Transforming Financial News into Market Strategies: A Hybrid Stock Prediction Model

Project report submitted to
Indian Institute of Information Technology, Nagpur, in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology
In
Computer Science and Engineering

by

Sairatna Kamble (BT21CSE081)
Avinash Wagh (BT21CSE108)
Sohum Hedaoo (BT21CSE115)

Under the guidance of

Dr. Jagdish Chakole

Indian Institute of Information Technology,

***Nagpur 441108* (India)**

2021-2025

Transforming Financial News into Market Strategies: A Hybrid Stock Prediction Model

Project report submitted to
Indian Institute of Information Technology, Nagpur, in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology
In
Computer Science and Engineering

by

Sairatna Kamble (BT21CSE081)
Avinash Wagh (BT21CSE108)
Sohum Hedaoo (BT21CSE115)

Under the guidance of

Dr. Jagdish Chakole

Indian Institute of Information Technology,

***Nagpur 441108* (India)**

2021-2025

Department of Computer Science and Engineering
Indian Institute of Information Technology, Nagpur

Declaration

We, Sairatna Kamble (BT21CSE081), Avinash Wagh (BT21CSE108), Sohum Hedaoo (BT21CSE115), hereby declare that this project work titled "Transforming Financial News into Market Strategies: A Hybrid Stock Prediction Model" is carried out by us in the Department of Computer Science and Engineering of Indian Institute of Information Technology, Nagpur. The work is original and has not been submitted earlier whole or in part for the award of any degree/diploma at this or any other Institution /University.

Date:

Sr.No.	Name	Enrollment No.	Signature
1	Sairatna Kamble	BT21CSE081	
2	Avinash Wagh	BT21CSE108	
3	Sohum Hedaoo	BT21CSE115	

Declaration

We, Sairatna Kamble, Avinash Wagh, Sohum Hedaoo, Enrollment No (BT21CSE081, BT21CSE108, BT21CSE115), understand that plagiarism is defined as any one or the combination of the following:

1. Uncredited verbatim copying of individual sentences, paragraphs or illustrations (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.

2. Uncredited improper paraphrasing of pages or paragraphs (changing a few words or phrases, or rearranging the original sentence order).

3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who did or wrote what. (Source: IEEE, the institute, Dec.2004) I have made sure that all the ideas, expressions, graphs, diagrams, etc. that are not a result of my own work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.

I affirm that no portion of my work can be considered as plagiarism and I take full responsibility if such complaint occurs. I understand fully well the guide of the thesis may not be in a position to check for possibility of such incidences of plagiarism in this body of work.

Date:

Sr.No.	Name	Enrollment No.	Signature
1	Sairatna Kamble	BT21CSE081	
2	Avinash Wagh	BT21CSE108	
3	Sohum Hedaoo	BT21CSE115	

Computer Science and Engineering,

IIIT, NAGPUR

<u>Certificate</u>

This is to certify that the project titled "Transforming Financial News into Market Strategies: A Hybrid Stock Prediction Model", submitted by **Sairatna Kamble (BT21CSE081), Avinash Wagh (BT21CSE108), Sohum Hedaoo (BT21CSE115**) in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering,** IIIT Nagpur. The work is comprehensive, complete and fit for final evaluation.

Date:

Dr. Jagdish Chakole
Assistant Professor, CSE
IIIT, Nagpur

Dr. Nishat Afshan Ansari
Head of the Department, CSE
IIIT, Nagpur

Acknowledgement

"Acknowledgement is an art, one can write glib stanzas without meaning a word, and on the other hand one can make a simple expression of gratitude"

It gives us a great sense of pleasure to present the report of the Project Work undertaken during Final Year Project. We owe a special debt of gratitude to our Project Mentor Dr. Jagdish Chakole, Department of Computer Science and Engineering, Indian Institute of Information Technology, Nagpur for his constant support and guidance throughout the course of our work. It is only his cognizant efforts that our endeavors have seen the light of the day. No amount of written expression is sufficient to show our deepest scene of gratitude to him.

We would also like to express our gratitude to Dr. Nishat Afshan Ansari, HOD, Dr. Tausif Diwan, Associate Dean, and Dr. Prem Lal Patel, Director, for their invaluable support throughout the course of this project. We are very thankful to the project review committee for their ever-lasting support and guidance on the ground in which we have acquired a new field of knowledge. We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project.

We also acknowledge with a deep sense of reverence, our gratitude towards our parents and members of our family, who have always supported us morally as well as economically. And last but not least, gratitude goes to all of our friends who directly and indirectly helped us to complete this project and this project report.

Sr.No	**Name**	**Enrollment No.**	**Signature**
1	**Sairatna Kamble**	**BT21CSE081**	
2	**Avinash Wagh**	**BT21CSE108**	
3	**Sohum Hedaoo**	**BT21CSE115**	

ABSTRACT

Stock market prediction is a challenging yet critical task due to the complex and volatile nature of financial markets, which are influenced by numerous factors such as market sentiment, company performance, government policies, and external economic events. Traditional stock price prediction methods often fail to account for the intricacies of market movements and investor behaviour. As financial markets become more data-driven, there is an increasing need for more sophisticated models that integrate quantitative and qualitative data sources. This project investigates the integration of financial news sentiment analysis with stock market data to improve the accuracy of stock price predictions. The problem addressed in this research is the inability of traditional models to effectively capture the dynamic interplay between market sentiment, news, and stock price movements. To solve this, we propose a hybrid methodology that utilizes Fin-BERT, a transformer-based model fine-tuned for financial sentiment analysis, to process both textual and numerical stock market data the textual data consists of financial news articles, which are analysed to determine sentiment and generate trading signals, while the numerical data includes historical stock prices and technical indicators, also processed by Fin-BERT to predict trading signals. These signals are then fused to provide a final prediction on whether to BUY, SELL, or HOLD a stock. The research methodology integrates the strengths of both sentiment analysis and stock market trading signal forecasting, offering a more comprehensive and robust approach to stock market prediction. Experimental results demonstrate that this hybrid model significantly improves prediction accuracy compared to traditional methods, making it a valuable tool for investors seeking to make informed trading decisions the implications of this discovery extend to more reliable investment strategies, highlighting the potential of combining ML with financial news sentiment analysis for more effective stock market forecasting.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

Abbreviation	Meaning
ML	Machine Learning
NLP	Natural Language Processing
Fin-BERT	Financial Bert
BERT	Bidirectional Encoder Representation from Transformer

CHAPTER 1
INTRODUCTION

The stock market operates as a complex and dynamic system influenced by an interplay of financial, economic, and psychological factors. Investor sentiment, shaped significantly by financial news and global events, plays a critical role in driving market behaviour. Predicting stock trading signals, such as BUY, SELL, or HOLD, is a daunting challenge due to the non-linear, noisy, and volatile nature of market data. Traditional predictive models often focus solely on numerical data, such as historical prices and technical indicators, neglecting the critical role that financial news sentiment plays in influencing market movements. This gap in understanding results in less accurate predictions and suboptimal decision-making for investors.

In recent years, advancements in natural language processing (NLP) and machine learning have enabled the development of models capable of analysing textual data, including news articles, to extract sentiment and correlate it with market trends. This project leverages Fin-BERT, a transformer-based language model optimized for financial text analysis, to evaluate the sentiment of news articles and classify them into positive, negative, or neutral categories. These classifications are further mapped to actionable trading signals.

To complement this analysis, numerical stock data—encompassing metrics like open and close prices, adjusted close, volume, and various technical indicators—is processed through machine learning models. By combining the insights from sentiment analysis and numerical forecasting, the system generates final trading signals based on consensus. This integrated approach captures the nuanced relationship between market sentiment and stock trends, enabling more reliable predictions.

An additional innovation of this project is its ability to estimate the potential profit or loss associated with each trading signal. This feature provides investors with a clearer picture of the financial impact of their decisions, enhancing the system's practical utility. By presenting this information in an accessible manner, the system bridges the gap between complex market analytics and actionable investment strategies, catering to both novice and experienced investors.

1.1 Objectives:

1. To develop a hybrid model combining sentiment analysis and numerical stock data for predicting trading signals.
2. To utilize Fin-BERT for analyzing financial news and correlating it with stock trends.
3. To provide actionable trading signals (BUY, SELL, HOLD) with potential profit or loss estimates.
4. To improve the accuracy and reliability of trading decisions by integrating textual and numerical data.

1.2 Applications:

1. Investor Decision-Making: Offers informed trading signals to optimize investment strategies.
2. Risk Management: Provides profit/loss estimates to help investors assess risks.
3. Market Analysis: Enhances understanding of the interplay between news sentiment and stock trends.
4. Financial Advisory Tools: Can be integrated into platforms to support financial advisors and traders.

1.3 Organization of the Thesis:

Chapter 1: Introduction

This section outlines the research problem, objectives, and the significance of the study. It provides an overview of the challenges in stock price prediction and the potential benefits of integrating financial news sentiment analysis with advanced machine learning techniques.

Chapter 2: Literature Review

This discusses previous studies and methodologies related to stock price prediction, sentiment analysis, and the application of machine learning and deep learning models in financial forecasting. This section highlights the gaps in existing research that the proposed work aims to address.

Chapter 3: Work-Done

This section is divided into four subsections. The Methodology subsection describes the step-by-step approach employed in the research, detailing the integration of sentiment analysis and numerical data processing. The Datasets subsection provides insights into the data used, including the NIFTY and Reliance datasets, and explains the preprocessing techniques applied. The Architecture of the Proposed Model subsection elaborates on the Fin-BERT model for sentiment analysis and the Transformer-based trading signal forecasting model. Additionally, the Web Application Development subsection highlights the development of a user-friendly interface to implement the proposed system for real-time usage.

Chapter 4: Result and Discussions

This section presents the performance metrics, comparative analysis, and findings of the research. It interprets the outcomes in the context of their practical implications for investment strategies.

Chapter 5: Summary and Conclusion

It summarizes the key contributions of the research, discusses its limitations, and suggests directions for future work, such as enhancing model robustness and expanding its applicability to other financial domains.

Finally, the References section lists all the scholarly works and resources cited throughout the thesis.

CHAPTER 2
LITERATURE REVIEW

1. Convolutional Neural Network-based a Novel Deep Trend Following Strategy for Stock Market Trading:

Predicting stock price trends grounded on current and literal trading exertion is a gruelling area of exploration. Recent advancements in machine literacy, particularly Convolutional Neural Networks (CNNs), have shown pledge in rooting meaningful features from raw data. A study introduced a Deep Trend Following (DTF) strategy, enhancing the traditional Trend Following approach, which assumes trends persist but struggles during trend reversals the DTF strategy integrates CNN- prognosticated future trends with current price trends for bettered decision- timber. Experimental results showed that the CNN-grounded classifier and the DTF strategy outperformed birth styles and traditional trading strategies, including Buy- and- Hold and simple Trend Following, when tested on stock indicators from the American and Indian requests. This exploration demonstrates the eventuality of combining prophetic models with traditional strategies to enhance stock request decision- timber.

2. Stock trend prediction using news sentiment analysis:

The Efficient Market Hypothesis (EMH) faces limitations in stock trend prediction, prompting alternative methods like financial news sentiment analysis. This study explored the relationship between news sentiment and stock trends using Random Forest (RF), Support Vector Machines (SVM), and Naïve Bayes models to classify news articles as positive or negative. RF and SVM outperformed Naïve Bayes, achieving over 80% accuracy—30% higher than random labelling these results highlight the potential of sentiment analysis and machine learning models in enhancing stock market predictions.

3. Stock trend prediction using news sentiment analysis:

This study addresses the limitations of the Efficient Market Hypothesis (EMH) by incorporating financial news sentiment analysis for stock trend prediction. Using Random Forest (RF), Support Vector Machines (SVM), and Naïve Bayes models to classify news articles as positive or negative, RF and SVM delivered strong performance, achieving over 80% accuracy—30% higher than random labelling the findings demonstrate the potential of sentiment analysis and machine learning models in enhancing stock market trend predictions.

4. Predicting stock price movements based on different categories of news articles:

Financial news articles have a significant influence on investors' decisions, thereby affecting market behaviour. This makes them a crucial source of data for financial predictions. Recently, various forecasting models have been developed using information from news articles. However, the potential benefits of incorporating different categories of news articles into these models have not been fully explored.

This research focuses on enhancing financial forecasting by simultaneously using news articles from five different categories, each relevant to varying levels of detail about a target stock the study employs the **Multiple Kernel Learning (MKL)** technique to integrate information from news articles categorized based on their relevance to the target stock, its sub-industry, industry, group industry, and sector. Separate kernels are used for each category to analyse the data.

the experimental findings demonstrate that using a combination of news from these five categories leads to a significant improvement in prediction accuracy compared to methods that use fewer categories of news articles.

5.The volatility of the stock market and news:

This study examines the impact of news events on stock market performance, focusing on indices like DJIA, NASDAQ, and S&P 500. Over ten weeks, it analyses the relationship between news, crude oil prices, and market volatility using a regression model the findings highlight that news significantly influences stock trends, driving market fluctuations the research also delves into risk assessment and market adaptation to news events, comparing news impact with traditional economic indicators using models like GARCH, offering a comprehensive view of how markets respond to news over time.

6. News analytics and sentiment analysis to predict stock price trends:

This study investigates how daily news events influence stock market performance, particularly indices like DJIA, NASDAQ, and S&P 500. Over a ten-week period, it examines the connection between news, crude oil prices, and market volatility using a regression model the findings reveal that news significantly impacts stock trends, driving market fluctuations the research also explores risk assessment, market adaptation, and probabilistic risk analysis, comparing the effects of news with traditional economic indicators through models like GARCH. It highlights the evolving role of news in shaping market behaviour.

7. Beyond Sentiment in Stock Price Prediction: Integrating News Sentiment and Investor Attention with Temporal Fusion Transformer:

This paper proposes a novel prediction model that integrates Fin-BERT-based news sentiment analysis with investor attention metrics, using an attention-based Temporal Fusion Transformer (TFT) framework. This approach not only enhances forecasting accuracy but also offers insights into the temporal dynamics that influence stock market behaviour the model’s effectiveness is demonstrated through the analysis of stock price data for 41 of the largest market capitalization companies from 2010 to 2021the results show that this approach outperforms existing deep learning methods, and the attention analysis highlights the importance of synthesizing both news sentiment and investor attention in accurately predicting stock prices.

8. Transformer-based attention network for stock movement prediction:

While previous methods have explored this fusion, challenges remain, particularly regarding the temporal dependence of financial data and the limited effectiveness of integrating text and stock prices.

To address these issues, this paper introduces a novel framework, the Transformer Encoder-based Attention Network (TEANet). This framework applies feature engineering on a limited dataset, utilizing a 5-day sample to effectively capture temporal patterns in financial data. TEANet utilizes the Transformer model along with multiple attention mechanisms for feature extraction and analysis, resulting in accurate predictions of stock movement. Simulations show that a trading strategy based on this model significantly boosts profits, demonstrating its practical application in stock trading.

9. Enhancing stock price prediction using GANs and transformer-based attention mechanisms:

This paper proposes a new approach to improve stock price prediction by combining Generative Adversarial Networks (GANs) with transformer-based attention mechanisms. GANs are used to create synthetic stock data while incorporating market sentiment. Attention mechanisms help focus on important features and patterns, aiding in the identification of key market indicators that influence stock prices the model also incorporates social media news to find sentiment, for better prediction accuracy and precision.

To overcome challenges like unrealistic data generation and overfitting often encountered with GANs and attention mechanisms, the authors use regularization techniques and integrate additional data sources. Experimental evaluations with real-world stock market data compare the proposed model's performance against traditional methods the findings offer valuable insights for investors, financial analysts, and stakeholders, helping refine investment strategies.

10. Transformer-Gated Recurrent Unit Method for Predicting Stock Price Based on News Sentiments and Technical Indicators:

The proposed method uses a classification model to derive daily sentiment scores, which are then refined using an active learning model to create sentiment indicators these indicators are integrated into stock price forecasting through a novel Transformer Encoder Gated Recurrent Unit (TEGRU) architecture the TEGRU model consists of a transformer encoder that captures patterns in time-series data using multi-head attention. This data is then passed to a Gated Recurrent Unit (GRU) layer to predict stock prices.

To evaluate the model's performance, the accuracy mean absolute percentage error (AcMAPE) is used, which is sensitive to misclassifications of stock price trends. Experimental results show that sentiment indicators enhance the forecasting model’s accuracy, with the TEGRU model outperforming other transformer-based architectures across five feature scenarios. Additionally, TEGRU helps minimize financial risk by determining the best-fit parameters for each stock issuer.

CHAPTER 3
WORK-DONE

This project involved developing a hybrid system to predict stock trading signals by integrating financial news sentiment analysis and numerical stock data. Multiple datasets were used to train the overall model architecture, emphasizing the importance of data quality in achieving accurate result. Better datasets lead to better outcomes. These datasets provided a rich source of information for training, ensuring the model could capture complex market dynamics. Emphasizing the importance of data quality, the project highlighted that better datasets yield more reliable and accurate predictions. Fin-BERT was utilized for analyzing news sentiment, while machine learning models processed stock metrics. The outputs of both models were fused to generate final trading signals (BUY, SELL, HOLD) along with potential profit or loss estimates. Comprehensive preprocessing and dataset splits ensured effective training and evaluation of the system, demonstrating its capability to support data-driven investment decisions. Beyond the model, the project also focused on building an accessible platform for end-users. A user-friendly web application was developed to serve as the interface for the system. This website allows users to input real-time financial news and stock data, view predicted trading signals, and assess the potential profit or loss from their investment decisions. The platform was designed to make the system's capabilities easily available to investors, offering a seamless experience with actionable insights. Overall, the project involved extensive research, advanced model development, and practical implementation through a web-based interface, providing a comprehensive solution for stock trading signal prediction and financial decision support. A significant emphasis was placed on exploring and comparing multiple datasets to evaluate their impact on the model's accuracy and reliability. Comprehensive research was conducted to identify and utilize the most relevant datasets, as the quality of input data directly influences the model's performance. The comparison of these datasets provided valuable insights into the nuances of market behaviour and their representation in various data sources.

3.1 Methodology:

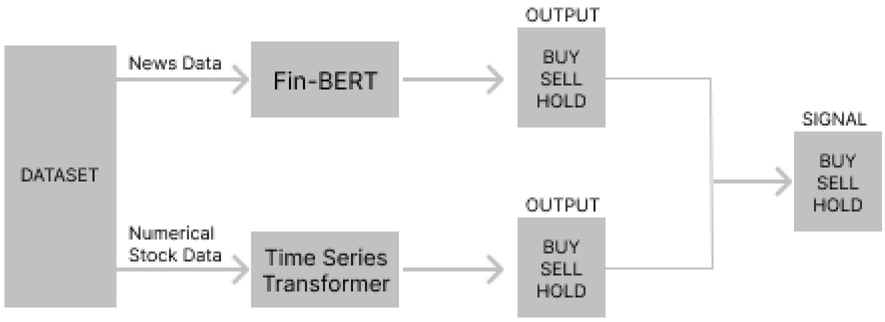

Figure 3.1 Diagram of proposed architecture of model

Fin-BERT Model

Fin-BERT is a tailored version of the BERT model, optimized specifically for analysing financial sentiment.

It is built to analyse and categorize the sentiment of financial content, including news articles, reports, and social media posts. In this project, Fin-BERT is used to analyse the sentiment of financial news and categorize it into three classes: positive, negative, and neutral these sentiments are then mapped to trading signals: BUY, SELL, and HOLD.

3.1.1 Preprocessing Financial News Data

Data Collection: Financial news articles are collected from various sources, including news websites, financial blogs, and social media platforms the dataset includes headlines and full articles.

Text Cleaning: The collected texts are cleaned to remove any irrelevant information, such as HTML tags, special characters, and stop words. This step ensures that the input to the Fin-BERT model is clean and relevant.

Tokenization: The cleaned texts are tokenized into words and sub-words using the BERT tokenizer. This step converts the text into a format that can be processed by the Fin-BERT model.

3.1.2 Training Fin-BERT

Model Architecture: Fin-BERT is based on the BERT architecture, which uses a transformer model with multiple layers of bidirectional self-attention. This architecture allows the model to understand the context of words in a sentence by looking at both the preceding and following words.

Fine-Tuning: The pre-trained Fin-BERT model is fine-tuned on a labeled dataset of financial news articles the labels indicate the sentiment of each article (positive, negative, or neutral). During fine-tuning, the model learns to associate specific words and phrases with their corresponding sentiments.

Output Layer: The final layer of the Fin-BERT model is a soft-max layer that outputs the probabilities of the three sentiment classes the predicted sentiment is determined by selecting the class with the highest probability.

3.1.3 Sentiment Classification

Sentiment Analysis: The fine-tuned Fin-BERT model is used to analyze the sentiment of new financial news articles. Each article is classified as positive, negative, or neutral.

Trading Signals: The classified sentiments are mapped to trading signals: positive sentiment corresponds to a BUY signal, negative sentiment corresponds to a SELL signal, and neutral sentiment corresponds to a HOLD signal.

3.2.1 Transformer based model to analyze stock price movement

This model is designed to analyze numerical stock market data and predict future stock price movements. It uses a transformer architecture to capture temporal dependencies and patterns in the data. The model is trained on historical stock prices and technical indicators to generate trading signals: BUY, SELL, and HOLD.

3.2.2 Preprocessing Stock Market Data

Data Collection: Historical stock market data is collected from financial databases, including metrics such as open, high, low, close, adjusted close, volume, and various technical indicators (e.g., MACD, Bollinger Bands, RSI).

Normalization: The collected data is normalized to ensure that all features are on a similar scale. This step helps the model to learn more effectively.

Feature Engineering: Additional features are created from the raw data, such as moving averages and momentum indicators these features provide more information to the model and improve its predictive performance.

3.2.3 Training the Transformer based model to analyze stock price movement

Model Architecture: This proposed model uses a transformer architecture with self-attention mechanisms to capture temporal dependencies in the data the model consists of an encoder and a decoder, each with multiple layers of self-attention and feed-forward neural networks.

Training Process: The model is trained on historical stock market data, with the input being a sequence of past stock prices and technical indicators, and the output being the predicted stock price movement (BUY, SELL, or HOLD) the training process involves minimizing a loss function that measures the difference between the predicted and actual stock price movements.

Output Layer: The final layer of the model is a soft-max layer that outputs the probabilities of the three trading signals the class with the highest probability is selected as the predicted signal.

3.2.4 Stock Price Prediction

Prediction: The trained Transformer model is used to predict future stock price movements based on new input data the model generates trading signals (BUY, SELL, HOLD) for each time step in the input sequence.

Trading Strategy: The predicted signals are used to make trading decisions. If the model predicts a BUY signal, the stock is bought; if it predicts a SELL signal, the stock is sold; and if it predicts a HOLD signal, no action is taken.

3.3.1 Fusion Model

the fusion model combines the outputs of the both models to make final trading decisions the fusion model takes into account both the sentiment of financial news and the predicted stock price movements to generate a more accurate and robust trading signal.

3.3.2 Combining Model Outputs

Consensus Mechanism: The fusion model uses a consensus mechanism to combine the outputs of the Fin-BERT for sentiment analysis of textual news data and Transformer model for stock trading prediction. If both models predict a BUY signal, the final signal is BUY; if both predict a SELL signal, the final signal is SELL; and if the predictions are conflicting, the final signal is HOLD.

Decision Making: The combined signals are used to make trading decisions. This approach leverages the strengths of both models and reduces the risk of making incorrect trading decisions based on a single model's prediction.

By integrating financial news sentiment analysis with stock trading forecasting, this methodology aims to enhance the accuracy of stock price predictions and provide a robust decision-making tool for investors the fusion model leverages the strengths of both textual and numerical data analysis, ultimately improving investment strategies and outcomes.

3.2 Datasets

This project leverages multiple datasets to train and evaluate the proposed model, each contributing uniquely to the system's capability to predict trading signals accurately. The datasets encompass both textual and numerical data, enabling a comprehensive analysis of market trends. Here's how these datasets were utilized in the project:

1. "NIFTY" Dataset (Hugging Face)

This dataset, containing textual and numerical stock market data from January 6, 2010, to September 21, 2020, was used to develop and fine-tune the sentiment analysis component of the model. The dataset was divided into training, validation, and test sets (70%, 15%, and 15%, respectively) to ensure robust evaluation and prevent overfitting. The textual data was processed using Fin-BERT for sentiment classification, while numerical data contributed to trend analysis.

2. NIFTY 50 Stock Market Data (2000-2023)

Sourced using the y-finance package, this dataset offered a rich set of 238,485 rows containing stock data for the top 50 companies on the NSE of India. It was instrumental in training the numerical model, providing metrics such as open, high, low, close, and volume. The large volume of data facilitated better generalization and helped identify patterns for generating accurate trading signals.

3. NIFTY-50 Stock Market Data (2000-2021)

Covering stock trends from January 1, 2000, to April 30, 2021, this dataset was used for comparative analysis alongside the 2000-2023 dataset. By examining overlapping periods, we validated the consistency of the model’s performance and explored the impact of historical data in forecasting stock trends.

4. Reliance Stock Prices with News Sentiment Dataset

This dataset, focusing on Reliance Industries Limited (RIL), played a crucial role in testing the model's performance on a single-company scale. Its integration of news sentiment and stock price trends made it ideal for evaluating the effectiveness of sentiment analysis in predicting trading signals. Interestingly, this dataset yielded exceptional results, demonstrating how domain-specific data can enhance accuracy.

5. News Sentiments India (2001-2022)

With its vast repository of news articles spanning over two decades, this dataset was pivotal in training the sentiment analysis model. The data helped refine the system's ability to discern market-relevant news and its polarity (positive, negative, or neutral). This analysis was then mapped to actionable signals like BUY, SELL, and HOLD, creating a bridge between textual information and trading decisions.

By integrating and comparing these datasets, the project ensured that the model's predictions were data-driven and accurate. Additionally, dataset comparisons revealed the significance of domain-specific data, such as the Reliance dataset, in achieving superior performance.

3.3 Architecture of proposed model:

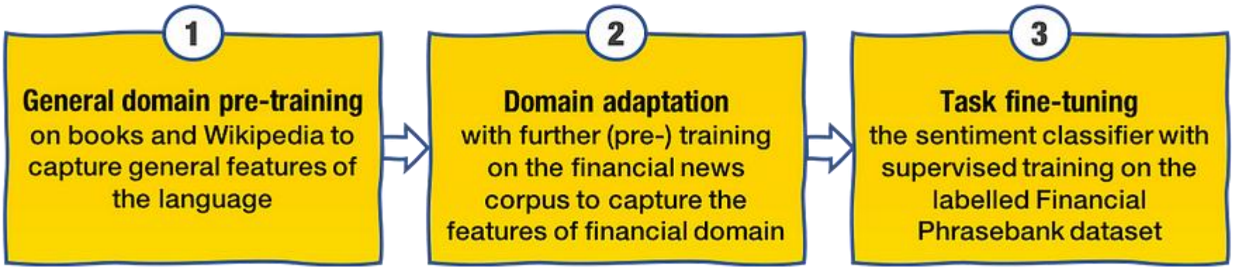

Figure 3.2 Diagram illustrating the three-stage process of developing a sentiment classifier

Fin-BERT is a domain-adapted variant of BERT specifically designed for financial sentiment analysis. It adapts the general-purpose BERT model to understand financial jargon and nuances effectively.

Fin-BERT is a specialized language model designed for financial text mining, built on the foundation of BERT (Bidirectional Encoder Representations from Transformers). Developed to address the unique challenges of financial text analysis, Fin-BERT leverages the power of pre-trained language models to enhance the understanding and processing of financial documents.

3.3.1 Background and Motivation

Financial text mining involves extracting valuable insights from various financial documents such as analyst reports, company announcements, and news articles. Traditional Natural Language Processing (NLP) models often struggle with financial texts due to the specialized vocabulary and context-specific language used in the financial domain. Moreover, the scarcity of labelled training data in finance further complicates the application of deep learning techniques.

Fin-BERT was developed as a domain-specific language model, pre-trained on extensive financial datasets to address these challenges. By focusing on financial texts, Fin-BERT aims to capture the nuances and intricacies of financial language, making it more effective for tasks such as sentiment analysis, sentence boundary detection, and question answering in the financial domain.

3.3.2 Architecture and Pre-training

Fin-BERT is based on the standard BERT architecture, which utilizes a multi-layer Transformer encoder the key innovation in Fin-BERT lies in its pre-training methodology. Unlike standard BERT, which uses Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) as pre-training tasks, Fin-BERT incorporates six self-supervised pre-training tasks designed to capture a broader range of language knowledge and semantic information.

1. **Span Replace Prediction**: Predicting masked spans of text to improve contextual understanding.
2. **Capitalization Prediction**: Identifying whether words are capitalized, which is crucial for recognizing named entities in financial texts.
3. **Token-Passage Prediction**: Determining the relevance of tokens within a passage to capture the main topics.
4. **Sentence De-shuffling**: Reordering shuffled sentences to learn sentence relationships.
5. **Sentence Distance Prediction**: Classifying the distance between sentences to understand their contextual proximity.
6. **Dialogue Relation Prediction**: Learning semantic relevance using question-answer pairs from financial dialogues.

these pre-training tasks enable Fin-BERT to learn from both general and financial domain corpora, enhancing its ability to understand and process financial texts effectively.

The diagram illustrates the encoder part of a Transformer model, which is essential for processing input sequences in natural language processing tasks. Here's a brief explanation:

1. **Input Embedding**: Transforms input tokens into dense vector forms for processing

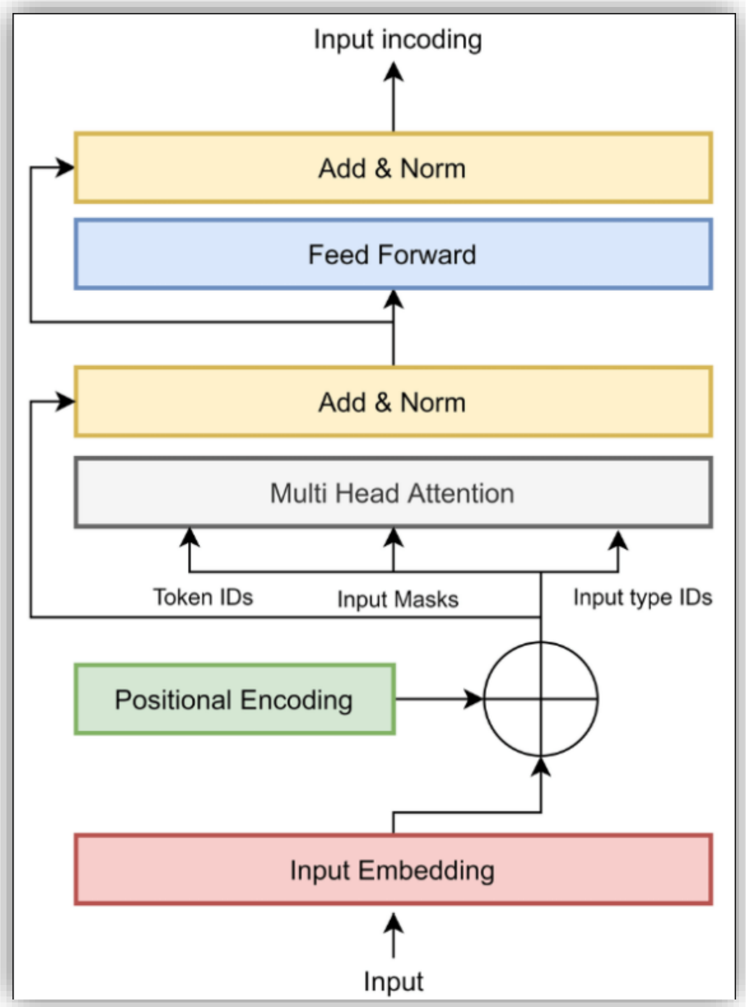

Figure 3.3 Architecture of Transformer's Encoder

2. **Positional Encoding**: Incorporates positional data into embeddings to maintain token order.
3. **Multi-Head Attention**: Enables the model to attend to multiple aspects of the input sequence at once.
4. **Add & Norm**: Applies residual connections and normalization to stabilize training.
5. **Feed Forward**: Processes the data through a feed-forward neural network.
6. **Add & Norm**: Another layer of residual connections and normalization.
7. **Output**: Produces the encoded representation of the input sequence, ready for the decoder.

This process enables the model to understand and encode the input data effectively.

3.3.3 Training Data

Fin-BERT was pre-trained on a diverse set of corpora, including:

- **General Domain Corpora**: English Wikipedia and Book-Corpus, totalling 13GB of text.
- **Financial Domain Corpora**: Financial news articles, Yahoo Finance articles, and Reddit Finance QA, totalling 48GB of text.

the combination of these corpora ensures that Fin-BERT captures both general language knowledge and domain-specific financial information.

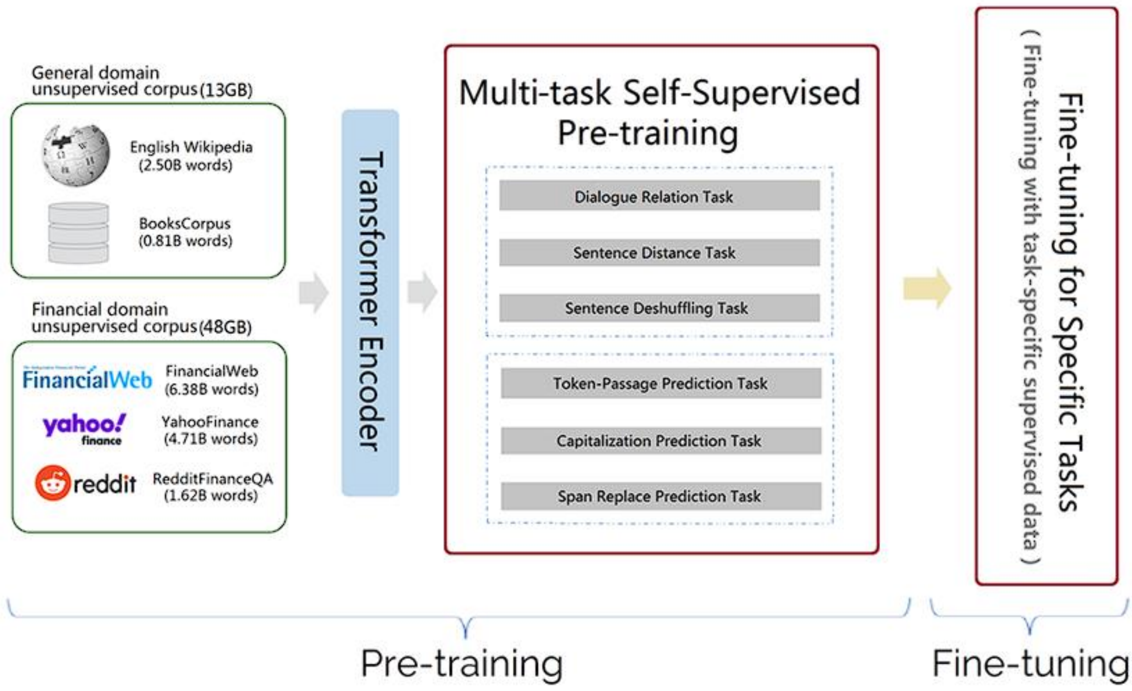

Figure 3.4 Pre-training and fine-tuning process of a transformer encoder using general and financial domain corpora, followed by multi-task self-supervised pre-training and task-specific fine-tuning.

3.4 Proposed Trading System:

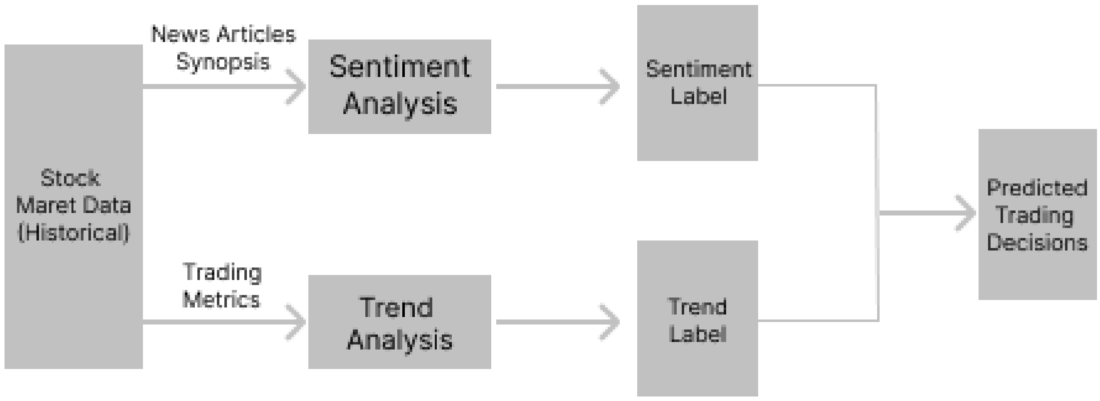

Figure 3.5 Diagram illustrating proposed trading system

This diagram represents a trading system that integrates sentiment analysis and trend analysis to generate predicted trading decisions. Historical stock market data serves as the foundation, providing input for both sentiment analysis and trend analysis the sentiment analysis module interprets the textual data from financial news or related sources to derive sentiment labels that reflect the overall market sentiment (e.g., positive, negative, or neutral). Simultaneously, the trend analysis module processes numerical stock data, such as historical price movements and technical indicators, to produce trend labels indicating whether the stock is in an uptrend, downtrend, or stable phase these outputs—sentiment and trend labels—are then fused in the final decision-making stage to produce predicted trading decisions. By combining insights from textual sentiment and numerical trend patterns, this system provides a holistic and data-driven approach to making informed trading decisions, balancing market sentiment with actual price trends.

3.5 Application Development:

In addition to the core development of the sentiment analysis and stock price prediction model, a key focus of our work was creating a user-friendly and interactive web application. Below is an outline of the web application development process tailored to our project

1. User Interface Design

Approach: Designed an intuitive, visually appealing user interface tailored for both casual investors and financial analysts.

Process: Wireframing and prototyping were carried out to ensure a seamless user experience. Iterative design cycles were used to refine the interface.

2. Frontend Development

Tools: The frontend was developed using modern frameworks like React.js combined with HTML, CSS, and JavaScript.

Features:

Interactive dashboards to display stock predictions and sentiment trends.

Real-time search functionality for stock-specific sentiment analysis and price forecasts.

Easy navigation and responsive UI for an enhanced user experience.

3. Backend Integration

Tools and Frameworks: Backend services were developed using Django to handle data processing, sentiment analysis, and stock prediction tasks.

API Development: Seamless communication between frontend and backend was ensured by implementing RESTful APIs these APIs handle tasks such as fetching sentiment results, retrieving historical stock data, and returning real-time predictions.

4. User Authentication and Authorization

Implementation: Secure user authentication was achieved with features like user registration, login/logout, and password encryption.

Authorization: Role-based access control ensures that only authorized users can access advanced analytics and administrative functions.

5. Responsive Design

Objective: Ensure the application performs optimally on various devices, including desktops, tablets, and mobile phones.

Approach: Implemented responsive design principles using CSS Flexbox and Grid Layout, with thorough testing across screen sizes and devices.

6. Testing and Quality Assurance

Testing Types:

Unit Testing: Validated individual components, ensuring robustness.

Integration Testing: Checked interactions between frontend, backend, and APIs.

User Acceptance Testing: Ensured the application met user expectations for usability and functionality.

Process: Rigorous bug fixing and performance optimizations were carried out.

7. Documentation

Documentation: Prepared a comprehensive user manual detailing how to use the application, including its key features and troubleshooting steps.

By creating this user-friendly web application, we have ensured that our stock price prediction system is accessible, interactive, and valuable for a wide range of users the platform empowers users with actionable insights by combining financial sentiment analysis with advanced stock forecasting tools.

CHAPTER 4
RESULT AND DISCUSSION

The experimental results of this work are divided into two primary tasks. In the first task, we focus on sentiment analysis using a custom-built model that classifies the sentiment of financial news and social media data into positive, negative, or neutral categories the sentiment analysis model was evaluated against baseline sentiment classifiers using precision, recall, and F1-score as performance metrics. Given the imbalanced nature of textual financial data, macro-average F1-score was employed as the primary metric to ensure a balanced evaluation across all classes. Our proposed model demonstrated superior performance compared to baseline classifiers due to its ability to leverage deep learning techniques and larger, diverse training data. Unlike the baseline models, which used traditional NLP-based feature extraction, our model incorporated a contextual understanding of financial language through advanced embeddings and sentiment-specific features, achieving higher accuracy in distinguishing market sentiment. Additionally, our model's ability to capture non-linear relationships and context-specific sentiment nuances contributed to its improved performance.

The second task involved trend analysis and its integration with the sentiment analysis results to develop a trading strategy the trend analysis module classified the stock price movements into upward, downward, or stable trends using numerical stock data and technical indicators the outputs from sentiment analysis and trend analysis were combined to generate a hybrid decision-making model for trading. This fusion-based trading strategy was tested against baseline methods, including the traditional trend-following approach and the buy-and-hold strategy.

The Fin-BERT model was fine-tuned using the NIFTY dataset alongside with Reliance dataset, starting from a pre-trained model provided by ProsusAI on Hugging Face.

Table 4.1 Performance Metrices for **NIFTY** dataset

Title	Accuracy	Precision	Recall	F1-Score
Performance Metrics of proposed Fin-Bert Model	0.4039	0.4310	0.4039	0.4102
Performance Metrics of proposed trading prediction system	0.5069	0.2570	0.5069	0.3410

Table 4.2 Performance Metrices for **Reliance** dataset

Title	Accuracy	Precision	Recall	F1-Score
Performance Metrics of proposed Fin-Bert Model	0.95	0.95	0.95	0.95
Performance Metrics of proposed trading prediction system	0.98990	0.99009	0.98990	0.98989

Observation :

- Domain-specific datasets, such as those focused on a single company like Reliance, enable models to learn patterns more easily due to the consistent, focused nature of the data. These datasets are more predictable and lead to higher performance because the stock price is often more closely tied to company-specific events and factors.

- Broader market datasets, like NIFTY, are inherently more complex because they represent a wide range of companies. These datasets require more fine-tuning, feature engineering to enhance prediction accuracy.
- Therefore, while domain-specific data can significantly improve model performance, broader datasets often demand more advanced techniques and larger datasets for improved accuracy and robustness.

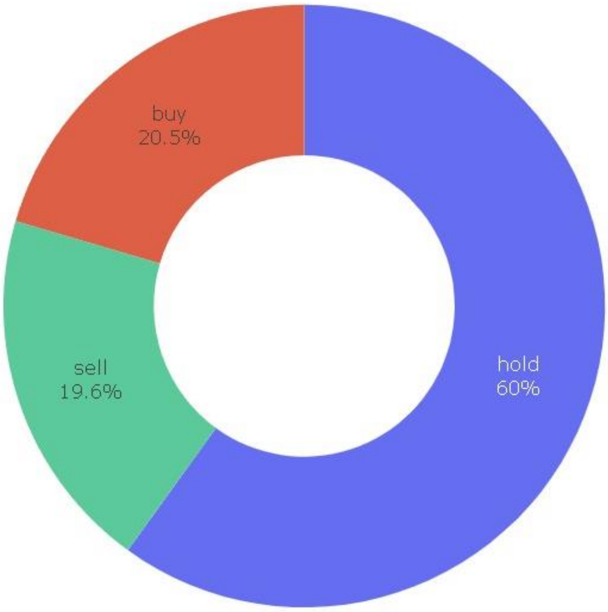

Figure 4.1 Distribution of Trade Signal after training 'NIFTY' financial stock data on model

The chart illustrates the distribution of trading decisions generated by the sentiment analysis model, categorized into **Buy**, **Sell**, and **Hold**. The largest portion, **60%**, is attributed to "Hold" decisions, indicating that the strategy predominantly recommended maintaining the current stock position. This suggests that the market conditions or sentiment analysis results were largely neutral or did not signal significant changes. **Buy** decisions accounted for **20.5%**, reflecting instances where the model identified opportunities for potential stock value growth. Meanwhile, **Sell** decisions constituted **19.6%**, suggesting moments when the model predicted a potential decline in stock value, prompting a recommendation to sell. This distribution highlights the strategy's cautious approach, favouring stability while selectively identifying buy or sell opportunities.

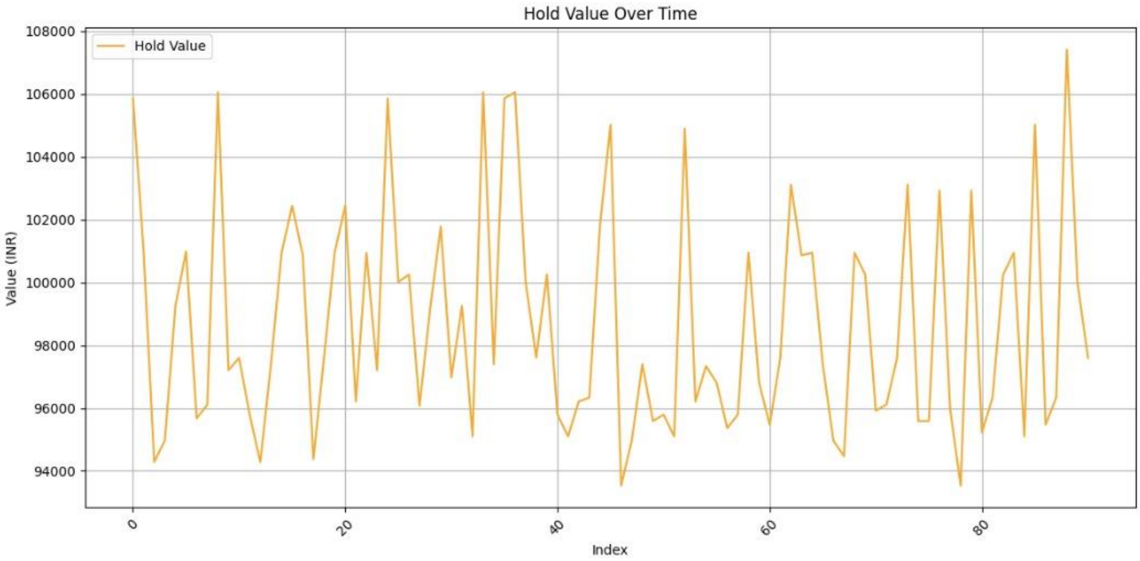

Figure 4.2 Hold Value Over Time (Reliance Dataset)

The fluctuations in the graph reflect the stock's daily price volatility. The portfolio's value shows significant oscillations, with peaks and troughs indicating price surges and declines, respectively. Despite occasional spikes, the general trend suggests an overall decline in value, as seen in the lower troughs and the eventual portfolio value being less than the initial investment. This outcome highlights the limitations of a passive holding strategy in volatile market conditions.

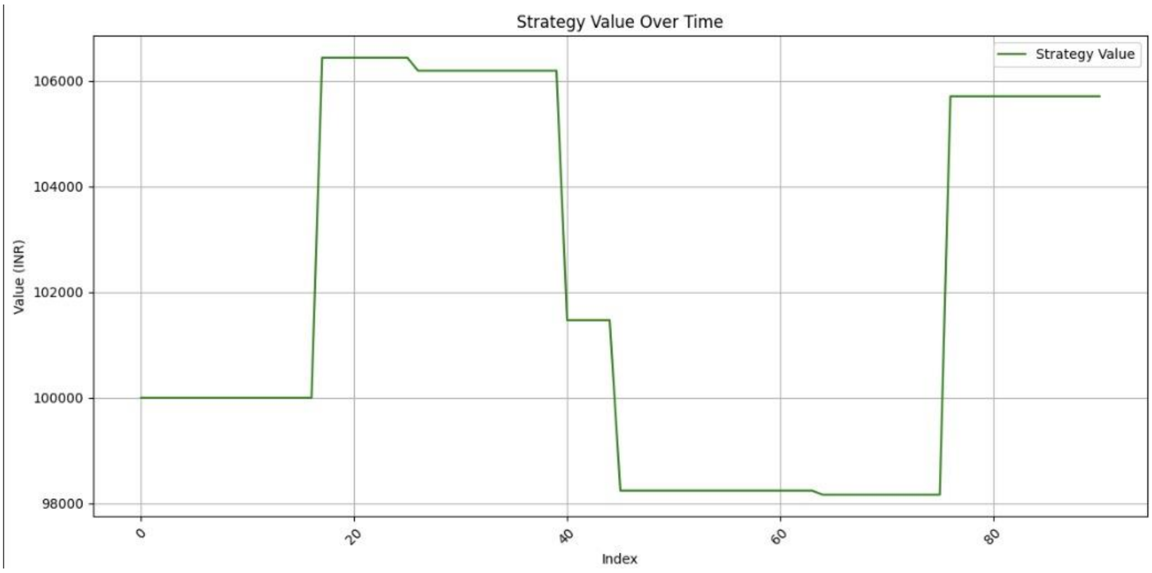

Figure 4.3 Enhanced Strategy Value (Reliance Dataset)

The strategy value remains stable for extended periods, with significant upward or downward changes corresponding to executed trades based on the sentiment and technical indicators. This highlights the strategy's controlled and calculated approach to trading.

The following graph compares the portfolio values of both strategies over time, showing the performance of the **Enhanced Strategy** (blue line) against the **Hold Strategy** (green line).

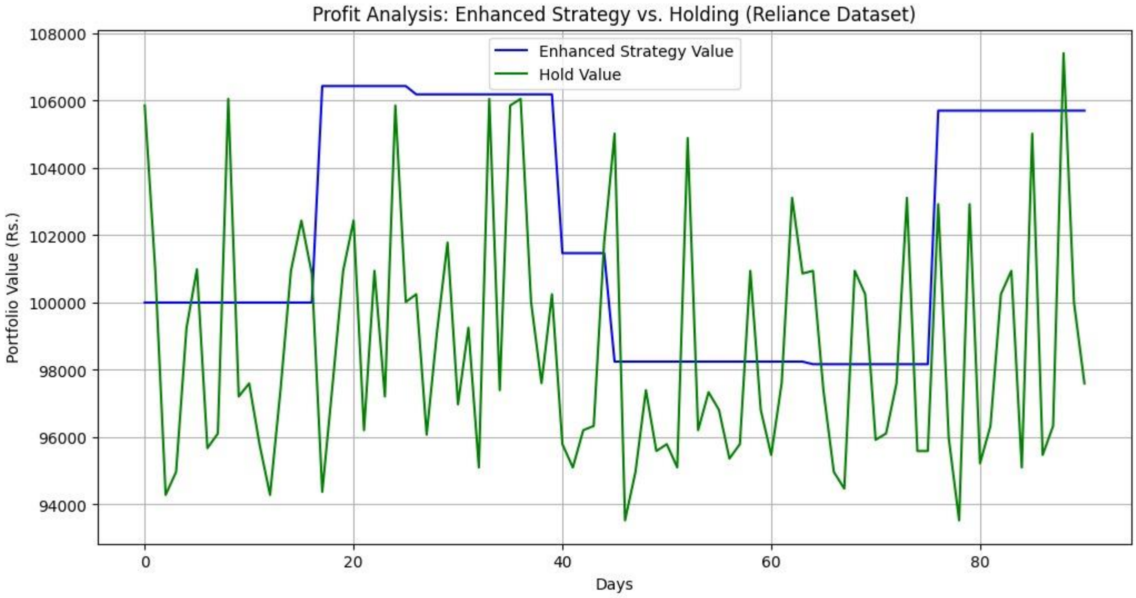

Figure 4.4 Portfolio value changes for the Enhanced Strategy and Hold Strategy over the analyzed period.

In this section, we present the results of a **Profit Analysis**, where we compare the portfolio performance of two strategies: one using **Fin-BERT sentiment analysis** along with technical indicators (Enhanced Strategy) and another employing a straightforward **Buy and Hold approach** over the same dataset (Reliance stock prices).

- **Enhanced Strategy:** Uses sentiment analysis to make daily buy, sell, or hold decisions, based on the sentiment of financial news and technical indicators.
- **Hold Strategy:** A simple approach where the stock is bought and held without any further trading decisions

Observations:

- The **Enhanced Strategy** demonstrates a **more stable performance** with minimal fluctuations in portfolio value.
- The **Hold Strategy** shows **greater volatility** due to the lack of active trading decisions, which results in a higher risk.
- The **Enhanced Strategy** appears to manage risks more effectively, yielding consistent returns compared to the **Hold Strategy**, which does not adjust for market conditions.

The **Enhanced Strategy**, which incorporates **Fin-BERT sentiment analysis** and **technical indicators**, significantly outperforms the **Hold Strategy** in terms of reducing risk and providing more consistent portfolio growth. This highlights the potential benefits of utilizing advanced predictive models in stock trading.

Table 4.3 Profit Comparison Between Enhanced Strategy and Hold Strategy (**Reliance** Dataset)

Initial Investment	Final Strategy Value	Final Hold Value	Total Profit Based on Sentiment Strategy	Total Profit by Holding
Rs. 100,000	Rs. 105,702.43	Rs. 97,598.34	Rs. 5,702.43	Rs. -2,401.66

Table 4.4 Profit Comparison Between Enhanced Strategy and Hold Strategy (**NIFTY** Dataset)

Initial Investment	Final Strategy Value	Final Hold Value	Total Profit Based on Sentiment Strategy	Total Profit by Holding
Rs. 1,13,680	Rs. 2,72,030	Rs. 2,72,029	Rs. 1,58,350	Rs. 1,58,349

Observations:

- The Reliance dataset achieved a solid profit of Rs. 5,702.43 while the NIFTY dataset only yielded highlights the superior performance of the sentiment-based strategy when applied to market indices.
- The Reliance dataset being domain-specific allowed the model to focus on relevant company news, leading to more accurate predictions and greater profitability. On the other hand, the NIFTY dataset, which is a composite of various companies across sectors, requires a more holistic approach.

- Domain-specific datasets yield better results due to the ability to leverage more targeted and relevant information, while broader datasets may require more complex models that can handle multiple data sources and external factors.

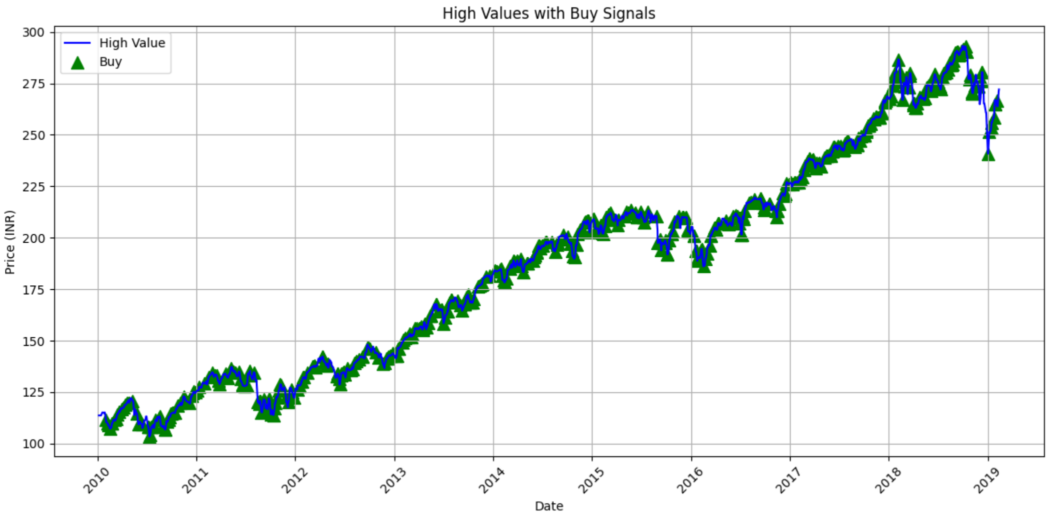

Figure 4.5 High values with Buy signals

This graph plots the daily high values of the NIFTY index (or the specific stock you're analyzing) as a blue line. Green upward-pointing triangles are overlaid on the line to represent the points where your trading strategy generated a "Buy" signal.

When you see a green triangle, it means your combined sentiment analysis (from Fin-BERT and numerical features) indicated a positive outlook, suggesting it might be a good time to buy the asset. These buy signals are often triggered when the price is trending upwards or showing signs of potential growth. These signals indicate periods where the combined sentiment analysis, derived from Fin-BERT and numerical features, suggested a positive outlook for the asset, potentially indicating an opportune time for purchase

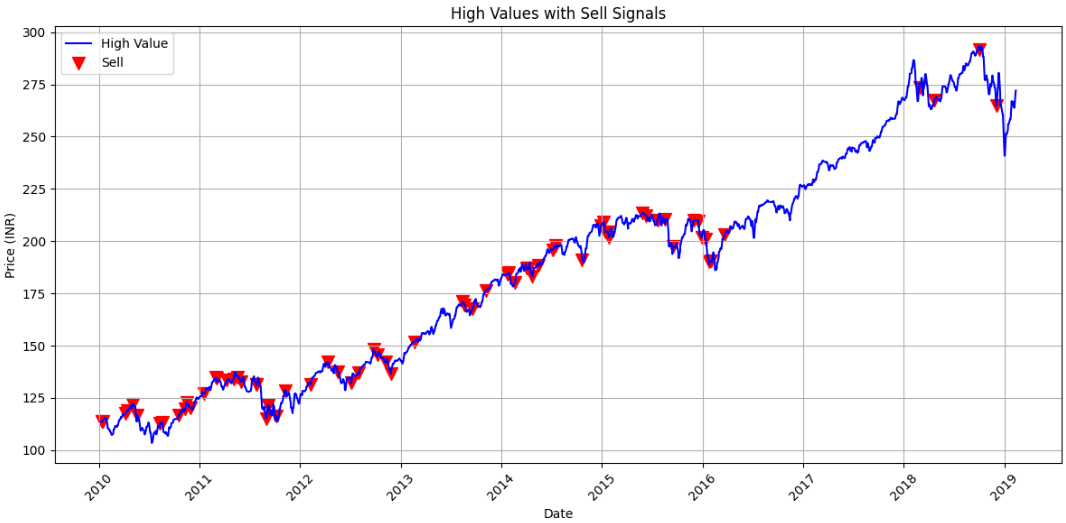

Figure 4.6 High values with Sell signals

Similar to the buy signal graph, this one displays the daily high values as a blue line. However, it uses red downward-pointing triangles to mark the points where a "Sell" signal was generated. Red triangles indicate that your combined sentiment analysis identified a negative outlook, suggesting it might be a good time to sell the asset to potentially avoid losses. Sell signals are often triggered when the price is trending downwards or showing signs of potential decline. These signals highlight periods when the hybrid model's combined sentiment analysis identified a negative outlook, potentially suggesting an advantageous time to sell the asset and mitigate potential losses.

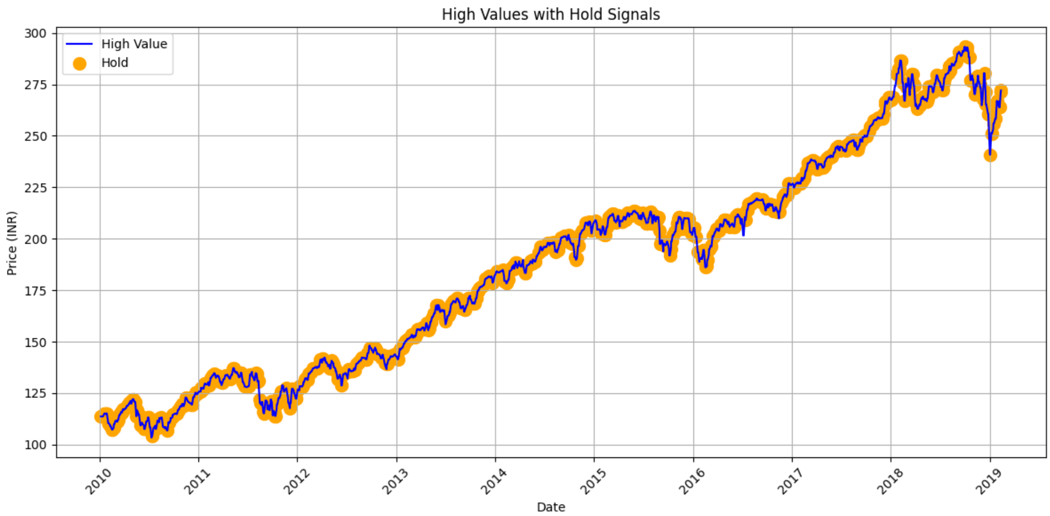

Figure 4.7 High values with Hold signals

This graph, again, uses a blue line to represent the daily high values. Orange circles are placed on the line to mark the points where a "Hold" signal was generated. Orange circles represent periods where your combined sentiment analysis didn't provide a clear buy or sell signal. This might indicate that the market sentiment is neutral or that there's not enough evidence to make a strong trading decision. During these periods, your strategy suggests holding your current position (neither buying nor selling). These signals represent periods where the hybrid model's combined sentiment analysis did not provide a clear buy or sell indication. This might signify neutral market sentiment or insufficient evidence for a strong trading decision, prompting the strategy to recommend maintaining the current position (neither buying nor selling).

These three graphs help you visualize how your combined sentiment analysis translates into trading actions. By overlaying the buy, sell, and hold signals on the price chart, you can see how these signals align with the actual price movements of the asset. This visualization can be helpful in evaluating the effectiveness of your trading strategy and understanding when it might be appropriate to enter or exit trades.

CHAPTER 5
SUMMARY AND CONCLUSION

This conversation encapsulates the meticulous process of building and refining a **stock price prediction model** that integrates textual sentiment analysis with numerical market data to enhance accuracy and decision-making capabilities. The hybrid approach, combining **Fin-BERT for sentiment analysis** and **Transformer-based models** for numerical data, has been a central focus, showcasing the synergy between textual insights and historical stock trends in predicting market movements effectively.

The comparative analysis of datasets, particularly the **Reliance-specific dataset** versus the broader **NIFTY dataset**, revealed the critical role of **domain-specific data** in achieving higher accuracy and profitability. The **Reliance dataset** yielded exceptional performance metrics and profitability, reflecting the value of focused, company-specific data for stock predictions. Conversely, the **NIFTY dataset**, encompassing multiple companies, highlighted the challenges of broader market data, such as handling diverse trends and achieving comparable precision. This reinforces the necessity for **advanced preprocessing, clustering by sectors, and techniques like multi-kernel learning** to address such complexities.

Performance metrics across different models underscored the effectiveness of combining textual and numerical data for robust financial forecasting. The integration of news sentiment with trading signal predictions was particularly impactful, translating into actionable insights for investors. The proposed model's adaptability was further demonstrated through web-based applications, which extend its utility to real-world financial decision-making.

Additionally, the discussion touched on the challenges and limitations of the current approach, such as managing overfitting, enhancing sentiment modeling, and ensuring robustness across different datasets. The conversation also emphasized the importance of incorporating **fake news detection mechanisms**, integrating macroeconomic indicators, and fine-tuning models to align with market dynamics.

To further enhance the proposed model, future research could focus on **expanding the dataset scope** to include global markets and cross-industry data, refining **sentiment analysis with multi-lingual capabilities**, and integrating **alternative data sources** like social media trends and macroeconomic indicators. Incorporating **state-of-the-art generative models** could improve data augmentation and trend analysis. Moreover, extending the model's application to other domains, such as commodities or real estate, and deploying it in real-time trading environments will ensure greater **scalability and versatility** in the ever-evolving financial landscape.

REFERENCES

[1] Chakole, Jagdish, and Manish P. Kurhekar. "Convolutional Neural Network-based a Novel Deep Trend Following Strategy for Stock Market Trading." *CIKM Workshops*, 2021.

[2] Kalyani, Joshi, Prof Bharathi, and Prof Jyothi. "Stock trend prediction using news sentiment analysis." *arXiv preprint arXiv:1607.01958*, 2016.

[3] Gu, Wen, et al. "Predicting stock prices with Fin-BERT-LSTM: Integrating news sentiment analysis." *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing*, 2024.

[4] Shynkevich, Yauheniya, et al. "Predicting stock price movements based on different categories of news articles." *2015 IEEE Symposium Series on Computational Intelligence*. IEEE, 2015.

[5] Goonatilake, Rohitha, and Susantha Herath. "the volatility of the stock market and news." *International Research Journal of Finance and Economics* 3.11 (2007): 53-65.

[6] Chowdhury, Spandan Ghose, Soham Routh, and Satyajit Chakrabarti. "News analytics and sentiment analysis to predict stock price trends." *International Journal of Computer Science and Information Technologies* 5.3 (2014): 3595-3604.

[7] Hajek, Petr, and Josef Novotny. "Beyond Sentiment in Stock Price Prediction: Integrating News Sentiment and Investor Attention with Temporal Fusion Transformer." *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Cham: Springer Nature Switzerland, 2024.

[8] Zhang, Qiuyue, et al. "Transformer-based attention network for stock movement prediction." *Expert Systems with Applications* 202 (2022): 117239.

[9] Li, Siyi, and Sijie Xu. "Enhancing stock price prediction using GANs and transformer-based attention mechanisms." *Empirical Economics* (2024): 1-31.

[10] Haryono, Agus Tri, Riyanarto Sarno, and Kelly Rossa Sungkono. "Transformer-gated recurrent unit method for predicting stock price based on news sentiments and technical indicators." *IEEE Access* (2023)