Stock Market Prediction and Analysis: A Comprehensive Review and Development of an Intelligent Stock Market Guide

Ayush Srivastava Ankush Gupta Harshvardhan Singh Aditya Dahiya

Shantanu Prakash

1. Problem Statement

Traditional methods of stock prediction, such as technical and fundamental analysis, rely on historical data and company fundamentals to forecast future stock prices. In our research initiative, we endeavor to craft an intelligent system employing natural language processing and machine learning techniques such as LSTM and Random Forest. This system will provide users, both novice investors and seasoned traders, with easily interpretable visual as well as textual insights to mitigate risk and make informed investment decisions. By summarizing past financial activities, integrating with the latest news, and offering personalized advice, the system seeks to enhance user engagement and facilitate better financial outcomes.

2. Motivation

Predicting stock market trends has always been a challenge for professionals in the financial sector, including statisticians, economists, and other experts. Stocks play a vital role in driving the economy, being bought, sold, and traded on various stock exchanges. However, the market is influenced not only by economic factors but also by psycho- logical and human elements, making it inherently unpre- dictable. Hence, our purpose of creating an interactive beginner-friendly chatbot service to develop a financial advisor which consider companies' past performance as well as relevant news around it to give sound and well-informed stock guidance and investment strategies

3. Literature Review

1. **Sentiment analysis of financial news using unsupervised approach**

   This paper explores Sentiment analysis and aims to determine the sentiment strength from a textual source for good decision making. This work focuses on application of sentiment analysis in financial news. The semantic orientation of documents is first calculated by tuning the existing technique for financial domain. The existing technique is found to have limitations in identifying representative phrases that effectively capture the sentiment of the text. Two alternative techniques - one using Noun-verb combinations and the other a hybrid one, are evaluated. Noun-verb approach yields best results in the experiment conducted.

2. **Stock market prediction using machine learning classifiers and social media news**

   This research paper explores the use of social media and financial news data to predict stock market trends. The authors collected data from Twitter, Yahoo Finance, and Business Insider for two years, and used sentiment analysis to assess the overall sentiment of the data. They then trained ML classifiers on the combined data set to predict future stock market trends. The results showed that the classifiers were able to predict stock market trends with some accuracy, but that the performance varied depending on the classifier and the stock market being predicted. The authors also found that social media data was more useful for predicting short-term trends, while financial news data was more useful for predicting long-term trends.

3. **Stock Market Prediction using Financial News Articles**

   This paper discusses extracting data from trusted news sources, cleaning the text using natural language processing, and applying sentiment analysis techniques to determine sentiment polarity. The review highlights the use of machine learning algorithms, such as Linear Regression, to predict stock prices based on historical patterns and news sentiment. It also mentions the integration of sentiment analysis with financial news articles, giving equal weightage to the model to enhance predictive models. The methodology emphasizes feature selection methods and classification techniques to improve the accuracy of stock price predictions.

4. **News Sensitive Stock Trend Prediction**

   The paper introduces a novel methodology for stock trend prediction that integrates incremental K-means clustering, new weighting schemes, and market simulation. They explored time series segmentation techniques and document clustering. It addresses the limitations of traditional approaches by incorporating incremental K-means clustering to filter news articles and align them with stock trends. They introduced a new weighting scheme that enhances feature importance identification within article collections, improving prediction accuracy. Their methodology also leverages agglomerative hierarchical clustering based on slopes and coefficients of determination to cluster interesting trends and align news articles accordingly.

5. **Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis**

   Previous studies have focused on analyzing the correlation between economic news sentiment and stock market behavior, utilizing techniques like SVM and KNN classifiers. These studies have shown accuracies ranging from 72.73% to 86.21% in predicting stock trends. In the research paper, researchers highlights the use of a Java-based machine learning toolkit for natural language processing to analyze textual data related to the stock market. They combine sentiment analysis of news articles with historical stock price data to increase accuracy. This paper emphasizes the importance of preprocessing techniques, feature weighting methods like TF-Idf, and the integration of sentiment analysis with technical analysis for robust stock market prediction models.

6. **NEU-Stock: Stock market prediction based on financial news**

   The paper addresses the enduring interest in forecasting stock price movements, particularly focusing on the influence of financial news on FPT Group's stock. Employing PhoBERT, a language model, the authors achieve a 93% accuracy in classifying the impact of financial article titles on stock prices. Subsequently, they introduce the NEU-Stock model, utilizing LSTM-Attention architecture to forecast the next day's stock price, integrating past prices and news impact. The model exhibits strong performance, as evidenced by high R2 coefficients and notable RMSE, underscoring its efficacy in stock price prediction.

4. Novelty

For news extraction, StockWorthy considers both multi-model news data considering articles on financial issues from the web soucrse like BBC, TOI, CNBC and more, as well as video transcripts from YouTube to judge the news sentiment around a particular company.

Apart from that, our stock market guide merges sequence models such as LSTM for stock prediction with traditional ML classifer (random forest) to nitegratefinancial news sentiment as well as macro-factors (crude oil prices, godl prices, inr-usd exchange price).

This model is first of its kind to integrate these two models in the regime of stock predictions.

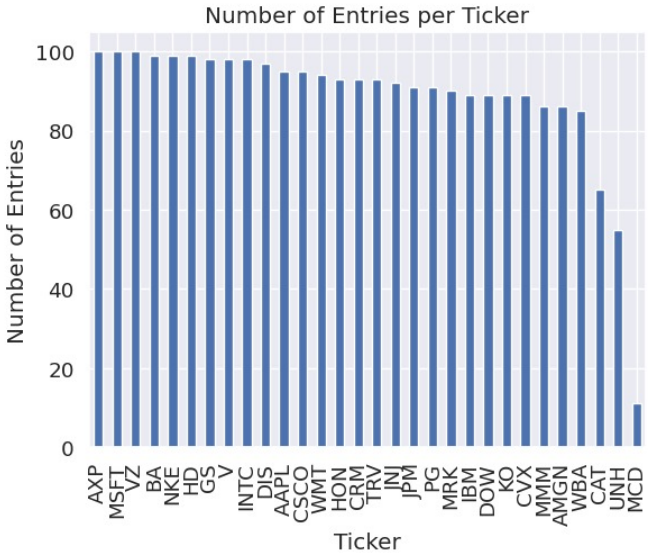


Figure 1. Number of entries per ticker

5. Data Analysis

5.1. Number of entries per company

To check class imbalance issues, we made a bar chart (Fig [3]) for the number of entries we had per company in our news dataset. Apart from some outliers, most of the companies had similar number of entries. Hence, there is no issue of class imbalance in the news dataset we have retrieved.

5.2. Covariance Heatmap

The darker the colour on the heat map, the more correlation there is between the emotions of the news sentiment analysis.

5.3. Frequency of values of Emotional Intensity

Since we used an external model for emotional analysis, we had to analyse its effectiveness and correctness. Hence, we have ploted the range of emotional scores for all the three classes of emotions in a combined line chart. Since all the three lines are similar in structure and magnitude, we can say that there is no imbalance in the magnitude of emotional score the model gives to the three emotions.

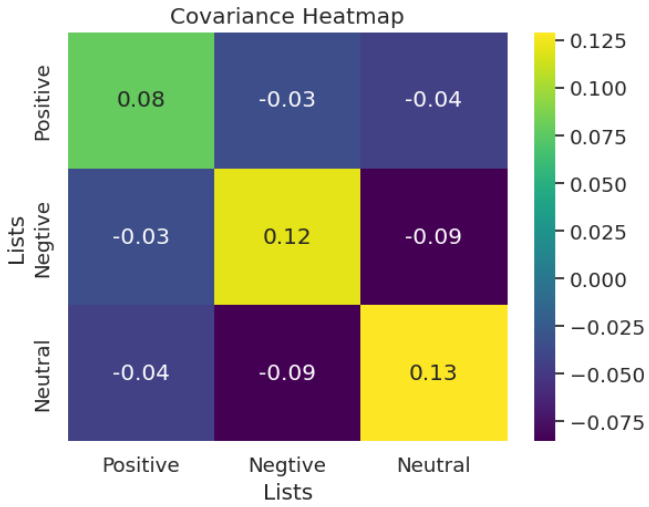


Figure 2. Covariance Heatmap

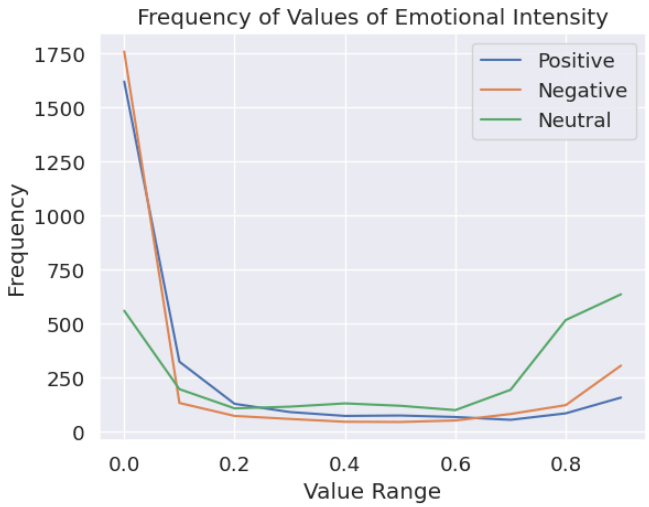


Figure 3. Frequency of values of Emotional Intensity

6. Final Project Review

6.1. Stock prediction

6.1.1 Data Collection and Preparation:

Utilized Yahoo Finance API via the ‘yfinance‘ library to retrieve historical stock price data for selected companies in the Dow Jones Industrial Average.

Specified a start date of March 1, 2015, and an end date of March 1, 2024.

Extracted stock price data for companies.

6.1.2 Data Preprocessing

Applied Min-Max scaling to normalize the data within a range of 0 to 1, ensuring consistency in the scale of features.

6.1.3 Model Building

Employed Long Short-Term Memory (LSTM) neural network architecture for time-series forecasting, implemented using TensorFlow and Keras.

Constructed a sequential model consisting of multiple LSTM layers with dropout regularization to mitigate overfitting.

Defined the network architecture with varying LSTM units to effectively capture temporal dependencies.

Compiled the model using the Adam optimizer and mean squared error loss function.

6.1.4 Training

Trained the LSTM model on each selected company’s training data from 2015 to 2021.

Configured the model to predict the stock price for the following day based on past closing prices and news impact.

6.1.5 Testing and Evaluation

Evaluated the model’s performance using test data from 2022 onwards.

Predicted the stock prices for the test period and compared them with actual values.

Calculated metrics such as coefficient of determination (R2) and root mean squared error (RMSE) to assess the model’s accuracy and reliability.

6.1.6 Result Analysis

Analyzed the model’s performance based on the visualizations and evaluation metrics, highlighting its effectiveness in forecasting stock prices.

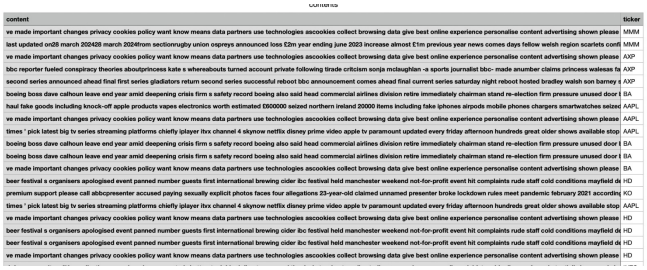
6.2. News Retrieval

6.2.1 NewsApi

We retrieved news articles from the internet using NewsApi’s JSON Api for 30 companies. The API was used to fetch news headline, description, content(truncated to 200 chars), url, author, date published and company tags for news articles related to any of the 30 companies from 8th February 2024 to 6th March 2024. These were then put into a csv file in a structured manner for sentiment analysis.

6.2.2 Google News

We first extracted all the URLs related to the 30 companies published on various reliable and free-to-use news agencies websites - BBC, Times Of India, CNBC, Livemint, Business Today, Finshots, Yahoo Finance, The Economist;



content	ticker
ve made important changes privacy cookies policy want know means data partners use technologies ascookies collect browsing data give best online experience personalise content advertising shown please	MMM
last updated on28 march 202428 march 2024from sectionrugby union ospreys announced loss £2m year ending june 2023 increase almost £1m previous year news comes days fellow welsh region scarlets confi	MMM
ve made important changes privacy cookies policy want know means data partners use technologies ascookies collect browsing data give best online experience personalise content advertising shown please	AXP
bbc reporter fueled conspiracy theories aboutprincess kate s whereabouts turned account private following tirade criticism sonja mclaughlan -a sports journalist bbc- made anumber claims princess walesas fa	AXP
second series announced ahead final first series gladiators return second series successful reboot bbc announcement comes ahead final current series saturday night reboot hosted bradley walsh son barney s	AXP
boeing boss dave calhoun leave end year amid deepening crisis firm s safety record boeing also said head commercial airlines division retire immediately chairman stand re-election firm pressure unused door	BA
haul fake goods including knock-off apple products vapes electronics worth estimated £600000 seized northern ireland 20000 items including fake iphones airpods mobile phones chargers smartwatches seized	AAPL
ve made important changes privacy cookies policy want know means data partners use technologies ascookies collect browsing data give best online experience personalise content advertising shown please	AAPL
times ' pick latest big tv series streaming platforms chiefly iplayer itvx channel 4 skynow netflix disney prime video apple tv paramount updated every friday afternoon hundreds great older shows available stop	AAPL
boeing boss dave calhoun leave end year amid deepening crisis firm s safety record boeing also said head commercial airlines division retire immediately chairman stand re-election firm pressure unused door	BA
boeing boss dave calhoun leave end year amid deepening crisis firm s safety record boeing also said head commercial airlines division retire immediately chairman stand re-election firm pressure unused door	BA
ve made important changes privacy cookies policy want know means data partners use technologies ascookies collect browsing data give best online experience personalise content advertising shown please	BA
beer festival s organisers apologised event panned number guests first international brewing cider ibc festival held manchester weekend not-for-profit event hit complaints rude staff cold conditions mayfield d	HD
premium support please call abbcpresenter accused paying sexually explicit photos faces four allegations 23-year-old claimed unnamed presenter broke lockdown rules meet pandemic february 2021 according	KO
times ' pick latest big tv series streaming platforms chiefly iplayer itvx channel 4 skynow netflix disney prime video apple tv paramount updated every friday afternoon hundreds great older shows available stop	AAPL
ve made important changes privacy cookies policy want know means data partners use technologies ascookies collect browsing data give best online experience personalise content advertising shown please	HD
beer festival s organisers apologised event panned number guests first international brewing cider ibc festival held manchester weekend not-for-profit event hit complaints rude staff cold conditions mayfield d	HD
beer festival s organisers apologised event panned number guests first international brewing cider ibc festival held manchester weekend not-for-profit event hit complaints rude staff cold conditions mayfield d	HD
ve made important changes privacy cookies policy want know means data partners use technologies ascookies collect browsing data give best online experience personalise content advertising shown please	HD

Figure 4. CSV files of content of news articles generated using google news

for a particular time frame, i.e. from 40 days of the request. Then, using the retrieved link, we extracted content published. After this retrieval, we performed basic preprocessing steps, i.e., tokenization, removal of stop words, punctuations, and new line characters on the text extracted. And stored the extracted data in the csv files for the given 30 companies.

6.2.3 YouTube

We extracted top Youtube video links for the 30 companies with the help of Youtube API. These results were ordered by relevance. After which we obtained summaries of these videos using summarize.text (A free text summarizer) and stored them in a csv file for the 30 companies. This was achieved using Selenium and Chrome WebDriver.

6.3. News Sentiment Analysis

To perform sentiment analysis on the financial news obtained, we used the FinBert Model which is built by fine tuning a Bert Model on a large financial corpus. For the financial news data fetched for the past one month, we obtained the sentiment scores(positive, negative and neutral) using softmax activation on the output of the FinBert model which are further utilized by the random regressor model.

The image of csv file is shown in figure 1.

6.4. Merging News Sentiments and Stock Predictions

6.4.1 Merging the datasets

For all dates of month February 2024, we took the average sentiments (positive. negative, neutral) of the news for the specific company, and added them in the stock predictions dataset.

Previous trends and news are not the only factors that affect stock prices, there are many other macro-factors/features that affect them such as the economy, the market itself, foreign policies/prices.

We took Gold prices, crude oil prices, price of the Indian rupee(compared to dollars) as a medium of verifying how the economy is performing and retrieved the macro-factor prices stated above for all dates into the dataset as well.

Combined the dataset of all the companies to train a random forest regressor.

6.4.2 Training and Testing

Created a 75-25 split training set to train the random forest.

Made predictions on the test set to check the accuracy of the model.

6.5. ChatBot using Gemini API

We use the GenerativeAI module with an API key, specifying generation configuration and safety settings. Employ pattern matching to identify company ticker symbols in user queries, retrieving relevant financial data. Model initialization initializes the GenerativeAI model and prompts a conversation. Contextual information, including user queries, is sent to the model, which generates responses based on the provided context and prompts. The resulting responses are printed to the console for potential further processing or display. Overall, we facilitate generating financial advice by leveraging contextual data and a GenerativeAI model.

6.6. News Articles retrieval using BM25

To extract the relevant news articles with respect to the input user query, the extracted ticker, input query and date is utilised to extract the relevant news articles for the user. We used 'rankbm25' library, 'BM250kapi' to provide scores for the articles in our dataset for retrieving relevant news articles from rows of our dataframe corresponding to the desired company and date range.

7. Results

We extracted the stock prices and split the dataset into train and test dataset as shown in Figure[2]. Then we used LSTM to predict the values of stock price based on the trend, the variation of stock price of 'Apple' is shown in Figure[3]. Then we fetched the news articles using API and after performing pre-processing on fetched data, we performed sentiment analysis and the result is shown in Figure[4]. Then the random forest predicts the stock prices of company based upon the LSTM and sentiment analysis as shown in Figure[5] and Figure[6]. We have got R2 score greater than 0.9, and RMSE for the model is around 13.445.

Utilising the gold rates help us improve our R2 score from 0.9 to a rough value around 0.92.

We have connected our backend code to a basic Frontend, the front end is similar to Figure[7].

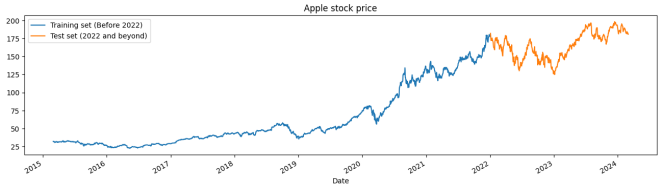


Figure 5. Extracting Stock Prices using API and splitting it into train and test data

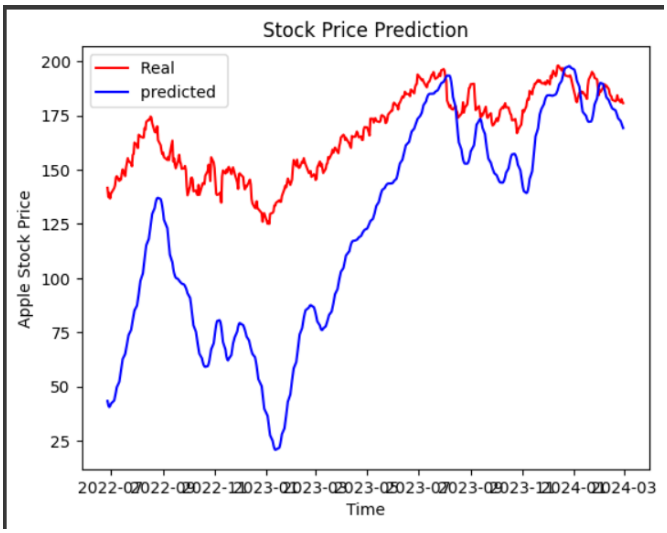


Figure 6. LSTM output for "Apple"

sno	18
headline	Game Font Forensics
description	VileR's blog: old school PCs, games, graphics,...
content	"THAT'S DPAINT COMIX! I can tell from some of ...
url	https://int10h.org/blog/2024/02/game-font-fore...
author	NaN
date	2024-02-18 13:50:23+00:00
ticker	AAPL
Positive	0.051831
Negative	0.023653
Neutral	0.924516

Name: 318, dtype: object

Figure 7. Sentiment Analysis output

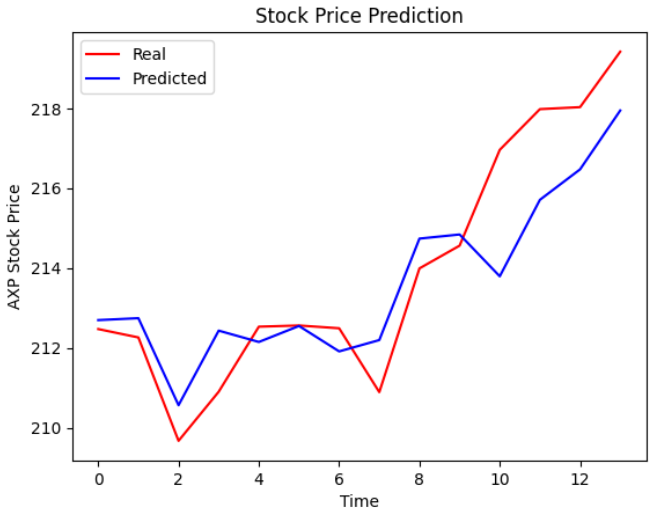


Figure 8. Random Forest output

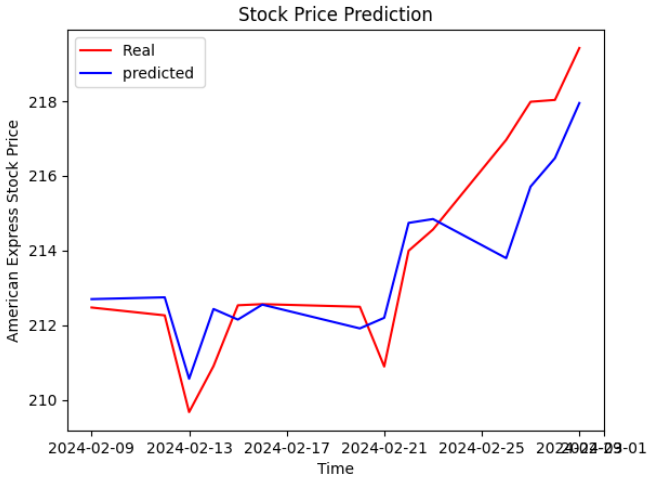


Figure 9. Random Forest output after adding Gold Rates

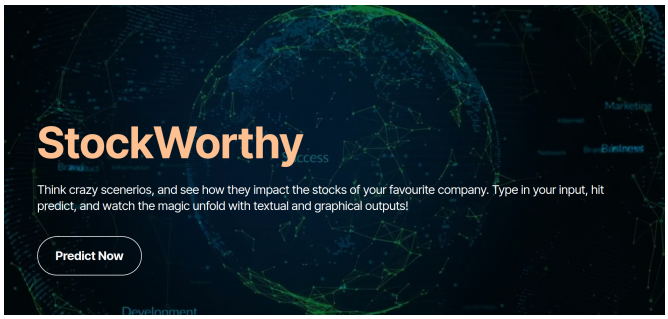


Figure 10. Basic Frontend

8. References

- Chen MS, ed. Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference; Proceedings. Springer; 2002.
- Faculty of Computers and Information Technology, Future University in Egypt, Khedr AE, S.E.Salama, Yaseen N. Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. IJISA. 2017;9(7):22-30. doi:10.5815/ijisa.2017.07.03
- M S. Stock Market Prediction using Financial News Articles. https://www.irjet.net/archives/V6/i12/IRJET-V6I12229.pdf
- Khan W, Ghazanfar MA, Azam MA, Karami A, Alyoubi KH, Alfakeeh AS. Stock market prediction using machine learning classifiers and social media, news. J Ambient Intell Human Comput. 2022;13(7):3433-3456. doi:10.1007/s12652-020-01839-w

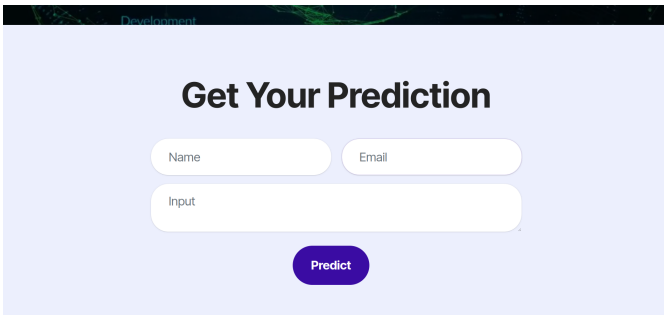


Figure 11. Basic Frontend

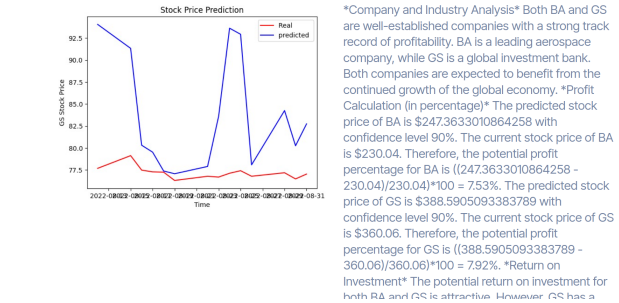


Figure 12. Basic Frontend

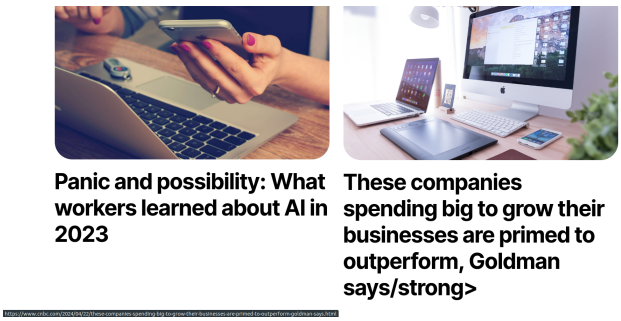


Figure 13. Basic Frontend

- Maqbool J, Aggarwal P, Kaur R, Mittal A, Ganaie IA. Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach. Procedia Computer Science. 2023;218:1067-1078. doi:10.1016/j.procs.2023.01.086
- Kalyani J, Bharathi ProfHN, Jyothi ProfR. Stock trend prediction using news sentiment analysis. Published online 2016. doi:10.48550/ARXIV.1607.01958
- NEU-Stock: Stock market prediction based on financial news https://ceur-ws.org/Vol-3026/paper24.pdf
- Sentiment analysis of financial news using unsupervised approach https://www.sciencedirect.com/science/article/pii/S1877050920307912?ref=pdf$_{d}ownloadfr = RR - 2rr = 86247f833c3fecac$