Stock Market Analysis Project

Sunanda K H
MFSDSAI
IIT Guwahati
Guwahati, India
h.sunanda@iitg.ac.in

***Abstract*— Predicting the future movements of the stock market is a complex task that has captivated investors and financial experts for decades. Traditional approaches to stock market prediction have relied primarily on historical financial data and technical analysis, often overlooking the significant influence of sentiment on market movements. Recent research has highlighted the potential of sentiment analysis to capture the emotional tone of public opinion and incorporate it into stock market prediction models. This project delves into the integration of sentiment analysis from news headlines into stock market prediction models, investigating the potential of sentiment analysis to enhance the accuracy and reliability of predicting future stock prices. To achieve this objective, the project tries to employ machine learning techniques, including ARIMA, Prophet, LSTM, Random Forest Regressor, and KNN, to develop prediction models that incorporate both historical stock prices and sentiment data extracted from news headlines. The sentiment analysis methodology utilizes natural language processing (NLP) techniques and sentiment lexicons to identify and extract sentiment from news headlines effectively. The performance of the sentiment-based prediction models is evaluated using metrics such as root mean squared error (RMSE), comparing their effectiveness to traditional models based solely on historical data. The results demonstrate that incorporating sentiment analysis into prediction models can improve their accuracy, suggesting that sentiment plays a significant role in influencing stock market movements. The project tries to contribute to a deeper understanding of the relationship between sentiment and stock price movements, providing valuable insights for enhancing stock market prediction models. By incorporating sentiment analysis from news headlines, investors and financial institutions can gain a more comprehensive understanding of market dynamics and make informed investment decisions.**

Keywords*— *Stock Market Prediction, Sentiment Analysis, Machine Learning, Random Forest Regressor, k-Nearest Neighbors (kNN), Long Short-Term Memory (LSTM), Auto-Regressive Integrated Moving Average (ARIMA), Prophet, Root Mean Squared Error (RMSE), News Headlines.

I. INTRODUCTION

Investing in the stock market involves understanding and predicting how stock prices will change. Traditionally, this prediction is based on economic data and the financial performance of companies. However with rise of social media and online news platforms has introduced new ways to find public opinion and reaction of public to the news of various stocks, which can also influence stock markets. This project tries to show the impact of these digital sentiments on stock prices, particularly focusing on HDFC Bank, one of the leading banks in India. The aim of this project is to determine whether analyzing sentiments from news headlines can help predict the future stock prices of HDFC Bank. This approach is based on the idea that public sentiment, as expressed through social media and news platforms, can provide insights into market trends. To conduct the analysis, there was the collection of news headlines related to HDFC Bank. Then sentiment analysis techniques was applied to this data. Sentiment analysis is a method used in natural language processing that identifies whether the text is positive, negative, or neutral. By quantifying the sentiments expressed in tweets and news articles, the aim was to uncover patterns and correlations with HDFC Bank's stock price movements. Few Machine Learning Models were used to predict stock prices based on the sentiment data collected. These models were chosen for their effectiveness in handling complex datasets and providing accurate predictions. By integrating sentiment data from digital platforms into traditional stock market analysis, we aim to enhance the accuracy of stock price predictions. In summary, this project is focused on understanding the relationship between digital sentiment and stock market behavior, done on the data of HDFC Bank.

II. MOTIVATION

The project involved collecting a dataset of news mentioning HDFC Bank and news headlines from The Economic Times over a specified period. Sentiment scores were calculated using the nltk library, which provided a

quantitative measure of sentiment polarity in the data. For the predictive analysis, there were a few Machine Learning Models that were employed. The models were trained and tested on historical stock price data from HDFC Bank, augmented with the derived sentiment scores.

Accurately predicting stock market trends is a complex yet highly sought-after endeavour, as it holds significant potential for both individual investors and financial institutions. While traditional approaches to stock market prediction rely on historical financial data and technical analysis, there is growing recognition of the role that sentiment plays in influencing market movements. News headlines, in particular, serve as a rich source of information about public perception and sentiment towards specific companies or the overall market. This project aims to investigate the potential of incorporating sentiment analysis from news headlines into stock market prediction models. Sentiment analysis has demonstrated its ability to capture the emotional tone of text, which can be indicative of public opinion and overall market sentiment. Studies have shown that news sentiment can have a significant impact on stock prices, suggesting that incorporating sentiment analysis into prediction models could enhance their accuracy. News headlines provide a continuous stream of up-to-date information about companies and the market, making them a valuable source of data for sentiment analysis. Additionally, the availability of real-time news feeds allows for the development of prediction models that can react swiftly to changing sentiment. Machine learning techniques, such as ARIMA, Prophet, LSTM, Random Forest Regressor, and KNN, have proven to be effective in modelling complex relationships between variables. These models can be trained on historical stock prices and sentiment data to identify patterns and make predictions about future price movements.

III. Problem Statement

Accurate stock market prediction remains a challenging task due to the complex interplay of various factors, including historical financial data, technical indicators, and market sentiment. Traditional stock market prediction models primarily rely on historical financial data and technical analysis, often overlooking the significant influence of sentiment on market movements. Sentiment analysis, particularly sentiment extracted from news headlines, offers a promising approach to capture the emotional tone of public opinion and incorporate it into stock market prediction models. This project aims to address if can sentiment analysis from news headlines be effectively integrated into stock market prediction models to enhance their accuracy and reliability. Specific challenges include Developing a robust methodology to effectively identify and extract sentiment from news headlines using natural language processing (NLP) techniques and sentiment lexicons, Analyzing the correlation between sentiment scores extracted from news headlines and stock price movements over various time periods, Integrating sentiment data into machine learning models, such as ARIMA, Prophet, LSTM, Random Forest Regressor, and KNN, to develop effective prediction models that consider both historical financial data and sentiment analysis, Assessing the accuracy and reliability of sentiment-based prediction models compared to traditional models that rely solely on historical data.

IV. Architectural Details

This research employs a structured approach to predict stock prices by analyzing online sentiments and applying machine learning techniques. The architecture is designed to process, analyze, and predict stock market trends for HDFC Bank using data online news sources.

A. Data Collection

There are two primary sources: Yahoo Finance and online news websites such as The Economic Times. For Twitter, there is a method that connects to the Twitter API and searches for all recent tweets that mention HDFC Bank. This has not been implemented due to the limitation of the API. For online news, a web scraping tool using BeautifulSoup was made to scrape the headlines of reliable financial news websites such as The Economic Times and collects headlines that are specifically about HDFC Bank and the Date Stamp.

B. Sentiment Analysis Process

This involves examining the words in each tweet or headline and deciding whether the overall sentiment is positive, negative, or neutral. Each piece of text is given a sentiment score based on its content. All the scores obtained for the news headlines in the day are averaged out to predict the final score and the corresponding Sentiment.

C. Machine Learning Models for Prediction

There was a use of five different models: Long Short Term Memory (LSTM), Auto Regressive Integrated Moving Average (ARIMA), Prophet, Random Forest Regressor and k-Nearest Neighbors (kNN). Random Forest is great for handling complex data and making accurate predictions, while kNN is simpler and good at finding patterns in data. These models are trained using the historical stock data. The models learn to see patterns in how the stock price changes.

D. Training and Evaluation of the Models

The models undergo a rigorous training phase using historical data, enabling them to discern associations

between sentiment trends and stock price movements of HDFC Bank. The dataset is bifurcated into training and testing segments, the latter being crucial to evaluate the predictive accuracy of the models. Root Mean Square Error (RMSE) is employed as the metric to gauge prediction accuracy, with lower RMSE values indicating higher precision in forecasts.

V. Methodology and Experiments

The methodology of this project revolves around integrating sentiment analysis with machine learning to predict stock market prices, particularly focusing on HDFC Bank. The process involves several key stages. Relevant financial news headlines are scraped from prominent financial news websites. Historical stock price data for HDFC Bank is sourced from reliable financial databases such as Yahoo Finance. The collected new headlines undergo preprocessing, which includes cleaning and normalizing the text. Using NLP tools like VADER, each piece of text is analyzed to assign a sentiment score, reflecting its positive, negative, or neutral tone. The models, *Long Short Term Memory (LSTM), Auto Regressive Integrated Moving Average (ARIMA), Prophet, Random Forest Regressor and k-Nearest Neighbors (kNN),* are employed for the prediction task. These models are chosen for their effectiveness in regression. The models are trained on a split dataset, comprising a majority portion of the historical data. The remaining portion of the data is used to validate the models. This involves testing the models' ability to predict stock prices accurately and comparing the predictions against actual stock market data. RMSE is used as the primary metric to evaluate the models' performance. It measures the average magnitude of the prediction errors, providing insights into the accuracy of the predictions.

TABLE I. RMSE

S.No	Model	RMSE
1	ARIMA	20.51007468005706
2	kNN	1.0425220598989475
3	LSTM	2.344422824595805
4	Prophet	23.89111928788275
5	RFR	0.5228798920630974

VI. Results and Conclusion

Random Forest Model eexhibited a high level of accuracy in predicting HDFC Bank's stock prices. The RMSE for this model was found to be 0.523, indicating a low prediction error. The inclusion of sentiment scores from news sources significantly influenced the stock price predictions. Models incorporating sentiment data were able to better capture the fluctuations in HDFC Bank's stock prices, as evidenced by a decrease in RMSE values compared to models without sentiment data. The project also highlights the potential for using real-time data from social media and online news to make timely predictions, offering a significant advantage in the fast-paced financial market. The research opens avenues for further exploration, such as integrating more diverse data sources, applying advanced machine learning and deep learning techniques, and exploring real-time predictive systems.