Stock Recommendation using Similarity and ML Techniques

Aruj Deshwal
IIITD, Delhi
New Delhi

Saatvik Bhatnagar
IIITD, Delhi
New Delhi

Ishaan Dayal
IIITD, Delhi
New Delhi

Hrishav Basu
IIITD, Delhi
New Delhi

Ipsita R
IIITD, Delhi
New Delhi

Hemang Dahiya
IIITD, Delhi
New Delhi

***Abstract*—In this paper we are going to use machine learning techniques that can effectively utilize a diverse range of data sources to provide more accurate predictions. This research paper aims to incorporate sequential models along with dynamic time warping and clustering to identify similar time series stock data and leverage financial news articles and social media data for sentiment analysis to make more accurate inferences about future stock prices.**

I. Introduction

For this project, we will use the price data of all the companies listed on the NYSE. We extract the price data for each day from NYSE's archive, it consists of the previous day's closing price, the current day's opening price, the current day's lowest price, the last traded price for the current day, the closing price for the current day, the average price of the stock, total turnover, number of trades, deliverable quantity and delivery percentage. This data is available for each day over a period of several years. For textual queries, we obtain data by using a web scraping tool to fetch news articles from reputed news websites like Bloomberg, CNN, Yahoo Finance, and Reuters.

We preprocess the data, by normalizing and sanitizing it and storing it in appropriate data structures for efficient queries.

We plan on combining the result of text-based and historical price-based predictions to arrive at a more accurate and efficient prediction. Historical price-based prediction is done using Dynamic Time Warping to measure the similarity of time series price data, by analyzing the DTW of stock prices, this helps in identifying similar patterns even if time series data is out of sync. DTW is widely used to predict the price but we use other models like Neural Networks and deep learning techniques like LSTM and choose the model on the basis of the metrics mentioned below

To evaluate the results of the models we will take our predictions of the stock prices and assess it in terms of the conventional classification metrics such as accuracy, precision, and recall for predicting the direction of stock price which is formulated in the form of a binary problem. We also take a look at regression metrics such as Mean Absolute Error(MAE), Mean Square Error(MSE), and Root Mean Square Error(RMSE). Furthermore, we analyze the profit made by our system to check the viability of our approach.

II. Problem Statement

The stock market is a highly unpredictable domain, and the ability to accurately predict future stock prices is highly valued by investors. However, predicting stock prices is a challenging task, and traditional methods are often not sufficient to make accurate predictions. The goal of this research paper is to develop an advanced machine learning-based approach that can utilize a diverse range of data sources to provide more accurate predictions. The paper proposes to incorporate LSTM, ANN, and SVM techniques, along with dynamic time warping and clustering, to identify similar time-series stock data and leverage financial news articles and social media data for sentiment analysis to make more accurate inferences about future stock prices. The research aims to evaluate the performance of this approach on historical price data of companies listed on the NYSE, and to test the viability of the approach by analyzing the profit made by the system.

III. Dataset

The dataset used in this research paper has been sourced from BBC News and the Yahoo Finance library. It comprises 10 years of historical stock price data, containing approximately 2500 rows with 6 columns in each row. The 6 values included in each row are the volume, open price, closing price, high, low, and adjusted closing price. To evaluate our model's performance, we have split the dataset into training and test sets, with a ratio of 90:10, respectively. Our model processes 30 days of stock data and predicts the prices of the 31st day. This process is then repeated for the subsequent interval. To establish a baseline for our results, we have tested our model on stock price data for five companies: Goldman Sachs, Wells Fargo and Co, Meta, and Roku.

IV. Related works

Yoo et al. [1] explores the efficacy of various methods of stock price prediction such as statistical time series prediction models such as Box-Jenkins univariate models which give 60% accuracy. It shows that neural networks are able to predict stock prices more accurately and robustly which shows correct stock movement 92% of the time. It also says that a very

common problem is overfitting. It also mentions that since SVM's are similar to solving linearly constrained quadratic programming problems they are resistant to the overtraining problems and solutions and relatively globally optimal.

Muhammad et al. [2] examines the predictive power of fundamental analysis on stock returns in non-financial sectors listed on the Karachi Stock Exchange in Pakistan. The study is based on four variables from five different areas of profitability, liquidity, solvency, and market-based ratios. The results indicate a significant positive impact of some fundamental indicators on stock return prediction, with market-based ratios and profitability having a greater impact than other ratios. However, some other variables such as current ratio, leverage, and earnings per share have shown insignificant coefficient values. The study suggests that fundamental analysis alone may not be enough for successful forecasting decisions, and both fundamental and technical analysis may be used for better prediction of future stock returns.

The research provides insights into the validity and forecasting quality of fundamental analysis in the Pakistani market, which may be helpful for investors, stock exchange dealers, and brokers to forecast stock prices and get excess returns. The study's limitations include the small sample size and the limited time horizon, which may be extended in future research. Overall, the research provides empirical evidence of the predictive power of fundamental analysis in the Pakistani market, which may be used as a basis for further research and investment decision-making.

Bemdt et al. [3] discusses the challenges of extracting valuable information from large databases, with a focus on the detection of patterns in time-series data. The text highlights the fact that most business transactions and scientific data are recorded by computers, leading to an exponential growth in the amount of data available. This creates a challenge for knowledge discovery research to develop methods for extracting valuable information from these databases.

The text goes on to discuss the different types of data, namely, categorical and continuous data, and how the addition of a temporal dimension creates time-series data. The detection of patterns in time-series data requires an approximate or "fuzzy" matching process. The example of lynx and snowshoe hare populations is used to illustrate this point.

The text then discusses the problem of pattern detection in time-series data and how humans are good at visually detecting patterns, but programming machines to do the same is a difficult problem. The technique of dynamic time warping is introduced as a way to align time series and a specific word template so that some distance measure is minimized. Overall, the text provides a good overview of the challenges of knowledge discovery from large databases and the need to develop new methods for analyzing time-series data.

Nikfarjam et al. [4] addresses the potential use of text mining techniques as a means of predicting stock market movements. In this study, the author examines the relationship between financial news and stock market trends and uses his text-mining techniques to extract important features such as vocabulary, sentiment, and subject matter from financial news articles. Various data mining algorithms such as decision trees, neural networks, and SVM are then applied to these extracted features to classify and predict stock prices. The research paper concludes that text mining techniques are a promising data source for predicting stock market trends. Additionally, the research highlights the importance of combining these text-mining techniques with machine learning algorithms to create more accurate and accurate predictive models. Overall, this research paper provides a comprehensive overview of the application of text mining techniques in stock market forecasting and highlights the potential benefits of integrating these techniques with existing forecasting models. The results of this study could have a significant impact on the trading and investment industry as it opens new avenues for predicting and forecasting stock market trends.

In "Stock Market Prediction Using LSTM Technique", Talati et al. [5] explores the use of Long Short-Term Memory (LSTM) neural networks for stock market prediction. The study utilizes historical stock prices and financial indicators to predict future prices and returns, with a particular focus on the Indian stock market. The authors provide a comprehensive literature review of prior studies on stock market prediction, highlighting the limitations of traditional statistical models in capturing the complex relationships between stock prices and financial indicators. They then introduce the concept of LSTM networks, which have been shown to perform well in time series prediction tasks. The study demonstrates the effectiveness of the LSTM model in predicting stock prices and returns, achieving higher accuracy than traditional statistical models. The authors also conduct an extensive analysis of the model's performance and provide insights into the importance of various financial indicators in predicting stock prices. Overall, the paper offers a valuable contribution to the literature on stock market prediction, highlighting the potential of LSTM networks in this domain and providing insights into the key factors that impact stock prices.

Khedr et al. [6] presents in the paper a new approach to predicting stock market trends by combining data mining techniques and news sentiment analysis. The proposed model offers a new method of analyzing stock market data that could improve the accuracy of predictions and has the potential to provide valuable insights for investors and financial analysts.

The paper's findings show that the proposed model outperforms other machine learning models, achieving an accuracy of 72.8%. This suggests that the model could be an effective tool for predicting stock market behavior. However, the paper also acknowledges the limitations of the model, such as the challenge of dealing with sudden changes in the stock market and the need for real-time news updates. These limitations provide a basis for future research to improve the model and

address the challenges of predicting stock market behavior.

Lee et al. [7] in their paper explores the impact of earnings news on small traders' stock trading behavior. The study focuses on intraday data from the NASDAQ stock exchange, using a sample of 97 firms with 1,346 earnings announcements over a period of six months. The study analyzes trading behavior of small traders, defined as those who hold less than 100 shares.

The methodology of the study includes event study analysis and regression analysis, and the results suggest that small traders respond to earnings news, with a significant increase in trading volume and volatility during the announcement window. The study finds that small traders' response to earnings news is not influenced by previous stock returns or volatility. The results also indicate that small traders have a tendency to trade in the same direction as the initial price movement following an earnings announcement.

Overall, the study contributes to the literature on behavioral finance by providing empirical evidence on the impact of earnings news on small traders' trading behavior. The study's methodology is robust and its results offer insights into the trading behavior of small traders in response to earnings news.

Borovkova et al. [8] in the research paper propose an approach to classify stock market movements using an ensemble of LSTM neural networks. The study employs high-frequency data from the NASDAQ stock exchange and uses an ensemble of five LSTMs to classify stock movements as either positive or negative.

The methodology of the study involves preprocessing the data and training the LSTMs on a sliding window of time series data. The results show that the ensemble of LSTMs outperforms other classification methods, with an accuracy of 65% on unseen data. The study also examines the importance of different technical indicators in predicting stock movements and finds that a combination of indicators produces the best results.

Overall, the study contributes to the literature on stock market prediction by offering a novel approach that leverages the power of LSTMs and ensemble learning. The methodology of the study is well-defined and the results are promising, indicating the potential for this approach to improve stock market prediction accuracy.

V. Baseline Results

A. Prediction of Stock Prices

The function run_model_company() is used for stock price prediction using the LSTM algorithm. It takes an input index of a company and extracts the stock data associated with that index. The stock data is first plotted for visualization purposes, and then preprocessed by scaling it to the range of 0 to 1 using MinMaxScaler. The scaled data is then divided into training and testing datasets in the ratio of 90:10. The training data is prepared by creating sliding windows of size 30, with

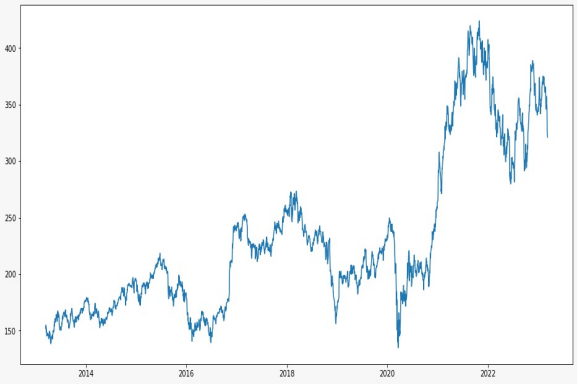


Fig. 1. Goldman Sachs

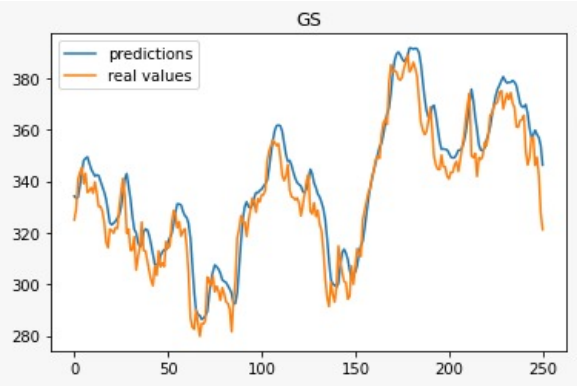


Fig. 2. Goldman Sachs

the corresponding output as a single value (which is the 31st day's value). A 2-layered LSTM neural network architecture is created with each layer consisting of 128 nodes, followed by two Dense layers with 32 and 1 nodes, respectively. The model is compiled with Adam optimizer and Mean Squared Error loss function, and then trained on the training data with a batch size of 1 and for a single epoch.

After training the model, the testing data is prepared and sliding windows of size 30 are created similar to the training

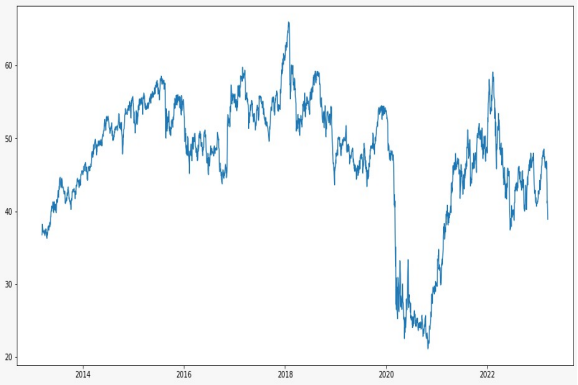


Fig. 3. Wells Fargo & Co

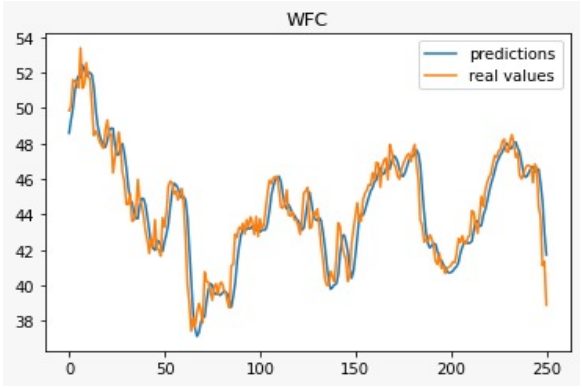


Fig. 4. Wells Fargo & Co

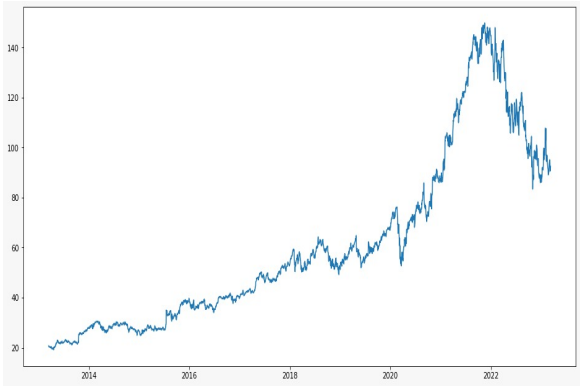


Fig. 5. Google

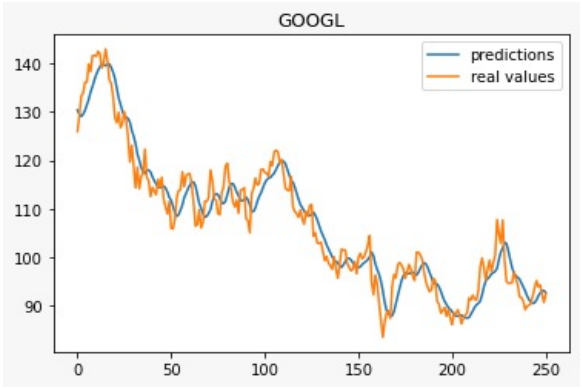


Fig. 6. Google

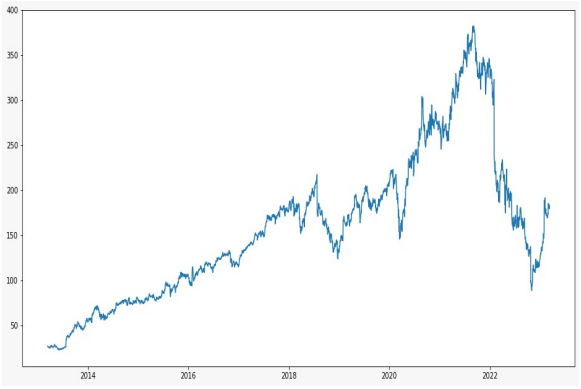


Fig. 7. META

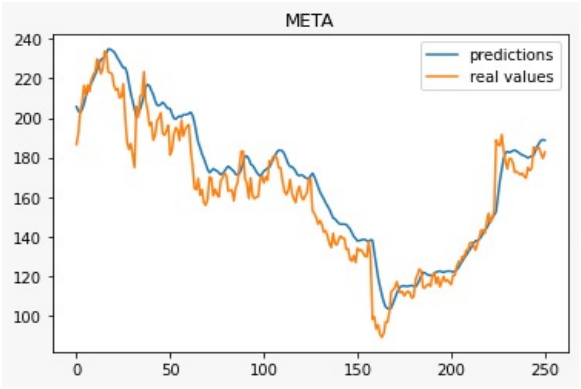


Fig. 8. META

data. The trained model is used to predict the stock prices for the testing dataset. The predicted values are then unscaled using the MinMaxScaler object used earlier. The Root Mean Squared Error and Mean Absolute Error of the predicted values compared to the actual values are calculated. Finally, the predicted and actual stock prices are plotted using Matplotlib.

In summary, the function predicts stock prices of a given company using LSTM neural networks and sliding window technique. It preprocesses the data using MinMaxScaler and divides it into training and testing datasets. The model is trained on the training data and then tested on the testing data. The accuracy of the model is evaluated using Root Mean Squared Error and Mean Absolute Error, and the predicted and actual stock prices are plotted for visualization purposes. The function can be used to predict the stock prices of multiple companies by looping through each company's index and calling the function.

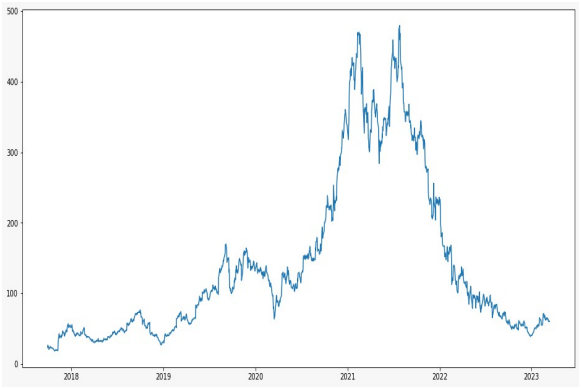


Fig. 9. Apple

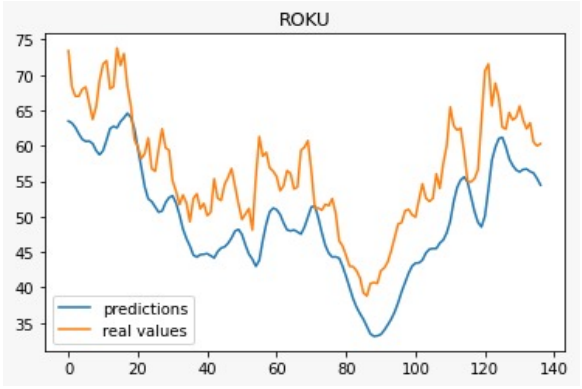


Fig. 10. ROKU

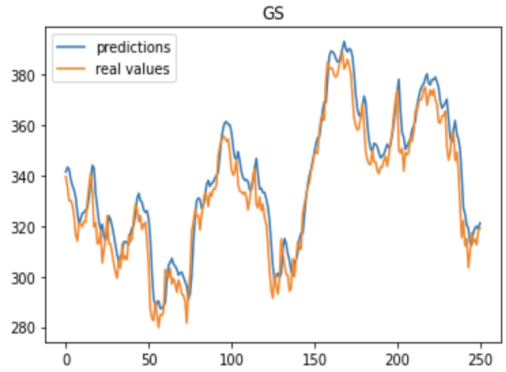


Fig. 11. Goldman Sachs updated

Company	Rmse Error	Mean absolute error
GS	9.901540149643976	7.839751224593813
WFC	1.118736376761082	0.8634557382044089
Google	12.406087325676825	3.1375015076413098
META	12.406087325676825	9.574828767206565
ROKU	7.822217043889508	6.962498797117359

Table 1. RMSE Error, Mean absolute error.

B. Updated Stock Prediction(Mid Project)

Several different model architectures were tested out for the LSTM models using different number of layers and different size of hidden layers but no significant improvement was found. Different values of hyperparameters were also tested such as the window size of number of days, batch size, number of epochs and learning rate were also experimentally tested out. The best results were seen on lr = 3e-4, batch size = 8, number of epochs = 3 and number of days = 30. Better results were seen on 4 out of 5 companies for both metrics of Root mean squared error and Mean Absolute error. Several different models were tried out and empirically the best results were seen on a GRU architecture with the 4 GRU units of hidden size 32 stacked together which were fed into a dense layer of size 1 which made the prediction of the closing price.

The Root Mean Square Error and the Mean Absolute Error values can be found in Table 2

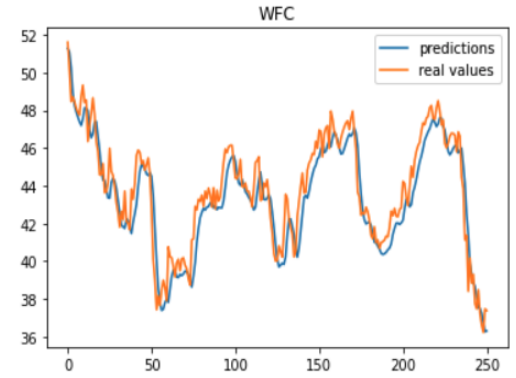


Fig. 12. Wells Fargo updated

Fig. 13. Google updated

Company	Rmse Error	Mean absolute error
GS	8.780	7.011
WFC	1.159	0.917
Google	3.639	2.936
META	8.912	6.603
ROKU	4.596	3.780

Table 2. Updated RMSE Error, Mean absolute error

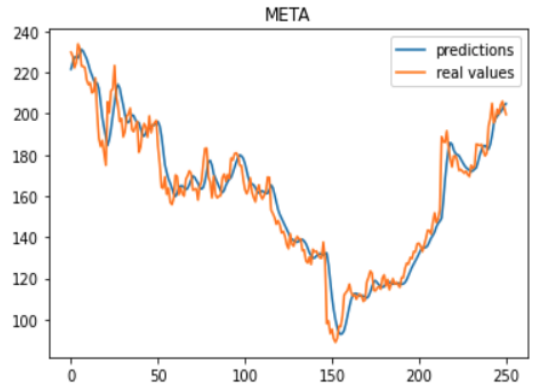


Fig. 14. Meta updated

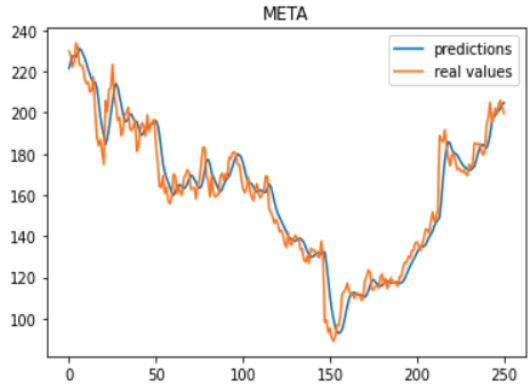


Fig. 15. Meta updated

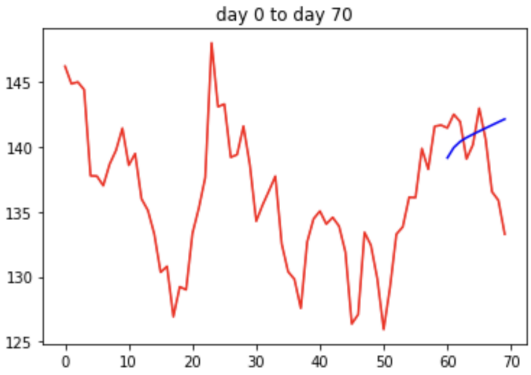


Fig. 16. Google 10 day prediction (day 0)

C. Updated Stock Prediction(Mid Project)

Some experiments were also done into predicted the stock closing price for google stock several days in the future but good results were not found due to the volatility of the stock market and we see that it predicts similar patterns across multiple months and that the model is overfitting. Fig 16 and 17 show two examples of such predictions. In the figure the red line show the real values and the blue line shows the predicted value across 10 days using only prior data and the predictions made the model itself as inputs.

D. Sentiment Analysis

The given Python function, get_sentimental_analysis(), extracts the text content from a web page, performs text cleaning, and uses TextBlob library to perform sentiment analysis on the extracted text. The function takes a single argument, url, which is the web address of the page to be analyzed.

The function begins by using the requests module to retrieve the HTML content of the web page. The BeautifulSoup module is then used to extract the text content of the article from the HTML. The extracted text is converted to lowercase and tokenized using the nltk.word_tokenize() function. Stop words and punctuation are removed from the tokenized text using the nltk.corpus.stopwords.words() function and the punctuation

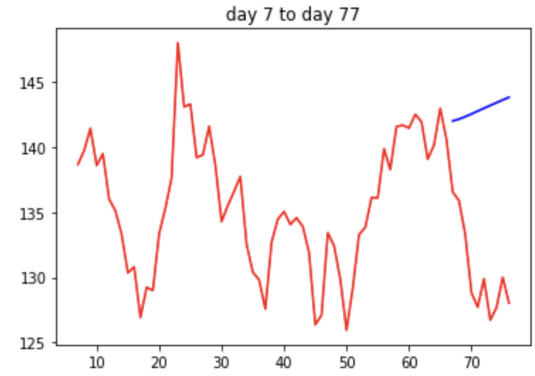


Fig. 17. Google 10 day prediction (day 7)

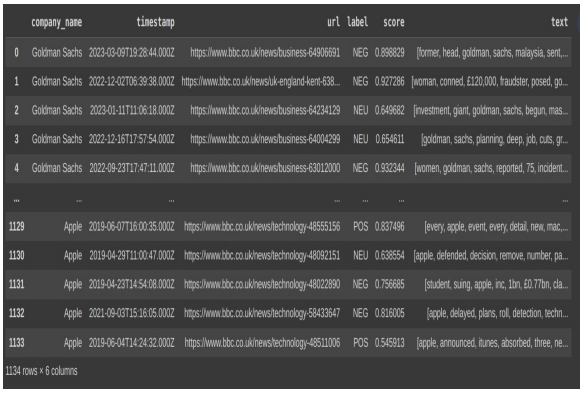
module, respectively. The cleaned text is then concatenated back into a single string and stored in the tokenized variable.

The finite automata/bertweet-base-sentiment-analysis model is a pre-trained natural language processing (NLP) model that is specifically designed for sentiment analysis. It is based on the BERTweet model, which is a variant of the popular BERT (Bidirectional Encoder Representations from Transformers) model, pre-trained on a large corpus of Twitter data.

The finite automata/bertweet-base-sentiment-analysis model is a pre-trained NLP model specifically designed for sentiment analysis based on the BERTweet model. It is fine-tuned using labeled data to classify text as positive, negative, or neutral. The model uses attention mechanisms and deep learning to analyze text and extract relevant features for sentiment determination. Self-attention mechanism allows focusing on different parts of the input text, and a multi-layer feed-forward neural network maps the input to an output sentiment class. This model is effective for analyzing sentiment in social media contexts, such as Twitter, and can handle complex language and nuanced expressions.

Next, the TextBlob() function from the TextBlob library is used to perform sentiment analysis on the cleaned text. The sentiment analysis returns two values, polarity and subjectivity, which indicate the sentiment and subjectivity of the text, respectively. The polarity score is a float value between -1 and 1, where negative values indicate negative sentiment, positive values indicate positive sentiment, and 0 indicates neutral sentiment. Subjectivity lies between [0,1]. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information.

Finally, the function returns a tuple containing the sentiment and subjectivity scores, as well as the timestamp and the original text content of the article. If the timestamp cannot be extracted from the HTML, an empty string is returned for the timestamp. The function can be called with the URL of any web page to perform sentiment analysis on its text content.



	company_name	timestamp	url	label	score	text
0	Goldman Sachs	2023-03-09T19:28:44.000Z	https://www.bbc.co.uk/news/business-64906691	NEG	0.898629	[former, head, goldman, sachs, malaysia, sent,...
1	Goldman Sachs	2022-12-02T06:39:38.000Z	https://www.bbc.co.uk/news/uk-england-kent-638...	NEG	0.927286	[woman, conned, £120,000, fraudster, posed, go...
2	Goldman Sachs	2023-01-11T11:06:18.000Z	https://www.bbc.co.uk/news/business-64234129	NEU	0.649682	[investment, giant, goldman, sachs, begun, mas...
3	Goldman Sachs	2022-12-16T17:57:54.000Z	https://www.bbc.co.uk/news/business-64004299	NEU	0.654611	[goldman, sachs, planning, deep, job, cuts, gr...
4	Goldman Sachs	2022-09-23T17:47:11.000Z	https://www.bbc.co.uk/news/business-63012000	NEG	0.932344	[women, goldman, sachs, reported, 75, incident...
...	...	...	...	...	...	...
1129	Apple	2019-06-07T16:00:35.000Z	https://www.bbc.co.uk/news/technology-48555156	POS	0.837496	[every, apple, event, every, detail, new, mac,...
1130	Apple	2019-04-29T11:00:47.000Z	https://www.bbc.co.uk/news/technology-48092151	NEU	0.638554	[apple, defended, decision, remove, number, pa...
1131	Apple	2019-04-23T14:54:08.000Z	https://www.bbc.co.uk/news/technology-48022890	NEG	0.756685	[student, suing, apple, inc, 1bn, £0.77bn, cla...
1132	Apple	2021-09-03T15:16:05.000Z	https://www.bbc.co.uk/news/technology-58433647	NEG	0.816005	[apple, delayed, plans, roll, detection, techn...
1133	Apple	2019-06-04T14:24:32.000Z	https://www.bbc.co.uk/news/technology-48511006	POS	0.545913	[apple, announced, itunes, absorbed, three, ne...

1134 rows × 6 columns

Fig. 18. Data Frame Description of the model

E. BERTweet Model

This model is better than TextBlob for sentiment analysis on financial news for stock prediction because it is specifically designed to handle the nuances and complexities of sentiment analysis in social media contexts, such as Twitter.

Financial news and stock prediction data often contain specialized terminology, slang, and other types of language that may not be captured by more general sentiment analysis tools. BERTweet, on the other hand, has been pre-trained on a large corpus of Twitter data, which includes similar language and expressions that are commonly used in financial news and discussions about stocks and investing.

Additionally, the finiteautomata/bertweet-base-sentiment-analysis model has been fine-tuned on a sentiment analysis task using a labeled dataset, which may make it more accurate and precise in predicting sentiment in financial news data. In contrast, TextBlob is a more general-purpose sentiment analysis tool that may not be as well-suited to the specialized language and context of financial news and stock prediction data.

VI. Contributions

- Aruj Deshwal: Literature review, LSTM , GRU Model
- Ishaan Dayal: Literature review, LSTM Model
- Saatvik Bhatnagar: Introduction, Implementation, LSTM Model
- Hemang Dahiya: Literature review, Sentiment Analysis
- Hrishav Basu: Literature review, Sentiment Analysis
- Ipsita R: Literature review, Documentation

References

[1] Yoo, P. D., et al. “Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation.” International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06), vol. 2, 2005, pp. 835–41. IEEE Xplore, https://doi.org/10.1109/CIMCA.2005.1631572. Springer, 2010.

[2] Muhammad, Shakeel, and Gohar Ali. ”The relationship between fundamental analysis and stock returns based on the panel data analysis; evidence from karachi stock exchange (kse).” Research Journal of Finance and Accounting 9.3 (2018): 84-96.

[3] Berndt, Donald J., and James Clifford. ”Using dynamic time warping to find patterns in time series.” KDD workshop. Vol. 10. No. 16. 1994.

[4] Nikfarjam, Azadeh, Ehsan Emadzadeh, and Saravanan Muthaiyah. ”Text mining approaches for stock market prediction.” 2010 The 2nd international conference on computer and automation engineering (ICCAE). Vol. 4. IEEE, 2010.

[5] Drashti Talati, Miral Pate and Bhargesh Patel. Stock Market Prediction Using LSTM Technique. International Journal for Research Pages 1-11

[6] A. E. Khedr, S. E. Salama, and N. Yaseen, “Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis,” International Journal of Intelligent Systems and Applications, vol. 9, no. 7, pp. 22–30, Jul. 2017

[7] [1]C. M. C. Lee, “Earnings news and small traders,” Journal of Accounting and Economics, vol. 15, no. 2–3, pp. 265–302, Jun. 1992, doi: https://doi.org/10.1016/0165-4101(92)90021-s.

[8] [1]S. Borovkova and I. Tsiamas, “An ensemble of LSTM neural networks for high-frequency stock market classification,” Journal of Forecasting, May 2019, doi: https://doi.org/10.1002/for.2585.