/COMMUNICATION PROOF ALONG WITH SCOPUS INDEXING PROOF (Conference Website OR Science Direct in case of Journal Publication).

DECLARATION

We/I hereby certify that the work which is presented in the Major Project-II/Research Work entitled A Comparative study on ML based stock prediction by incorporating Sentiment analysis FinBERT in fulfilment of the requirement for the award of the Degree of Bachelor/Master of Technology in M tech and submitted to the Department of SE, Delhi Technological University, Delhi is an authentic record of my/our own, carried out during a period from 2023-25, under the supervision of Dr. Rahul.

The matter presented in this report/thesis has not been submitted by us/me for the award of any other degree of this or any other Institute/University. The work has been published/accepted/communicated in SCI/ SCI expanded/SSCI/Scopus indexed journal OR peer reviewed Scopus indexed conference with the following details:

Title of the Paper: A Comparative Study of Sentiment-Driven Intelligent System Using FinBERT
Author names (in sequence as per research paper): Vibhor Singh, Dr. Rahul
Name of Conference/Journal: AEDC - 2025
Conference Dates with venue (if applicable): 28-29 June 2025, Virtual mode
Have you registered for the conference (Yes/No)?: Yes
Status of paper (Accepted/Published/Communicated): Accepted
Date of paper communication: 18/5/2025
Date of paper acceptance: 28/5/2025
Date of paper publication: yet to be done

Singh

Student(s) Roll No., Name and Signature

VIBHOR SINGH
23/DSC/01

SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I, further certify that the publication and indexing information given by the students is correct.

Place: New Delhi

Date: 30/05/2025

Supervisor Name and Signature Dr. Rahul 30/5/25

NOTE: PLEASE ENCLOSE RESEARCH PAPER ACCEPTANCE/ PUBLICATION/COMMUNICATION PROOF ALONG WITH SCOPUS INDEXING PROOF (Conference Website OR Science Direct in case of Journal Publication).