A Share Price prediction technique for the Indian Stock Market using Sentiment Analysis

Project report submitted for the partial fulfillment of requirement for the degree of **Bachelor of Technology (B. Tech)** in **Computer Science and Engineering** of the University of Calcutta

Submitted by

Swarup Das	**Akash Das**	**Shounak Roy Chowdhury**
B.Tech. 8th Semester	B.Tech. 8th Semester	B.Tech. 8th Semester
Exam Roll Number:	Exam Roll Number:	Exam Roll Number:
T91/CSE/206013	T91/CSE/206001	T91/CSE/2060
Registration No: D01-1111-0042-20	Registration No: A01-1112-0850-18	Registration No: D01-1211-0148-20

Under the guidance of

Prof. Sankhayan Chowdhury

Department of Computer Science and Engineering

University of Calcutta

JD-2, Sector-III, Salt Lake City, Kolkata 700106

May, 2024

Department of Computer Science and Engineering

University of Calcutta

JD-2, Sector-III, Salt Lake City, Kolkata 700106

Certificate

This is to certify that **Swarup Das**, **Akash Das**, and **Shounak Roy Chowdhury**, students of the 8th semester of Bachelor of Technology (B.Tech.) in Computer Science and Engineering in the Department of Computer Science and Engineering, University of Calcutta, have completed their project report entitled "**A share price prediction technique for the Indian Stock market using sentiment analysis**" under my supervision and guidance, for the partial fulfillment of their degree of Bachelor of Technology in Computer Science and Engineering at the University of Calcutta.

_____

Examiner(s)
Place: Kolkata, West Bengal, India
Date:

_____

Chairman
Board of Studies
University of Calcutta
Place: Kolkata, West Bengal, India
Date:

_____

Supervisor
Prof. Sankhayan Chowdhury
Department of Computer Science and Engineering
University of Calcutta
Place: Kolkata, West Bengal, India
Date:

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Stocks, also known as shares or equities, represent ownership in a company. When an individual buys stocks, they acquire a piece of the company and become a shareholder [5]. Stocks are traded on stock exchanges, which are marketplaces where buyers and sellers come together to conduct transactions. There are several different types of stocks. Stocks can be categorized based on several parameters like ownership, market capitalization, dividend payout etc. The stock market lets an individual trade in bonds, mutual funds, derivatives, shares of a company, etc. whereas a share market only allows the trading of shares. A Stock Exchange is a vital component of a stock market that facilitates the transaction between traders of financial instruments and targeted buyers.[1]

When a company decides to go public, it offers its shares for the first time through an IPO(Initial Public Offering). This process involves regulatory approval, setting a price, and finally selling the shares to investors. After the IPO, stocks are traded on the stock market. Investors buy and sell stocks through brokers, who execute trades on their behalf. Once a trade is executed, the ownership of the stocks is transferred from the seller to the buyer. In India, the settlement cycle is T+2, meaning transactions are settled within two business days. Companies may distribute a portion of their profits to shareholders in the form of dividends, providing regular income to investors. The working of the stock exchange is depicted in Fig. 1.1.

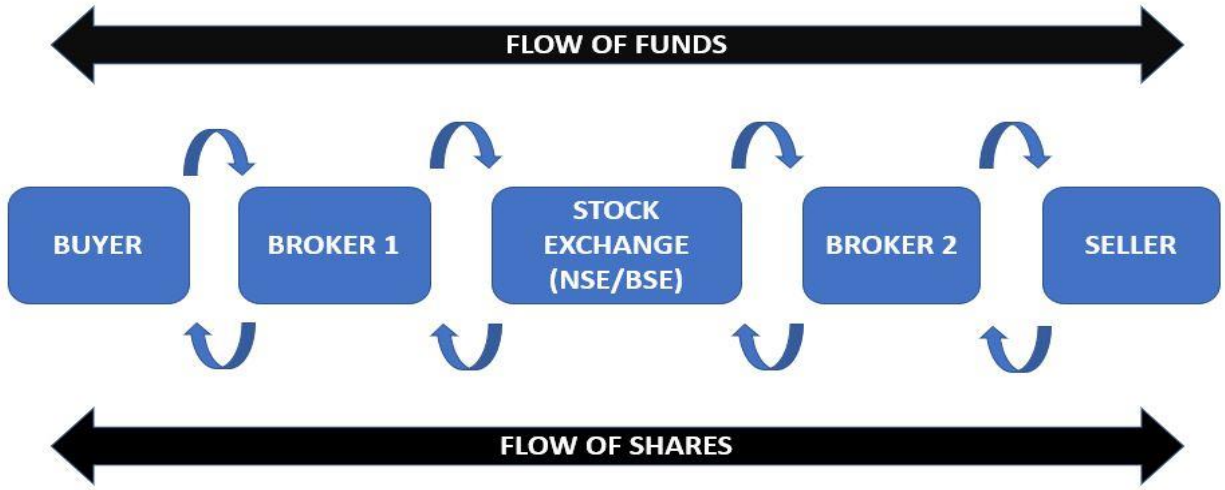


Fig. 1.1 - Working of the stock exchange[1]

Stock indices are indicators that reflect the performance of a group of stocks. They provide a snapshot of market trends and help investors gauge overall market performance[1]. In India, there are two primary stock indices. The S&P BSE Sensex, also known as the Sensex, comprises 30 of the largest and most actively traded stocks on the Bombay Stock Exchange (BSE). It is a key indicator of the Indian stock market's health[1,7]. The Nifty 50, managed by the National Stock Exchange (NSE), includes 50

major stocks from various sectors, providing a comprehensive overview of the market[1,7]. A snapshot of the Nifty 50 historical chart as of 5th June 2024 at 15:32 is shown in Fig. 1.2.

There are several financial metrics and Key Performance Indicators(KPIs) that are crucial tools used to evaluate the performance and health of a company. These indicators provide insights into various aspects of a company's operations, financial health, and overall success, helping stakeholders make informed decisions. Liquidity metrics, such as the Current Ratio and Quick Ratio, assess a company's ability to meet short-term obligations[5,8,9]. The Debt-to-Equity Ratio provides insights into long-term financial stability by comparing a company's total liabilities to its shareholder equity[8,9]. The Earnings Per Share (EPS) and Price-to-Earnings (P/E) Ratio, provide insights into a company's market valuation and profitability per share.[8,9]

Fig. 1.2 - Nifty 50 historical chart[12]

A share market is, well, a market. Hence, like any other market, the price of products being sold in the market is determined by the demand and supply of the said product[1]. If the demand for a particular stock increases for any reason, the stock price starts rising. Similarly, if there is a drop in demand for a particular share, we see that fewer bidders are attracted pulling the stock price low. Investor, as well as market sentiment, drives the demand. The key factors that influence stock price are company-related, industry-related factors, geopolitical factors, microeconomic factors, etc. A major example in this context is the crash of the share market on June 4, 2024, due to the results of the General Elections 2024 declared on that day. By crashing up to 8.5% intraday on Tuesday, Nifty suffered its biggest fall since March 2020 when the whole world was stressed about Covid-19. While the Sensex fell up to 6,234 points during the day, Nifty bulls saw an erosion of nearly 2,000 points at the day's low. On BSE, over 3,400 stocks fell with 745 hitting the lower circuit limit and 285 stocks touching 52-week

low levels. The market capitalization of all listed companies on BSE declined by Rs 45.56 lakh crore to Rs 380.35 lakh crore during the day. Exit poll predictions unleashed the bulls on Dalal Street on Monday with predictions that the ruling NDA may win over 350 seats in the elections. However, the actual results caused the crash of the market.

In the stock market, there are two types of market trends, namely Bullish and Bearish trends. The Bullish trend denotes that the stock prices are rising whereas the Bearish trend indicates that the stock prices are going down. Several parameters define the price of a stock. In a historical chart, these parameters are often depicted in the form of a 'candlestick'[Fig. 1.3], where a green candlestick denotes a bullish trend and a red candlestick denotes a bearish trend. These parameters form the prerequisites for our project. They are as follows:

- **Open** - The price at which a stock first trades upon the opening of an exchange on a trading day.
- **High** - The high price is the highest price at which a stock trades during the trading day.
- **Low** - The low price is the lowest price at which a stock trades during the trading day.
- **Close** - The closing price is the final price at which a security trades at the end of the trading day.
- **Adjusted close(Adj close)** - The adjusted closing price reflects any corporate actions, such as dividends, stock splits, or bonus issues, that occurred after the market close. It is the closing price adjusted to incorporate these factors and is often used to calculate returns over time.
- **Volume** - Volume or trading volume, means the number of shares traded over a particular period, typically a trading day. It includes every share that is bought and sold during the period in review.

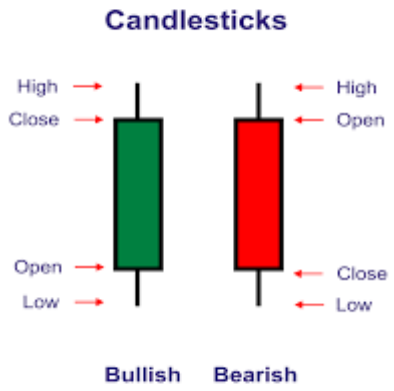


Fig. 1.3 - A candlestick generally used in a historical chart[13]

Stock analysts attempt to determine the future activity of an instrument, sector, or market. It is the practice of using information and analyzing data to make investment decisions. One popular form of stock analysis is fundamental analysis, the practice of using financial activity to forecast stock prices. Another popular form of stock analysis is technical analysis, the reliance on historical stock price activity to predict future price activity. The notion of stock analysis relies on the assumption that

available market information can be used to determine the intrinsic value of a stock. In the primary methods, investors use financial statements, stock price movements, market indicators, or industry trends to make investment decisions. Much of this strategy relies on leveraging historical information. For instance, investors may analyze a company's stock based on its financial performance. Various technical analysis tools allow trading decisions to be made without emotion. These stock analysis tools also help generate buy and sell indications and aid in discovering fresh trading possibilities. Some of the best tools for stock analysis are Screener Plus, Thinkorswim, Active Trader Pro, Slope of Hope, and Interactive Brokers.

The objective of the project is to predict the stock price for a given range of dates based on news articles and numerical parameters(namely, Open, High, Low, Close, Adj Close, and Volume), through machine learning and sentiment analysis approaches. In the next chapter, we will be focussing on the background of the project and the motivation that led us to develop this project.

Chapter 2

Background and Motivation

In this chapter, we are going to discuss the various aspects that motivated us to develop this project. In the ever-evolving landscape of financial markets, staying ahead of trends and making informed investment decisions is a complex challenge. The amalgamation of technological advancements and data-driven strategies has opened new avenues for investors seeking a competitive edge. Traditional methods for predicting stock prices have largely relied on quantitative data, such as historical prices, trading volumes, and various financial indicators. In recent years, sentiment analysis has gained traction as a complementary approach for stock market prediction. Sentiment analysis involves extracting subjective information from textual data, such as news articles, social media posts, etc. to gauge the mood or sentiment of the market as positive, negative, or neutral. The underlying hypothesis is that market sentiments can influence investor behavior and consequently affect stock prices.

Despite the global interest in stock market prediction, we found there is a notable gap in research focused specifically on the Indian stock market. There is very little or almost no work on the Indian stock market. The Indian stock market is one of the largest and fastest-growing markets in the world, yet it has received relatively less attention in the domain of stock prediction. This presents a unique opportunity to develop and test new prediction techniques tailored to the characteristics and dynamics of the Indian market. The primary motivation for this project is to fill the research gap related to sentiment analysis-based stock prediction in the Indian stock market. By focusing on an under-researched market, the project aims to provide novel insights and contribute to the existing body of knowledge.

Traditional stock prediction models often overlook the wealth of information embedded in textual data. We propose a novel approach that fuses sentiment analysis of textual data with historical stock prices to enhance prediction accuracy. The fusion of these two data types can provide a more holistic view of the factors driving stock prices, capturing both quantitative trends and qualitative sentiments. Integrating sentiment analysis with historical price data has the potential to improve prediction accuracy. Sentiments expressed in news articles and other textual sources can offer early indicators of market movements, complementing the information derived from historical price trends. This multimodal approach can lead to more robust and reliable prediction models and help in making informed decisions. By leveraging sentiment analysis, the model can provide early warnings of potential price movements, helping investors strategize their trades more effectively. This is particularly relevant in the Indian market, where investor sentiment can be significantly influenced by local news and events. Belov

Therefore, our project aims to develop a share price prediction technique for the Indian stock market by integrating sentiment analysis with historical stock prices. The motivation behind this approach is to address the existing research gap, enhance prediction accuracy through the fusion of diverse data sources, and provide practical benefits for investors in the Indian stock market. By pioneering this novel approach, the project seeks to contribute valuable insights and methodologies to the field of financial data science. In the next chapter, we will be discussing the related works and the present state-of-the-art solutions.

Chapter 3

Literature Survey

In this chapter, we are going to delve into the present state-of-the-art available in this domain. We have gone through research papers and can conclude that the state-of-the-art solutions are mainly of 3 types. We can classify them as using numerical data(historical stock prices) only, using textual data(sentiment analysis) only, and using a fusion of both of them. The first method focuses on analyzing past stock prices and trading volumes to forecast future movements. Techniques like technical analysis and time series analysis (e.g., ARIMA, GARCH, and LSTM networks) are commonly used. These models identify patterns and trends in historical data, assuming that past performance can predict future behavior. The second approach leverages textual data from sources like news articles, social media, and financial reports to gauge market sentiment. Natural Language Processing (NLP) techniques extract sentiment scores, which are then used in machine learning models (e.g., logistic regression, SVM, neural networks) to predict stock price movements. Sentiment analysis captures qualitative factors and provides early indicators of market trends based on public sentiment. Lastly, the hybrid approach combines historical price data with sentiment analysis from textual sources to leverage the strengths of both data types. This method aims to provide a more comprehensive and accurate prediction model by integrating quantitative and qualitative factors. Techniques involve data integration where historical prices and sentiment scores are used as input features for advanced machine learning models, including ensemble methods and deep learning frameworks capable of handling multimodal data. This fusion approach offers a holistic view of the market, potentially improving prediction accuracy, but it requires sophisticated integration techniques and is computationally more complex.

We first consider those works that used only textual data for stock price prediction. In the paper of Qianyi Xiao and Baha Ihnaini[2], the authors formulated the problem as a classification task and made a model that can predict the next day's stock movement i.e. uptrend or downtrend. This work is an example of how they used VADER, FinBERT, and Loughran-McDonald Word Dictionary to analyze sentiment from unstructured data. It was found that VADER is useful for handling Tweet data and FinBERT is useful for handling news data. The authors checked tweets, and news in two intervals for a normal day. They also considered the holiday's effect and day-of-the-week’s effect on the stock market. Their model checks the price of a stock and analyzes that stock-related news or tweet data at a given interval of times - Natural hours division and Opening hours division. The weighted sentiment data are found by summing tweets and news data together. Finally, the extracted sentiment from sources is given as an input in different Machine Learning models (Random Forest, Naïve Bayes, Logistic Regression, KNN, SVM, Tree). Naive Bayes was found to be the best classifier algorithm. It may be the case that their source dataset is biased as Naive Bayes always gives the best results when one class is much more prevalent than the others. Their model gets the best 62.4% accuracy. Compared with Kabbani and Usta’s result (Kabbani & Usta, 2022), which gets 63.6% as the best accuracy by using seven features (including sentiment score and technical factors), their model gets the best 62.4% accuracy with only news and tweets sentiment features.

In another study[3], the aim is to investigate the impact of President Trump's tweets on the stock market, particularly the direction of the Dow Jones Industrial Average (DJIA) on an hourly basis. The authors seek to understand whether and how Trump's tweets, when echoed by financial news, can be used to improve the accuracy of stock market predictions. The study focuses on capturing the influence of Trump's market-related tweets and their reinforcement by financial news to predict hourly

stock market trends. The authors developed a deep information echoing model using a bi-directional long short-term memory (BiLSTM) network. They too framed the problem as a classification task. This model integrates Trump's tweets and financial news to capture the echoing effects that influence market trends. The model employs hierarchical attention strategies to highlight crucial words and constructs effective representations of each period. Additionally, pre-training and fine-tuning techniques were used to enhance model training, along with a temporal weight function to adjust for shifts in tweet topics over time. Extensive experiments were conducted on real-world DJIA data to validate the model's effectiveness. The proposed deep information echoing model outperformed other baseline methods, achieving an accuracy of 60.42% in predicting stock market trends. The model also demonstrated significant accumulated profits in trading simulations, confirming that Trump's tweets contain indicative information for short-term market movements. The analysis revealed that tweets about finance, trade, and politics had a more substantial impact on the market. The study concluded that incorporating Trump's tweets into prediction models enhances the accuracy and provides better interpretability of market movements.

We discuss a work that used a fusion of textual data along with historical stock prices. In this study[4], the authors predict stock value by using sentiment data along with the historical Close Price and other Macro parameters to make predictions using the Random Forest Model. They used LSTM to predict stock values by analyzing historical closing stock values. LSTM is used as it performs well for time series data. Instead of using 3 sentiment outputs - negative, positive, and neutral, they also used another Compound sentiment. By calculating the closing stock price with 4 sentiment features and 4 external features in the Random Forest Model the result is found. The authors aimed to develop trading strategies for the real-world application of developed models. They used two regression models, LSTM and Random Forest, to predict the next day's value, with RMSEs as evaluation metrics. However, certain confidence levels were not achieved due to the regression exercise. A simple intuitive analysis of RMSE values was conducted to gauge the appropriateness of the predicted values. The authors compare the RMSE values for both LSTM and Random Forest. Sometimes Random Forest works well and sometimes LSTM. The results showed that the Mean Absolute Percentage Error (MAPE) varied between 1.36% and 1.81% for LSTM and 1.25% to 3.76% for Sentiment Analysis using the Random Forest Model. HDFC Bank performed better than LSTM, while SBI was the best-performing stock in both models. A 95% confidence level could be considered an approximate fit for the model's workings. The study did not develop any trading strategies but attempted to forecast future price trends instead of just predicting a single-day price. However, the trend forecasting results were not satisfactory, and significant changes may be needed for future results.

We can conclude that Xiao and Ihnaini's study[2] demonstrated the effectiveness of sentiment analysis alone in predicting stock movements, achieving the highest accuracy. The Trump tweet study[3] offered valuable insights into the specific impacts of influential tweets but with slightly lower accuracy. The final study[4] highlighted the benefits and challenges of combining textual sentiment with historical price data, achieving variable results depending on the model and stock analyzed. Each study underscores different facets of utilizing textual data for stock market prediction, from sentiment analysis and influential tweets to the fusion of historical and textual data.

Hence, our literature survey highlights the extensive research conducted on stock price prediction using various methodologies, including those based solely on historical prices, sentiment analysis of textual data, and the fusion of both data types. While significant advancements have been made globally, particularly in markets like the US and Europe, there is a noticeable gap in research focused on the Indian stock market. Most existing studies either overlook the unique characteristics of

the Indian financial landscape or do not leverage the powerful combination of historical and sentiment data specific to this market.

Therefore, this gap underscores the need for our project, whose goal is to develop a robust stock price prediction model tailored to the Indian stock market. By integrating sentiment analysis on financial news articles related to a particular stock with its historical stock price data, our approach promises to provide more accurate and reliable predictions. This project not only addresses the scarcity of localized research but also aims to enhance the decision-making tools available to investors and analysts within the Indian stock market. Through this innovative fusion of data sources, we hope to contribute significantly to the field of financial forecasting in India, offering a model that captures the intricate dynamics of one of the world's most vibrant and rapidly growing stock markets. In the next section, we will be concentrating on the scope of our work, focussing on its various aspects.

Chapter 4

Scope of the Work

In this chapter, we will discuss the scope of our work in brief. We are performing technical analysis, where we are relying on historical stock prices to predict the future stock price. Along with the historical values, we are integrating the sentiment scores of news articles. This integration aims to enhance the accuracy and reliability of stock price forecasts by leveraging both quantitative and qualitative data sources. Given historical stock price data and sentiment scores extracted from news articles, we aim to develop a predictive model for forecasting stock prices over a specified range of dates. The primary objective is to leverage the temporal patterns in historical stock prices along with sentiment analysis of news articles to predict future stock prices accurately. The project seeks to address the challenge of stock price prediction, which is inherently complex due to market volatility and the influence of external factors such as news sentiment. To achieve this, we will employ Natural Language Processing (NLP) techniques to perform sentiment analysis on the collected textual data, extracting sentiment scores that reflect the market's mood toward specific stocks. These sentiment scores will be integrated with the historical price data to create a comprehensive dataset. The integrated dataset will then be used to develop and train advanced machine learning models.

Given:

- Historical stock price data: $X = \{x_1, x_2, \ldots, x_n\}$ with n data points.
- Sentiment scores from news articles: $Y = \{y_1, y_2, \ldots, y_m\}$ with m data points.
- Target dates for prediction: $T = \{t_1, t_2, \ldots, t_k\}$

Let D be the merged dataset formed by matching the dates between the historical stock price data X and sentiment scores from news articles Y. The goal is to develop a predictive model that estimates the target stock prices T based on the merged dataset D. Mathematically, the problem can be formulated as finding a function f such that:

$$T = f(D)$$

where:

- $T = \{t_1, t_2, \ldots, t_k\}$ represents the target stock prices for a given range of dates.
- D is the merged dataset containing historical stock prices and sentiment scores, formed by matching the dates between X and Y.

Since our goal is to predict the exact future price of a stock, this is a regression task. The developed model will be evaluated using performance metrics like Mean Absolute Error (MAE), R-squared (R^2), and Mean Squared Error (MSE). These metrics quantify the prediction error between the estimated stock prices and the actual target prices over the given range of dates. The MAE measures the average magnitude of the errors between the predicted and actual values without considering their direction. It provides a straightforward measure of model prediction accuracy. The R^2 metric, also known as the coefficient of determination, is a statistical measure that explains the proportion of variance in the dependent variable that is predictable from the independent variables. The MSE measures the average of the squared differences between predicted and actual values, giving more weight to larger errors due to the squaring process. In the next chapter, we will be discussing our proposed solution in detail focussing on each of the steps involved.

Chapter 5

Proposed Solution

In this chapter, we will be discussing the intricacies of our proposed solution. The outline of the proposed solution is depicted in the flowchart in Fig. 5.1. The input to the problem is the news data gathered from various news websites. These news data are pre-processed to obtain the news articles about a particular concern. We perform sentiment analysis on these news articles to obtain a sentiment score of the emotional tone present in the text. The historical prices of the relevant stock are collected and clubbed with the sentiment scores of the news articles.

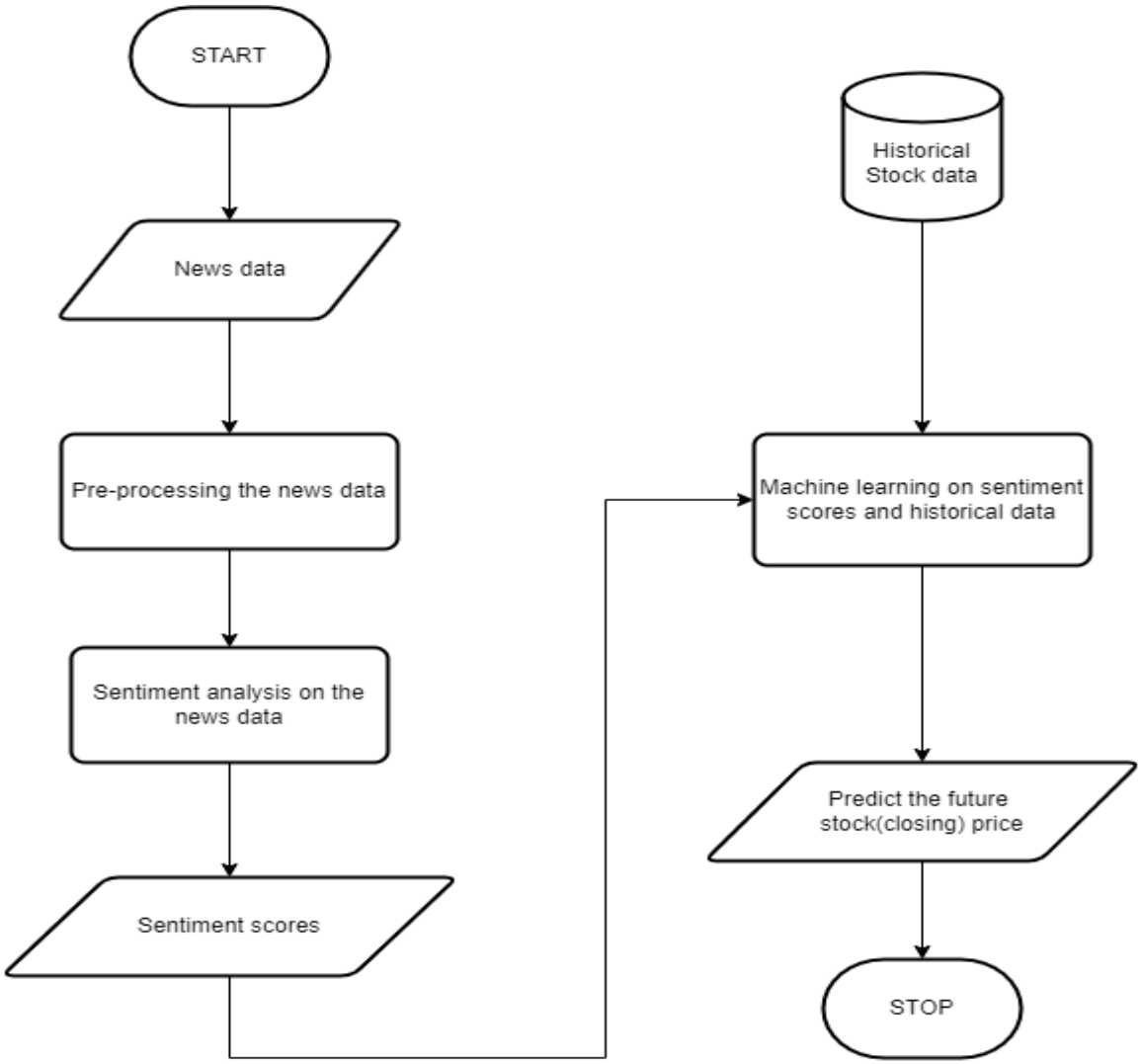


Fig. 5.1 - Flowchart for the proposed solution

The dataset obtained by this integration is then fed into a machine learning model. This model is trained to identify patterns and relationships within the data. The trained model is used to predict the future stock price based on the analyzed data. The process ends, having produced stock price predictions that incorporate both quantitative historical data and qualitative sentiment analysis from news articles. This comprehensive approach leverages both the numerical data and the textual data from news articles to

enhance the accuracy of stock market predictions, aiming to support more informed investment decisions. In the upcoming subsections, we will be focussing on each step of the proposed solution.

5.1 Pre-processing the data

In this subsection, we will discuss the step of pre-processing the data. The input to the process is the news data from multiple news websites. We have used economictimes.com[Fig. 5.2], moneycontrol.com, and timesofindia.com to extract the textual data that contains news of many relevant stocks. We collect news articles from these websites by the technique of web scraping. Web scraping is performed using a well-known tool, which is the BeautifulSoup library and requests package of Python.

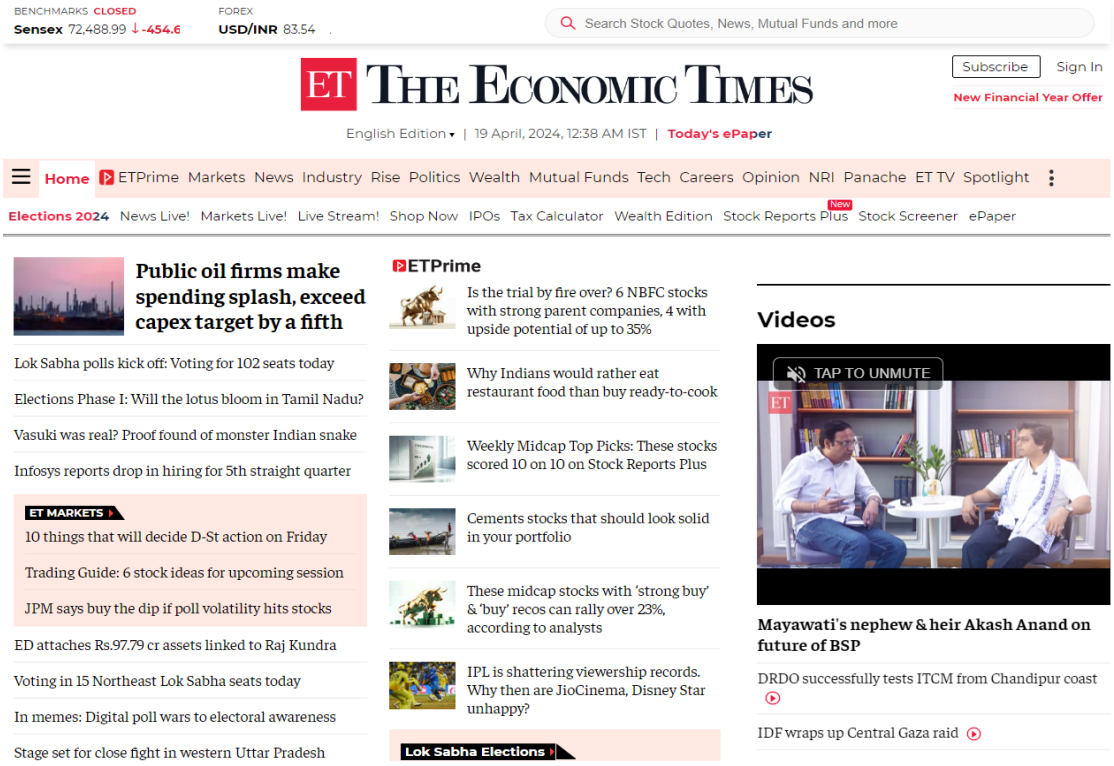
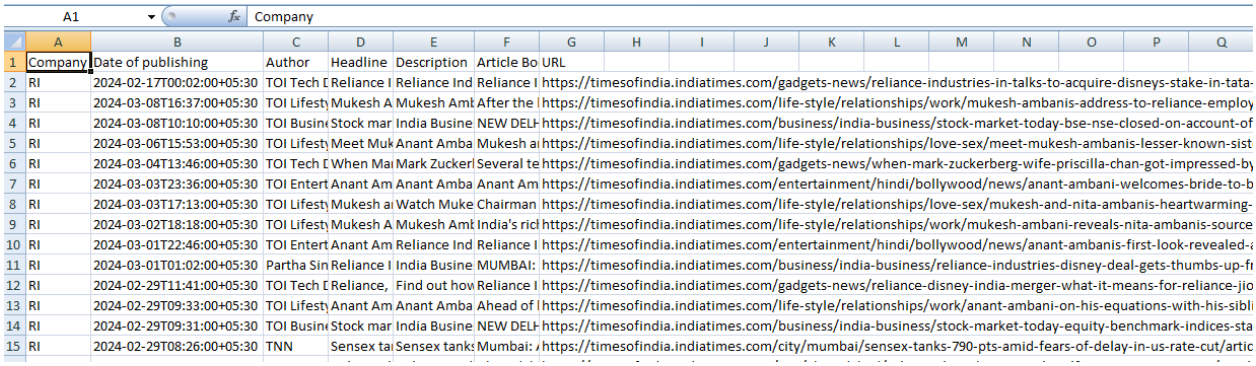


Fig. 5.2 - News data before pre-processing

The input to BeautifulSoup is the raw HTML content of the webpage and the parser to be used. The BeautifulSoup object represents the parsed document as a whole. Therefore, we can use the object to extract relevant information(viz. Date of publishing, Author, Headline, Description, Article Body, and URL) from the website about the concerned stock and store it in a CSV file[Fig. 5.3].



	Company	Date of publishing	Author	Headline	Description	Article Bo	URL
2	RI	2024-02-17T00:02:00+05:30	TOI Tech D	Reliance I	Reliance Ind	Reliance I	https://timesofindia.indiatimes.com/gadgets-news/reliance-industries-in-talks-to-acquire-disneys-stake-in-tata-
3	RI	2024-03-08T16:37:00+05:30	TOI Lifesty	Mukesh A	Mukesh Amb	After the	https://timesofindia.indiatimes.com/life-style/relationships/work/mukesh-ambanis-address-to-reliance-employ
4	RI	2024-03-08T10:10:00+05:30	TOI Busine	Stock mar	India Busine	NEW DELH	https://timesofindia.indiatimes.com/business/india-business/stock-market-today-bse-nse-closed-on-account-of
5	RI	2024-03-06T15:53:00+05:30	TOI Lifesty	Meet Muk	Anant Amba	Mukesh a	https://timesofindia.indiatimes.com/life-style/relationships/love-sex/meet-mukesh-ambanis-lesser-known-sist
6	RI	2024-03-04T13:46:00+05:30	TOI Tech D	When Mai	Mark Zucker	Several te	https://timesofindia.indiatimes.com/gadgets-news/when-mark-zuckerberg-wife-priscilla-chan-got-impressed-by
7	RI	2024-03-03T23:36:00+05:30	TOI Entert	Anant Am	Anant Amba	Anant Am	https://timesofindia.indiatimes.com/entertainment/hindi/bollywood/news/anant-ambani-welcomes-bride-to-b
8	RI	2024-03-03T17:13:00+05:30	TOI Lifesty	Mukesh a	Watch Muke	Chairman	https://timesofindia.indiatimes.com/life-style/relationships/love-sex/mukesh-and-nita-ambanis-heartwarming-
9	RI	2024-03-02T18:18:00+05:30	TOI Lifesty	Mukesh A	Mukesh Amb	India's rich	https://timesofindia.indiatimes.com/life-style/relationships/work/mukesh-ambani-reveals-nita-ambanis-source
10	RI	2024-03-01T22:46:00+05:30	TOI Entert	Anant Am	Reliance Ind	Reliance I	https://timesofindia.indiatimes.com/entertainment/hindi/bollywood/news/anant-ambanis-first-look-revealed-a
11	RI	2024-03-01T01:02:00+05:30	Partha Sin	Reliance I	India Busine	MUMBAI:	https://timesofindia.indiatimes.com/business/india-business/reliance-industries-disney-deal-gets-thumbs-up-fr
12	RI	2024-02-29T11:41:00+05:30	TOI Tech D	Reliance,	Find out how	Reliance I	https://timesofindia.indiatimes.com/gadgets-news/reliance-disney-india-merger-what-it-means-for-reliance-jio
13	RI	2024-02-29T09:33:00+05:30	TOI Lifesty	Anant Am	Anant Amba	Ahead of	https://timesofindia.indiatimes.com/life-style/relationships/work/anant-ambani-on-his-equations-with-his-sibli
14	RI	2024-02-29T09:31:00+05:30	TOI Busine	Stock mar	India Busine	NEW DELH	https://timesofindia.indiatimes.com/business/india-business/stock-market-today-equity-benchmark-indices-sta
15	RI	2024-02-29T08:26:00+05:30	TNN	Sensex ta	Sensex tanks	Mumbai: /	https://timesofindia.indiatimes.com/city/mumbai/sensex-tanks-790-pts-amid-fears-of-delay-in-us-rate-cut/artic

Fig. 5.3 - News data after pre-processing

5.2 Sentiment Analysis of the News Articles

In this subsection, we concentrate on sentiment analysis, a popular task in natural language processing(NLP) whose goal is to classify the text based on the mood or mentality expressed in the text, which can be positive, negative, or neutral. It is the process of analyzing text data to determine the sentiment or opinion expressed within it[7] Using NLP techniques, sentiment analysis algorithms classify text as positive, negative, or neutral, providing insights into the emotional tone of the text. This often involves techniques like bag-of-words or word embeddings to represent the text data numerically[6]. It helps to understand the opinions, emotions, and attitudes of individuals or groups towards a particular topic, product, service, or brand and provides valuable insights for businesses, marketers, and researchers to gauge public perception, identify trends, assess customer satisfaction, and make data-driven decisions.[6] Mathematically, sentiment analysis can be formulated as a classification problem where given a piece of text x, the goal is to predict its sentiment label y from a set of predefined sentiment categories. This can be represented as:

$$y = f(x)$$

where $f(x)$ is a function that maps the input text x to its corresponding sentiment label y. This function $f(x)$ can be implemented using various machine learning algorithms such as Support Vector Machines (SVM), Logistic Regression, or Transformers.

Bidirectional Encoder Representations from Transformers(BERT) is a transformer-based model that learns contextualized embeddings for words[6]. It considers the entire context of a word by considering both left and right contexts, resulting in embeddings that capture rich contextual information. BERT's architecture consists of only encoders and the input received is a sequence of tokens i.e. Token embeddings, Segment embeddings, and Positional embeddings[10]. These tokens are fed to the Transformer encoder, which are first embedded into vectors and then processed in the neural network. The output is a sequence of vectors, each corresponding to an input token, providing contextualized representations[11]. FinBERT is a specialized version of the BERT model, fine-tuned for financial texts. It is designed to understand the unique language and nuances found in financial documents, making it particularly well-suited for sentiment analysis tasks in this domain.[10,11] Since news articles related to stocks contain financial texts, FinBERT is used to perform the sentiment analysis of the news articles.

In the context of each statement from news collected from the news websites using web scraping, we intend to find positive(1), negative(2), or neutral(0) inclinations toward the concerned company using FinBERT. Depending on economic news, buyers buy or sell shares of the company. So

if we can extract sentiment from news articles then we can easily predict the future trend of a company in the share market based on that news article. A snapshot of the dataset is shown in Fig. 5.4. If there are multiple news articles available related to the particular stock on a trading day, then we take the average of the sentiment scores of all the news articles for that day. If there are no news articles present for a trading day, we set the sentiment score as '-1'[Fig. 5.5].

Date	Open	High	Low	Close	Adj Close	Volume	Company	Date of Pu	Author	Headline	Descriptic	Article Bo	URL	score
26-02-2024	2987.1	2989.05	2965	2974.65	2974.65	3756553	RI	2024-02-2	TOI News	What is Va	India New	NEW DELH	https://tir	2
26-02-2024	2987.1	2989.05	2965	2974.65	2974.65	3756553	RI	2024-02-2	Bloomber	Reliance I	India Busi	Walt Disn	https://tir	2
27-02-2024	2966.05	2999.9	2956.1	2971.3	2971.3	5413022	RI	2024-02-2	TNN	Reliance I	Reliance I	Ahmedab	https://tir	2
27-02-2024	2966.05	2999.9	2956.1	2971.3	2971.3	5413022	RI	2024-02-2	TNN	Reliance I	Reliance I	Ahmedab	https://tir	2
27-02-2024	2966.05	2999.9	2956.1	2971.3	2971.3	5413022	RI	2024-02-2	Reuters	Nita Amba	Nita Amba	Nita Amba	https://ec	1
28-02-2024	2966	2982.55	2900.35	2911.25	2911.25	4323975	RI	2024-02-2	Sagar Mal	Reliance I	Primark h	NEW DELH	https://ec	0
28-02-2024	2966	2982.55	2900.35	2911.25	2911.25	4323975	RI	2024-02-2	Javed Farc	Reliance,	Reliance I	Reliance I	https://ec	0
28-02-2024	2966	2982.55	2900.35	2911.25	2911.25	4323975	RI	2024-02-2	Akash Poc	RIL signs d	The deal,	Reliance I	https://ec	1
28-02-2024	2966	2982.55	2900.35	2911.25	2911.25	4323975	RI	2024-02-2	ET Online	Reliance,	Reliance I	Reliance I	https://ec	0
28-02-2024	2966	2982.55	2900.35	2911.25	2911.25	4323975	RI	2024-02-2	Navdeep S	Sensex, N	From the S	Tracking n	https://ec	1
28-02-2024	2966	2982.55	2900.35	2911.25	2911.25	4323975	RI	2024-02-2	Javed Farc	RIL-Disney	Nita Amba	Nita Amba	https://ec	2
29-02-2024	2930	2957.95	2909.05	2921.6	2921.6	11814488	RI	2024-02-2	TOI Tech D	Reliance,	Find out h	Reliance I	https://tir	2
29-02-2024	2930	2957.95	2909.05	2921.6	2921.6	11814488	RI	2024-02-2	TOI Lifesty	Anant Am	Anant Am	Ahead of	https://tir	2
29-02-2024	2930	2957.95	2909.05	2921.6	2921.6	11814488	RI	2024-02-2	TOI Busine	Stock mar	India Busi	NEW DELH	https://tir	1
29-02-2024	2930	2957.95	2909.05	2921.6	2921.6	11814488	RI	2024-02-2	TNN	Sensex tai	Sensex tai	Mumbai: /	https://tir	1
29-02-2024	2930	2957.95	2909.05	2921.6	2921.6	11814488	RI	2024-02-2	TOI Tech D	Reliance,	Find out h	Reliance I	https://tir	2
29-02-2024	2930	2957.95	2909.05	2921.6	2921.6	11814488	RI	2024-02-2	TOI Lifesty	Anant Am	Anant Am	Ahead of	https://tir	2
29-02-2024	2930	2957.95	2909.05	2921.6	2921.6	11814488	RI	2024-02-2	TOI Busine	Stock mar	India Busi	NEW DELH	https://tir	1
29-02-2024	2930	2957.95	2909.05	2921.6	2921.6	11814488	RI	2024-02-2	TNN	Sensex tai	Sensex tai	Mumbai: /	https://tir	1

Fig. 5.4 - A snapshot of the dataset with the sentiment scores of the news articles

5.3 Prediction of stock prices using Machine Learning

In this subsection, we discuss the prediction techniques of stock prices using machine learning. We have used the Bidirectional Long Short Term Memory(Bi-LSTM) algorithm. It is a type of recurrent neural network (RNN) that processes sequential data in both forward and backward directions and overcomes the limitations of traditional RNNs in capturing long-term dependencies in sequential data. It combines the power of LSTM with bidirectional processing, allowing the model to capture both the past and future context of the input sequence. The input to the machine learning model is the historical stock price combined with the sentiment scores[Fig. 5.5]. The historical prices are downloaded from the website of Yahoo Finance[12] which delivers hours of live, daily market coverage, with expert analysis and real-time market data

In the context of stock price prediction, which is a challenging time series forecasting problem, Bi-LSTMs offer significant advantages due to their ability to process information in both forward and backward directions[6]. Stock price prediction relies heavily on capturing temporal dependencies and patterns within historical price data. Traditional LSTMs are well-suited for this task as they can manage long-term dependencies and remember information over extended sequences. However, they process the input sequence in a single direction, typically from the past to the future. This unidirectional approach might overlook important patterns that could be captured by also considering the future context relative to a given point in time[6].

Date	Open	High	Low	Close	Adj Close	Volume	Company	Date of Pu	Author	Headline	Descriptic	Article Bo	URL	score	average_score
06-02-2024	2883.7	2883.7	2839.65	2855.6	2855.6	4523992	RI	2024-02-0	Nikhil Aga	Jio Financ	Jio Financ	Shares of	https://ec	0	0.5
07-02-2024	2871.85	2899	2858.5	2884.3	2884.3	4648284	RI	2024-02-0	Nikhil Aga	LIC owner	LIC's share	'India Bull	https://ec	1	1
08-02-2024	2900	2918.95	2855.05	2900.25	2900.25	7347317	RI	2024-02-0	Sangita M	Adani Pov	Lenders h	Mumbai: A	https://ec	2	2
09-02-2024	2908	2943.95	2901.9	2921.5	2921.5	6278399	RI	2024-02-0	Nikhil Aga	Chris Woc	Chris Woc	Jefferies'	https://ec	2	1.5
12-02-2024	2921.5	2922	2884.7	2904.7	2904.7	3337215	RI	2024-02-1	Arindam M	Governme	The gover	The gover	https://ec	1	1
13-02-2024	2911	2958	2908	2930.2	2930.2	3857797	RI	2024-02-1	PTI	Reliance I	India Busi	NEW DELH	https://tir	0	0.833333
14-02-2024	2915	2967.3	2915	2962.75	2962.75	3558944	RI	2024-02-1	Aseem Gu	Reliance I	India Busi	MUMBAI:	https://tir	2	1
15-02-2024	2966.7	2969.45	2933.05	2941.2	2941.2	5003391	RI	2024-02-1	Nikhil Aga	RIL turns t	Reliance I	With its m	https://ec	2	1.25
16-02-2024	2952.95	2954	2917.1	2921.15	2921.15	4883749	RI	2024-02-1	Navdeep S	Investors	Among th	Domestic	https://ec	0	1
19-02-2024	2924.1	2959	2907.05	2948	2948	3364914								-1	-1
20-02-2024	2950.05	2951	2923.6	2942.05	2942.05	3558748	RI	2024-02-2	ET Bureau	In a first, (	This is the	The Centr	https://ec	2	1.142857
21-02-2024	2948	2977.05	2915.1	2935.4	2935.4	6360146	RI	2024-02-2	Bloomber	BharatGPT	India Busi	A consorti	https://tir	2	1.5
22-02-2024	2936.3	2969.9	2916	2963.5	2963.5	9246346	RI	2024-02-2	Bloomber	RIL-backe	India Busi	A consorti	https://tir	2	1.4
23-02-2024	2979	2995.1	2966.7	2987.25	2987.25	7219292								-1	-1
26-02-2024	2987.1	2989.05	2965	2974.65	2974.65	3756553	RI	2024-02-2	TOI Lifest	Animals a	Anant Am	The year 2	https://tir	2	2
27-02-2024	2966.05	2999.9	2956.1	2971.3	2971.3	5413022	RI	2024-02-2	TNN	Reliance I	Reliance I	Ahmedab	https://tir	2	1.666667
28-02-2024	2966	2982.55	2900.35	2911.25	2911.25	4323975	RI	2024-02-2	Sagar Mal	Reliance I	Primark h	NEW DELH	https://ec	0	0.666667
29-02-2024	2930	2957.95	2909.05	2921.6	2921.6	11814488	RI	2024-02-2	TOI Tech I	Reliance,	Find out h	Reliance I	https://tir	2	1.5
01-03-2024	2927	3000	2925	2981.1	2981.1	5587933	RI	2024-03-0	TOI Entert	Anant Am	Reliance I	Reliance I	https://tir	2	0.7
04-03-2024	2980.95	3024.9	2974.45	3014.8	3014.8	5012210	RI	2024-03-0	TOI Tech I	When Mai	Mark Zuck	Several te	https://tir	2	2

Fig. 5.5 - Snapshot of the dataset containing historical prices and average sentiment scores.

Bi-LSTMs address this limitation by incorporating two LSTM layers: one that processes the input sequence from past to future, and another that processes it from future to past. This dual processing enables the model to understand the context more comprehensively, as it can leverage information from both directions[6]. In stock price prediction, this means a Bi-LSTM can better identify trends, reversals, and other patterns that might not be as evident when looking solely at past data. We can conclude that Bi-LSTMs enhance stock price prediction by leveraging their bidirectional processing capabilities, allowing the model to capture more nuanced patterns within the data. This results in potentially more accurate and robust predictions, making Bi-LSTMs a valuable tool in the domain of time series forecasting for financial markets.

5.4 Accuracy and Scoring Metrics

In this subsection, we will focus on the different accuracy and scoring metrics available for assessing the performance of the model. Evaluating a Bi-LSTM model for stock price prediction involves using various metrics to assess the accuracy and reliability of its forecasts. These metrics help determine how well the model captures the underlying patterns in the data and how effectively it predicts future stock prices. The most common evaluation metrics for time series forecasting include Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared(R^2) score, Root Mean Squared Error (RMSE), etc.

Mean Absolute Error (MAE) measures the average magnitude of errors between the predicted and actual stock prices without considering their direction. It is calculated as the average of the absolute differences between predicted and actual values. MAE provides a straightforward measure of prediction accuracy, where lower values indicate better performance. It is particularly useful when the cost of errors is linear, meaning that all errors are equally significant. Mathematically, it is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Mean Squared Error (MSE), on the other hand, squares the differences before averaging them. This penalizes larger errors more than smaller ones, making MSE sensitive to outliers. While this can be beneficial for highlighting significant prediction errors, it might not be ideal if the dataset contains

many outliers that could skew the evaluation. MSE is useful when large errors are particularly undesirable, and minimizing them is a priority. Mathematically, it is defined as:

$$\text{MSE} = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$$

R-squared(R^2) score, also known as the coefficient of determination, is another important metric used to evaluate the performance of a regression model, including Bi-LSTM models for stock price prediction. R-squared measures the proportion of the variance in the dependent variable (actual stock prices) that is predictable from the independent variables (historical stock prices and sentiment scores). Mathematically, it is defined as:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

In the above formulae, y_i is the actual value of the i^{th} observation, $\hat{y}_i$ is the predicted value of the i^{th} observation, and $\bar{y}$ is the mean value. In addition to these primary metrics, it is also important for us to consider the model's performance over different time horizons and across various market conditions. For instance, evaluating the model's predictive accuracy during volatile and stable market periods can provide insights into its robustness and reliability. Additionally, visualizing predicted versus actual stock prices through plots can help us qualitatively assess the model's performance and identify any systematic biases or trends in the predictions.

Therefore, we conclude that a comprehensive evaluation of a Bi-LSTM model for stock price prediction should involve a combination of these metrics to provide a well-rounded assessment of its accuracy, robustness, and reliability. By leveraging multiple evaluation criteria, one can gain deeper insights into the model's strengths and areas for improvement, finally leading to more informed and effective forecasting. In the next chapter, we will discuss the experimental findings obtained in our project.

Chapter 6

Experimental Findings

This chapter presents the experimental findings, showcasing the performance of our model across various metrics and highlighting the impact of incorporating sentiment analysis on prediction accuracy. The purpose of the project is to forecast the closing price by analyzing historical stock prices and incorporating sentiment analysis of news articles. The historical stock prices and news articles are utilized to develop a comprehensive understanding of market sentiment and its impact on stock movements. The dataset consists of two primary components: historical stock prices, providing a chronological record of past market behavior, and news articles sourced from relevant financial publications. The historical prices span across a period of 5 years from February 24, 2020, to March 07, 2024, containing data of 1003 trading days, downloaded from [12]. The news articles are also collected for the same period.

Sentiment analysis was carried out on the news articles to gauge the perception of the public towards the concerned stock. FinBERT was used to perform the sentiment analysis. The values corresponding to neutral, positive, and negative sentiments are 0, 1, and 2, respectively. For example, we consider the news published in the timesofindia.com on 11th September 2023, authored by PTI, "Markets News: MUMBAI: Equity benchmark indices continued their upward trend for the seventh consecutive day on Monday, as investors maintained their positive outlook". This news article has a sentiment score of 1, denoting a positive sentiment.

The Bi-LSTM model was executed 10 times and the scoring metrics were noted as shown in Table 6.1.

Execution No.	**R^2 score**	**MAE**	**MSE**
1	92.57%	48.06	3805.38
2	93.77%	44.13	3193.56
3	84.11%	74.18	8144.71
4	93.04%	46.41	3564.61
5	91.59%	53.55	4308.75
6	93.21%	46.92	3477.79
7	95.25%	37.97	1806.37
8	93.49%	45.41	3489.21
9	93.66%	44.27	3248.50
10	93.32%	45.18	3422.89

Table 6.1 - R^2 scores, MAE, and MSE values obtained from our model

Since we have not used any seed value in the model, the values of these metrics keep changing. Therefore, we will take the average of all the executions we have made and the following values were obtained as shown in Table 6.2.

Metric	**Average value obtained**
R^2 score	92.40%
Mean Absolute Error	48.60
Mean Squared Error	3846.17

Table 6.2 - Average values of the R^2 score, MAE and MSE

The graph plotting the predicted values and the actual values is visualized which is shown in Fig. 6.1.

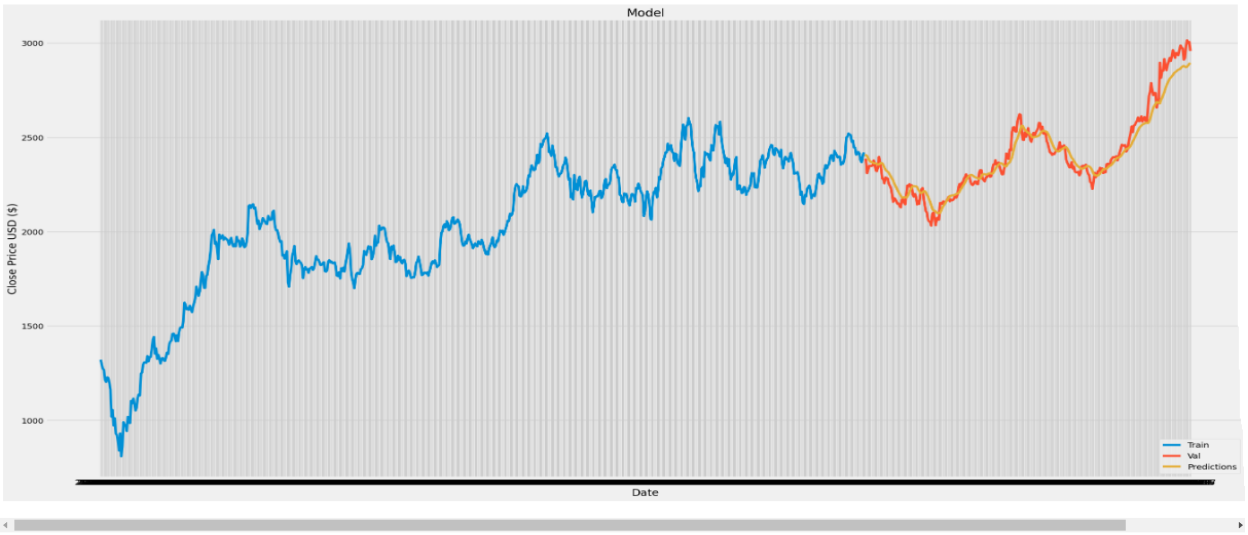


Fig 6.1 - Data visualization of the predicted and the actual values

Therefore, an R-squared value(average) of 92.40% in our results means that the model explains 92.40% of the variance in stock prices, indicating a strong fit and high predictive power. Thus the model effectively captures the underlying patterns in the data, resulting in predictions close to actual stock prices. However, it still leaves 7.60% of the variance unexplained, which could be due to random noise or external factors. An MAE of 48.60 means that, on average, the predictions made by the model are off by 48.60 units from the actual stock prices. An MSE of 3846.17 indicates that the average of the squared differences between the predicted and actual stock prices is 3846.17 units squared. It suggests that there are substantial discrepancies between the predicted and actual stock prices, particularly if these numbers are squared.

Chapter 7

Conclusion

The results confirm that sentiment analysis can be a valuable tool for understanding and predicting stock market trends in the Indian context, although normal LSTM can predict them. The positive, negative, and neutral sentiments influence the market considerably. However, the moderate correlation and R^2 values indicate that while sentiment is influential, it is not the sole determinant of market movements. This aligns with existing literature suggesting that stock markets are driven by a combination of factors, including macroeconomic indicators, company performance metrics, and global market trends. On a short-term basis, we can capture the trend of the market in a better way using the FinBERT sentiment analyzer.

7.1 Novelty

The novelty of this project lies in its comprehensive approach to sentiment analysis and its application of advanced NLP techniques. This contribution enhances the understanding of the complex relationship between market sentiment and stock market behavior, particularly in the context of the Indian market. This project leverages advanced Natural Language Processing (NLP) techniques, such as the FinBERT tool, to accurately interpret financial language nuances. We are considering the news articles for each day and computing their sentiment score using FinBERT. This approach not only enhances predictive modeling accuracy but also offers practical applications for investors, analysts, and policymakers, making sentiment analysis a valuable tool for understanding and anticipating market behavior.

7.2 Problems faced

Several challenges were encountered during the project, particularly in data collection and processing. One of the primary issues was acquiring and cleaning a comprehensive dataset that included financial news articles and historical market data. The unstructured nature of text data required significant preprocessing efforts to ensure accuracy in sentiment analysis. Another major challenge was the integration of sentiment scores with market data to establish meaningful correlations. The dynamic and multifaceted nature of stock market movements meant that sentiment alone could not fully explain price fluctuations. Moreover, the development of robust models that could account for these variances without overfitting required meticulous tuning and validation, making the analytical process more intricate.

7.3 Current Drawbacks

Despite the promising results, the study has several limitations. The reliance on sentiment scores derived from textual data may introduce bias, as not all market-relevant information is captured in news articles. The variation in sentiment sensitivity across different sectors can be attributed to the nature of the businesses and their market perception. Additionally, the model's explanatory power indicates that incorporating other variables, such as macroeconomic indicators, could improve predictive accuracy.

7.4 Future scope of improvement

As of now, we have worked on Reliance Industries, but other companies are to be analyzed. We did not complete the analysis of the market as a whole. We are trying to find a better representation of the news articles and incorporate macroeconomic factors and geopolitical events. Future research could explore more sophisticated machine learning models, to capture temporal dependencies in sentiment data. Additionally, expanding the dataset to include other sources of sentiment, such as investor forums, social media inputs, and expert analyses, could provide a more comprehensive view of market sentiment.

In conclusion, we can say that while sentiment analysis is a potent tool for analyzing the Indian stock market, it should be used in conjunction with other analytical methods to achieve more robust and accurate predictions. The insights gained from this study can inform better investment strategies and risk management practices, contributing to a more nuanced understanding of market dynamics.

References:

[1] Groww: https://groww.in/, last visited on 21/05/2024 10:00 pm

[2] Stock trend prediction using sentiment analysis, Qianyi Xiao and Baha Ihnaini, Published online 2023 Mar 20. doi: 10.7717/peerj-cs.1293

[3] Dancing with Trump in the Stock Market: A Deep Information Echoing Model, KUN YUAN, GUANNAN LIU, JUNJIE WU, Beihang University, HUI XIONG, Rutgers University, ACM Trans. Intell. Syst. Technol. 11, 5, Article 62 (July 2020), doi: https://doi.org/10.1145/3403578

[4] Stock Price Prediction Using Sentiment Analysis and Deep Learning for Indian Markets, Narayana Darapaneni, Northwestern University/Great Learning, Evanston, US, Anwesh Reddy Paduri, Himank Sharma, Milind Manjrekar, Nutan Hindlekar, Pranali Bhagat, Usha Aiyer, Yogesh Agarwal, Great Learning, Bangalore, India

[5] Investopedia: https://www.investopedia.com/, last visited on 01/06/2024 9:30 pm

[6] GeeksforGeeks: https://www.geeksforgeeks.org/, last visited on 01/06/2024 9:35pm

[7] Wikipedia : https://www.wikipedia.org/, last visited on 31/05/2024 8:00 pm

[8] Oracle NetSuite: https://www.netsuite.com/portal/resource/articles/accounting/financial-kpis-metrics.shtml, last visited on 13/05/2024 8:30 pm

[9] The Motley Fool: https://www.fool.com/investing/2018/03/21/9-essential-metrics-all-smart-investors-should-kno.aspx, last visited on 13/05/2024 8:30 pm

[10] Medium: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270, last visited on 16/05/2024 01:00 am

[11] TechTarget: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model#:~:text=BERT%2C%20which%20stands%20for%20Bidirectional,calculated%20based%20upon%20their%20connection, last visited on 16/05/2024 01:35 am

[12] Yahoo Finance: https://finance.yahoo.com/, last visited on 22/05/2024 12:00 pm

[13] Google Images: https://images.google.com/, last visited on 09/06/2024 04:25 pm

Appendix:

GitHub link: https://github.com/swarupcs/Stock-Prediction-Project