

Stock Market companion: real-time data, hybrid analysis, and enhanced user interaction

Ms Sneha Jha, Ms Madhurima Rawat , Mr Geetanshu Dev Meshram

Department of Data Science , Chhattisgarh Swami Vivekanand Technical University
Department of Data Science , Chhattisgarh Swami Vivekanand Technical University
Department of Data Science , Chhattisgarh Swami Vivekanand Technical University

Abstract

Stock price prediction has long been a major focus in financial sciences due to the need to anticipate movements in an unpredictable market. Traditional models often rely only on numerical data, overlooking external factors like news and social media. This project introduces a hybrid stock price prediction model that combines historical stock data with sentiment analysis of financial texts for improved accuracy. While initially demonstrated on eight major tech companies, the system is flexible enough to predict the stock prices of any publicly traded company, including popular names like Apple, Nvidia, Tesla, and Microsoft. Numerical data is processed using machine learning algorithms in Scikit-learn, and natural language processing (NLP) techniques assign sentiment scores to textual data. For real-time data handling and visualization, InfluxDB and Grafana are used alongside a detailed Power BI dashboard, which presents key insights, sentiment trends, and predictive analytics in an intuitive format. A Streamlit application allows users to interact with the model easily, and a Flask API backend supports scalable integration and future development. The project is open-source, available on GitHub with complete documentation, ensuring accessibility for collaboration and further innovation. By combining structured numerical analysis with unstructured sentiment data, the project offers a comprehensive and scalable approach to stock market forecasting.

Keywords: Stock Market , Streamlit , Grafana

Introduction

The stock market plays a critical role in global economies, acting as a barometer for economic health and offering opportunities for investment and wealth creation. However, predicting stock price movements is one of the most challenging tasks due to the inherent volatility, randomness, and the influence of numerous factors. This research paper introduces a “hybrid model” that integrates numerical and textual data for advanced stock price prediction forecasting. The numerical component utilizes historical stock prices and market trends to analyze quantitative patterns. The textual component employs sentiment analysis on financial news and social media data to gauge market sentiment and its impact on stock prices. By combining these two dimensions, the hybrid model aims to provide a more comprehensive and accurate prediction mechanism compared to traditional methods. The significance of this study lies in its potential to bridge the gap between quantitative and qualitative factors in stock forecasting. The hybrid approach can empower investors, traders, and financial institutions to make informed decisions based on a more holistic understanding of market behavior. Furthermore, the integration of machine learning algorithms enhances the model’s adaptability to complex patterns and changing market conditions, making it a valuable tool in dynamic financial environments.

Literature Review:

An extensive review of existing research and methodologies in the field of stock market prediction was carried out. This helped in identifying the strengths and limitations of various approaches. In the study [1], the research focuses on developing advanced deep learning models for live stock price predictions, specifically introducing the Fast RNN model, which demonstrates a low root mean squared error (RMSE) of 0.02068 and a computation time of 3.35 seconds. Additionally, a hybrid model combining Fast RNN, CNN, and Bi-LSTM is proposed, achieving an RMSE of 0.02181 and a computation time of 18.18 seconds. These proposed models significantly outperform traditional forecasting methods such as ARIMA and FBProphet, which have higher RMSE values of 0.796 and 0.935, respectively, showcasing their effectiveness for real-time stock predictions. The study[2] evaluates the performance of the MVL-SVM model, which integrates multi-view learning with support vector machines (SVM) for stock price prediction. The model achieves an impressive accuracy of nearly 88%, significantly outperforming other baseline models, including ARIMA and traditional SVM approaches, which show lower accuracies around 70%. Additionally, the MVL-SVM model based on news and daily returns demonstrates superior prediction capabilities, with average and median accuracies consistently above 0.8767 across various time frames, indicating its effectiveness in leveraging financial news and market data for improved trading strategies. The research [3] presents a hybrid stock price prediction model that combines Prediction Rule Ensembles (PRE) and Deep Neural Network (DNN) techniques, authored by Srivinay, B.C. Manujakshi, M.G. Kabadi, and N. Naik, and published on April 20, 2022. The model utilizes moving average indicators over 20, 50, and 200 days to identify stock trends and demonstrates improved prediction accuracy, achieving a Root Mean Square Error (RMSE) of 5.60 for ICICI Bank and 6.30 for SBI Bank, among others. Overall, the proposed hybrid model's RMSE scores are 5% to 7% lower than those of existing DNN and ANN models, indicating its effectiveness in stock price forecasting. The study[4] explores the use of deep learning techniques to predict stock prices by integrating numerical data (historical stock prices) and textual data (news articles). The research emphasizes the importance of combining these data types to improve prediction accuracy, although it does not provide specific accuracy metrics or detailed drawbacks. In the study [5] the authors present an ensemble model that combines deep learning methods to utilize both numerical stock prices and news articles for predictions. The model achieves an accuracy of 85.5%, but it requires significant computational resources and large datasets for training. The complexity and resource demands are noted as potential drawbacks. This study[6] proposes an ensemble approach, integrating Transformer models, ARIMA, and Linear Regression to predict stock prices. It highlights the potential of combining these diverse methods for better accuracy but does not specify the exact accuracy achieved. The complexity and interpretability challenges of the ensemble model are mentioned as drawbacks. This research[7] integrates sentiment analysis of news articles with technical analysis of historical stock prices using deep learning models to predict stock market

movements. The study suggests improved prediction capabilities but does not provide specific accuracy figures. The sensitivity to the quality of textual data is considered a limitation. It [8] focuses on employing NLP techniques to analyze financial news articles for predicting stock market trends. While it emphasizes the relevance of incorporating textual data, it does not specify accuracy metrics. The quality and timeliness of news data are critical factors affecting the model's performance. This [9] introduces a deep neural generative model that uses news articles to forecast stock prices. The paper highlights the potential of generative models in this context but does not provide detailed accuracy results. Preprocessing and handling ambiguous news content are noted as challenges. The research [10] explores the use of NLP and deep learning to predict stock prices based on financial news. It emphasizes the combination of textual and numerical data for enhanced predictions but lacks specific accuracy metrics. The specialized nature of the financial news corpus limits the model's broader applicability.

Methodology:

DFD : The Data Flow Diagram (DFD) is used to visualize the system processes, data flow, and interaction with external entities. Two levels of DFDs are presented: Level 0 (context diagram) and Level 1 (decomposition diagram).

DFD Level 0: Context Diagram Description:

The Level 0 DFD provides a high-level view of the system, representing it as a single process. It highlights the interaction with external entities and data flow.

Components:

Stock Data Source: Represents the source of historical stock data, such as open, high, low, and volume for seven technology companies.

User: The user interacts with the system to view the predicted close prices.

Process: Stock Price Prediction System: The core process that receives stock data, processes it, and predicts the close price.

Stock Data Store: Stores the historical stock data needed for processing.

Prediction Data Store: Stores the predicted close prices for user access.

Data Flows:

Input: Data fields (open, high, low, volume) flow from the Stock Data Source to the system.

Output: Predicted close prices flow from the system to the User.

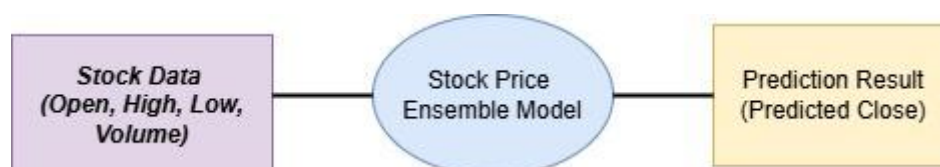


Figure 4.1 DFD Level 0 diagram

DFD Level 1: Decomposition Diagram

Description: The Level 1 DFD provides a detailed view of the core system by breaking it down into four sub-processes. It captures the flow of data between these processes and the corresponding data stores.

Sub-Processes:

- **Data Collection:** Collects historical stock data (open, high, low, volume) from the stock data source. Stores the collected data in the Stock Data Store.
- **Data Preprocessing:** Cleans and preprocesses the collected data by handling missing values, scaling features, and splitting data into training and testing sets. Outputs preprocessed data for further analysis.
- **Regression Model:** Applies a regression model (using Scikit-learn) to predict the close price based on the preprocessed data. Outputs predicted close prices to the Prediction Data Store.
- **Store Predictions:** Saves the predicted close prices in the Prediction Data Store for visualization and analysis by the user.
- **Data Stores:** Stock Data Store: Stores input data fields (open, high, low, volume).
Prediction Data Store: Stores predicted close prices for future use or visualization.
- **Data Flows:**
 - Data flows from the Stock Data Store to Data Preprocessing.
 - Preprocessed data flows into the Regression Model.
 - Predicted results flow from the Regression Model to the Prediction Data Store.
 - Users retrieve predictions from the Prediction Data Store.

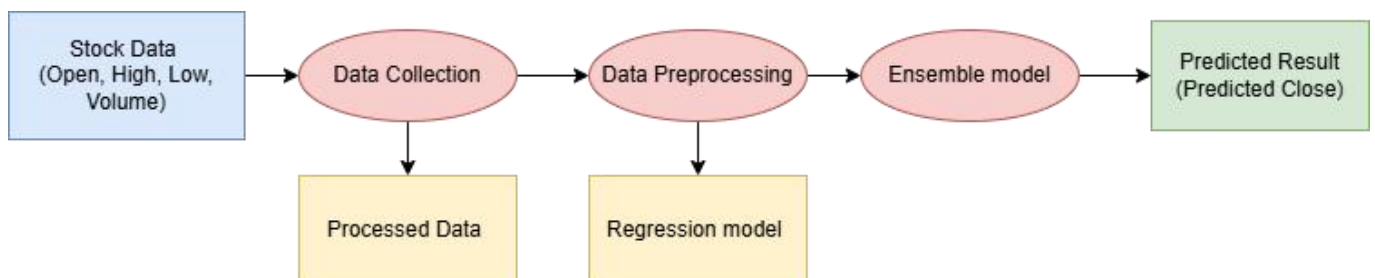





Figure 1. DFD Level 1 diagram

Data Collection: Historical stock data is sourced from reputable financial databases such as **Yahoo Finance, Kaggle and macrotrends**. These platforms provide extensive data on stock prices, trading volumes, and other relevant financial metrics.

Table 1 Companies dataset summary

Company	Data Range	Dataset Shape	Starting Stock Date	Current Stock Date	Starting Stock Price	Current Stock Price
 Alphabet Inc. (Google) [GOOG]	2014-03-27 : 2024-10-17	(2659, 5)	2014-03-27	2024-10-17	\$27.8542	\$164.51
 Amazon.com Inc. [AMZN]	1997-05-16 : 2024-10-17	(6901, 5)	1997-05-16	2024-10-17	\$0.0863	\$187.53
 Apple Inc. [AAPL]	1980-12-12 : 2024-10-17	(11053, 5)	1980-12-12	2024-10-17	\$0.0992	\$232.15
 Meta Platforms [META]	2012-05-18 : 2024-10-17	(3124, 5)	2012-05-18	2024-10-17	\$38.1174	\$576.93
 Microsoft Corp. [MSFT]	1986-03-13 : 2024-10-17	(9728, 5)	1986-03-13	2024-10-17	\$0.0603	\$416.72
 Netflix Inc. [NFLX]	2002-05-23 : 2024-10-17	(5640, 5)	2002-05-23	2024-10-17	\$1.1964	\$687.65
 Nvidia Corp. [NVDA]	1999-01-22 : 2024-10-17	(6477, 5)	1999-01-22	2024-10-17	\$0.0377	\$136.93

 Tata Consultancy Services [TCS]	2013-11-01 : 2024-10-17	(2758, 5)	2013-11-01	2024-10-17	\$543.0	\$11.8
---	----------------------------	-----------	------------	------------	---------	--------

Dataset Description

This dataset contains historical stock market data for 8 companies: **GOOG, AMZN, AAPL, META, MSFT, NFLX, NVDA, TCS**. The dataset provides crucial financial metrics, including daily **Open, High, Low, Close, Adjusted Close, Volume**, and **Date** for each stock. Our target is to build a predictive model to forecast the Closing Price of a stock based on the Open, High, and Low prices, along with the specific company data.

Dataset Fields: This section provides an in-depth explanation of each field in the stock market dataset. Understanding these fields is crucial for effectively analyzing and modeling stock price movements.

The goal is to predict the Closing Price based on the Open, High, and Low prices. Additionally, company data is used as an input feature to improve the model's learning capability. The dataset combines stock data from companies like GOOG, AMZN, AAPL, META, MSFT, NFLX, NVDA, TCS enabling the model to generalize across diverse data sources.

Table 2 Overview of Key Fields and their Importance in Stock Market Predictions

Field	Definition	Details	Usage/Significance
Date	The specific calendar day when the trading activity occurred.	<p>Format: Typically, YYYY-MM-DD (e.g., 2024-04-27).</p> <p>Importance: Tracks stock performance over time and identifies trends, seasonal patterns, and cyclical behaviors.</p> <p>Usage: Time-based aggregations like daily, weekly,</p>	Essential for aligning data from different sources and performing time-series analysis.

		or monthly analysis.	
Open	The price at which a stock starts trading when the market opens for the day.	<p>Determination: First trade executed during the trading session.</p> <p>Influencing Factors: Overnight news, pre-market sentiment, and global events.</p> <p>Importance: Serves as a reference for the day's activity and identifies opening trends in technical analysis.</p>	Indicates how the market opens, providing context for the day's price movements.
High	The highest price at which the stock traded during the trading day.	<p>Measurement: Recorded as the peak price reached from open to close.- Significance: Reflects maximum buying interest.</p> <p>Usage: Identifying resistance levels and comparing momentum with previous highs.</p>	Assesses price volatility and provides insights into market optimism.
Low	The lowest price at which the stock traded during the trading day.	<p>Measurement: Lowest price reached from open to close.</p> <p>Significance: Reflects maximum selling pressure.</p> <p>Usage: Identifying support levels and gauging bearish momentum.</p>	Helps define the trading day's price range and identifies points of significant selling activity.
Close	The final price at which the stock trades when the market closes for the day.	<p>Determination: Based on the last executed trade of the day.-</p> <p>Importance: Used as a benchmark for performance evaluation.</p>	Widely regarded as the most critical price for performance reporting and analysis.

		Usage: Calculating daily returns and forming the basis for indicators like moving averages.	
Volume	The total number of shares traded during the trading day.	Measurement: Sum of shares bought and sold. Significance: Reflects activity and liquidity. Usage: High volume confirms trends or identifies potential reversals/breakouts.	Indicates market interest and can highlight significant price movements.

This table summarizes the information systematically for easier understanding and reference.

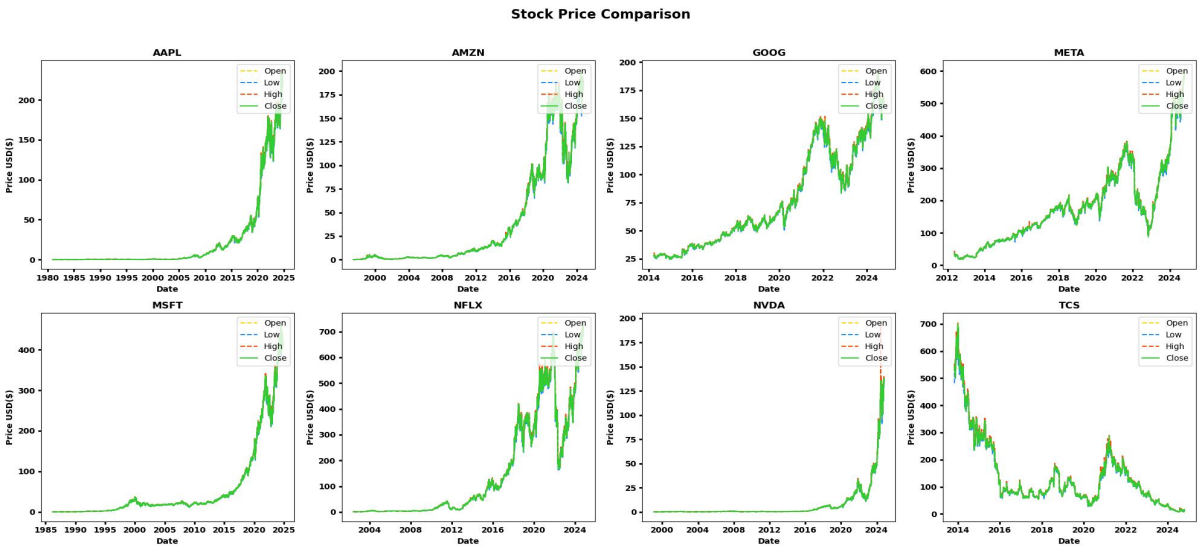


Figure 2 Stock Price Comparison of all 8 companies

The plots in Figure 2 illustrate the opening, high, low, and closing prices for all companies. By including these key price points alongside the original closing prices, the visualization provides a comprehensive comparison of price fluctuations and highlights the range of trading activity for each company over time.

Its primary goal is to preprocess these historical datasets, clean the data (e.g., handling missing values, dealing with outliers), and perform an exploratory data analysis (EDA). The goal of EDA is to summarize the main characteristics of the dataset and uncover underlying patterns, trends, and relationships.

Visualizations such as time series plots, histograms, and correlation heatmaps will help to assess the distribution of prices, trading volumes, and the relationships between stock prices and volume.

Thus, the initial data loading phase involves reading CSV files for Google and Apple. For Google, the dataset contains columns such as date, open, high, low, close, and volume, with data starting from March 2014. For Apple, the dataset spans a longer timeframe, starting from December 1980, and shares a similar structure. Key insights, such as Apple's starting price of \$0.0992 in 1980 and its current price of \$232.15 in October 2024, demonstrate the significant growth of these companies over time.

Moving Averages

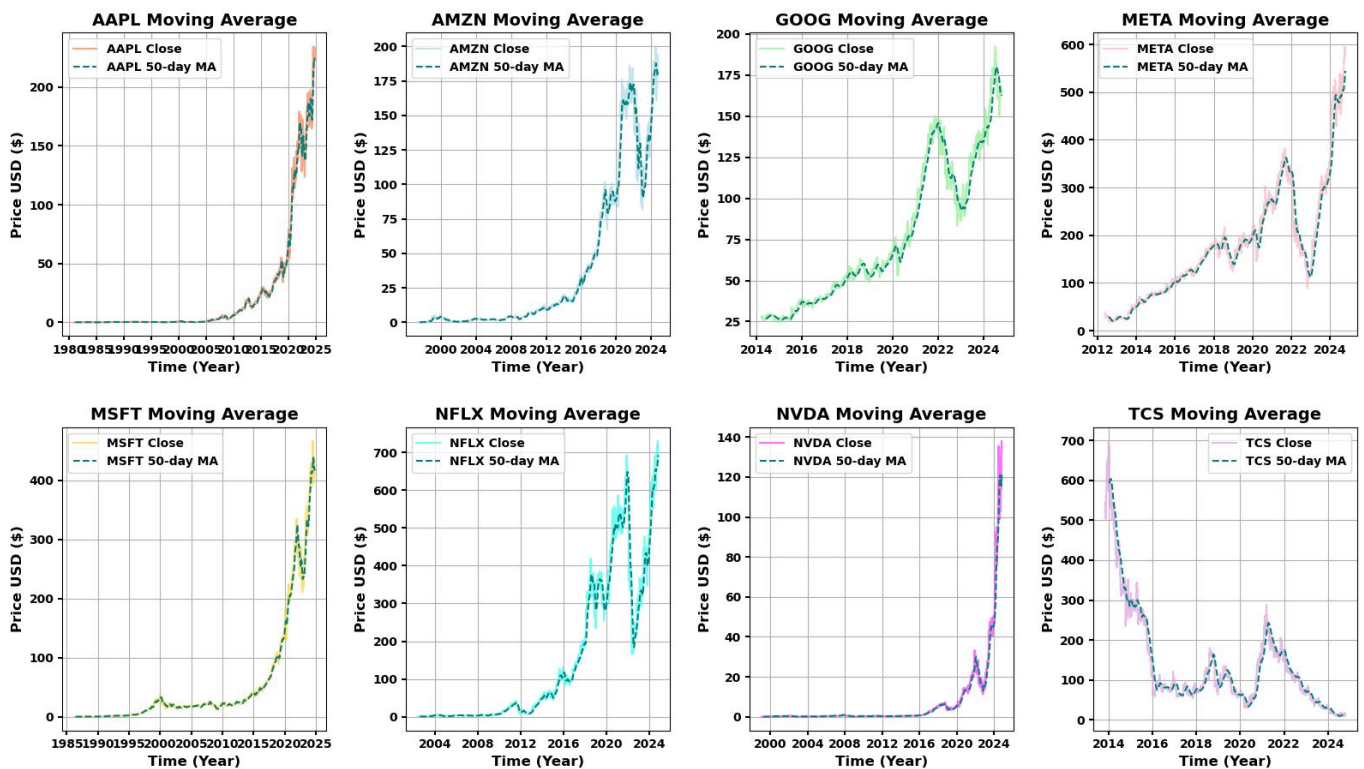


Figure 3. Close Price Moving Average Comparison of all 8 companies

The plots in Figure 3 depict the moving averages of all companies, overlaid with their original closing prices. This approach offers a clearer perspective on price trends by smoothing short-term fluctuations, enabling better insights into the overall performance and market behavior of each company. The EDA serves as a foundational step for building models that can predict stock price movements, identify market trends, and make informed trading decisions. The preprocessing and EDA ensure that the data is in a clean and usable format for future analysis, such as forecasting stock prices using machine learning models or conducting more advanced financial analysis. These datasets are uniformly structured, featuring price data (open, high, low, close) and trading volume, making them suitable for comparative

and trend analysis. It sets the stage for subsequent steps, which include cleaning the data, performing exploratory data analysis (EDA), and deriving meaningful insights through statistical and visual techniques. Thus , the EDA serves as a foundational step for building models that can predict stock price movements, identify market trends, and make informed trading decisions. The preprocessing and EDA steps outlined in the notebook ensure that the data is in a clean and usable format for future analysis, such as forecasting stock prices using machine learning models or conducting more advanced financial analysis.

Result and Discussion

Trends and Insights:

Companies like NVDA and META exhibited rapid price growth in recent years, reflecting high volatility influenced by technological advancements and market dynamics.TCS, on the other hand, showed a steady upward trend, suggesting consistent performance and less susceptibility to market shocks.

Statistical Analysis:

Measures such as mean, median, and standard deviation revealed the overall behavior of each stock. Skewness and kurtosis values indicated whether the stock prices were symmetrically distributed or had heavy tails, helping to guide feature engineering.



Figure 4. InfluxDB database showing all measurements

As shown in the above snapshot, there are four distinct measurements, each serving a specific purpose:

- **stock_price:** Stores numerical data related to stock prices.
- **model_prediction:** Contains numerical data representing model predictions.
- **textual_analysis:** Holds text-based analytical data.
- **hybrid_model:** Combines numerical and textual data for hybrid analysis.

These measurements are designed to categorize and store data based on its type and usage. Numerical data, such as stock prices and model predictions, enables quantitative analysis and trend visualization. Textual data, processed under textual_analysis, provides insights derived from qualitative information like sentiment or news. The hybrid_model integrates both numerical and textual elements, offering a comprehensive perspective for advanced analysis and decision-making. **Numerical Analysis**

The **Numerical Analysis** section is dedicated to examining quantitative data such as stock prices and related metrics. It includes various panels with graphs and charts, such as line plots, bar charts, and histograms, to display trends, price movements, and fluctuations over time.



Figure 5 Numerical Analysis Panel Google Close data

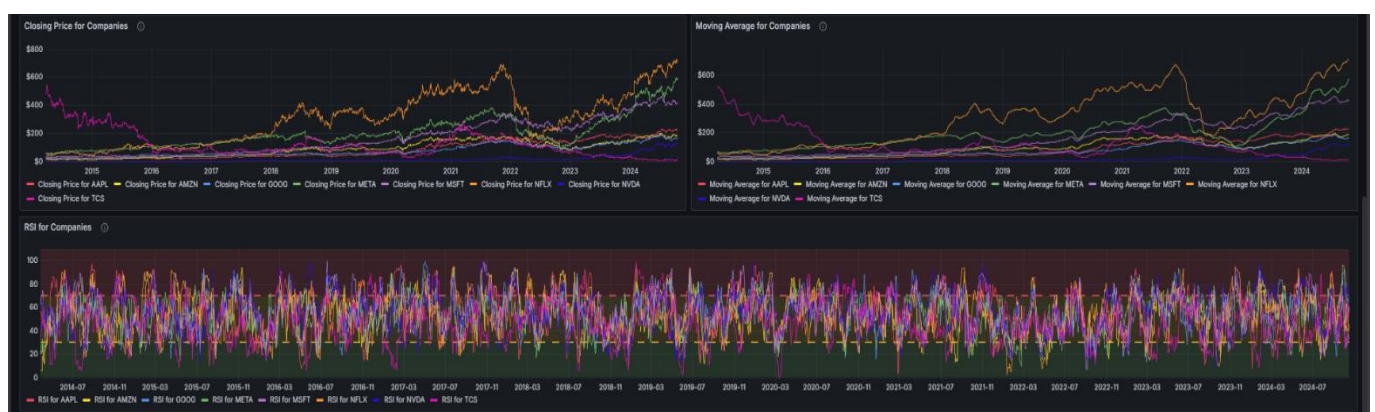


Figure 6 EDA of 'close' data of all company

This section enables users to identify patterns and gain insights into numerical data performance, essential for making data-driven financial and business decisions.

Model Prediction (Numerical Data)

The **Model Prediction (Numerical Data)** section highlights the results from predictive models and their comparisons to real-world data. This section features line plots, scatter plots, and error analysis charts that display model performance, accuracy, and prediction intervals.

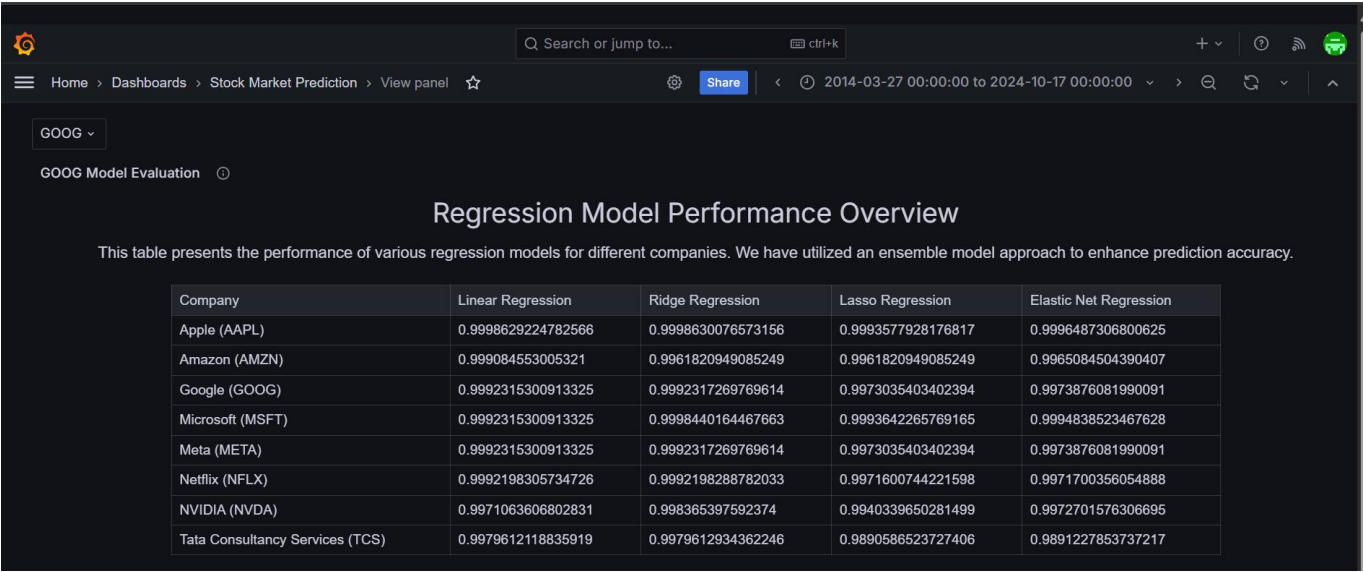


Figure 7 Numerical Model Evaluation

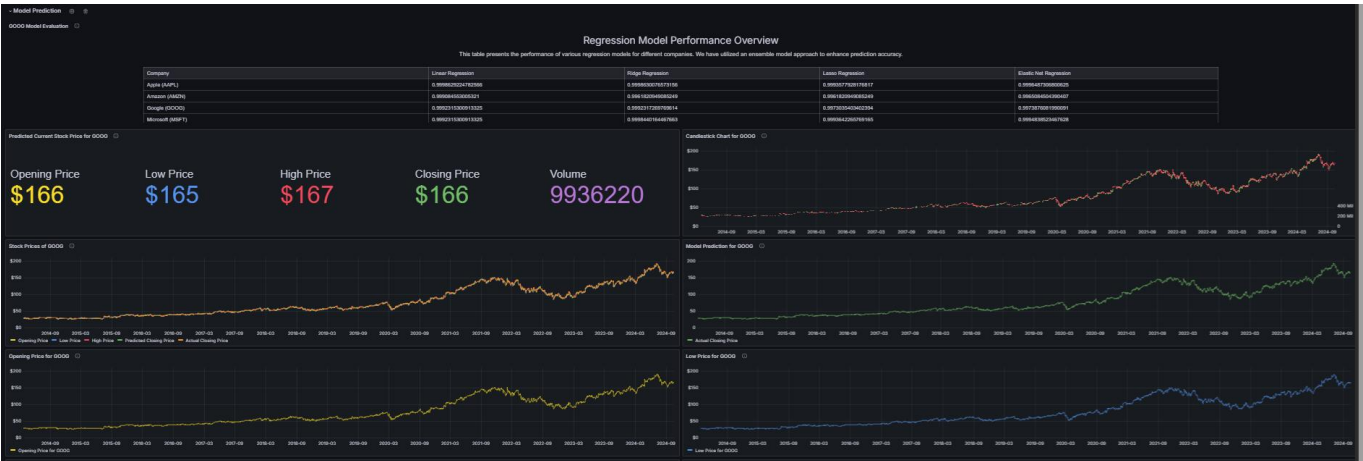


Figure 8 Numerical Model Visualization

Users can assess the reliability and accuracy of the models, helping to evaluate their suitability for forecasting future trends and informing strategic planning.

Textual Analysis: The **Textual Analysis** section is designed for the processing and visualization of text-based data. It showcases panels featuring word clouds, sentiment analysis graphs, and frequency charts that present key themes, keywords, and sentiment scores extracted from textual sources such as news articles, reports, or social media posts.

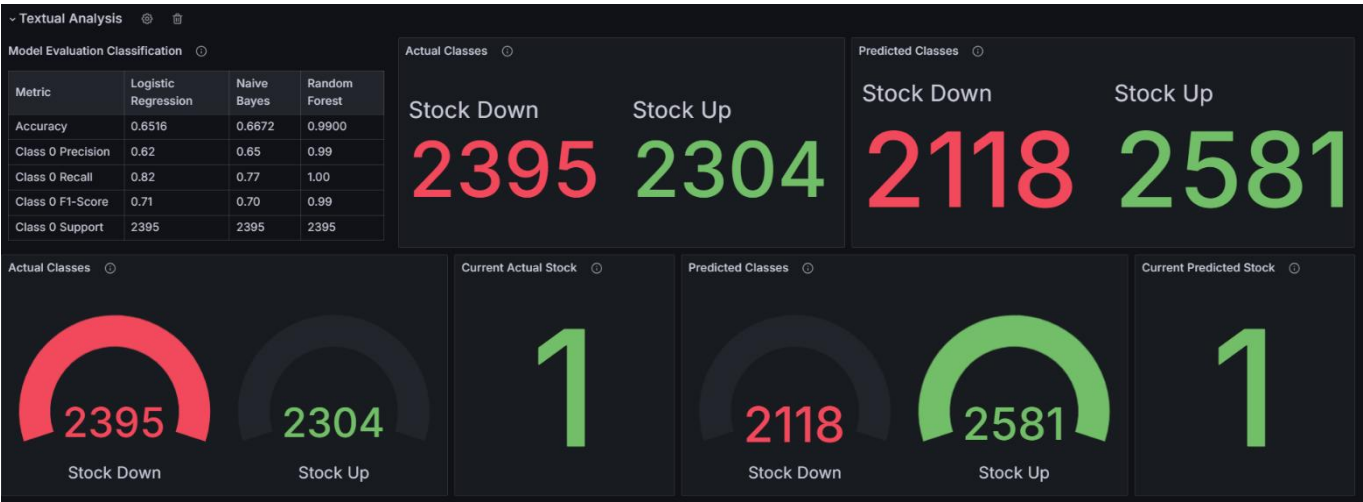


Figure 9 Textual Model Visualization

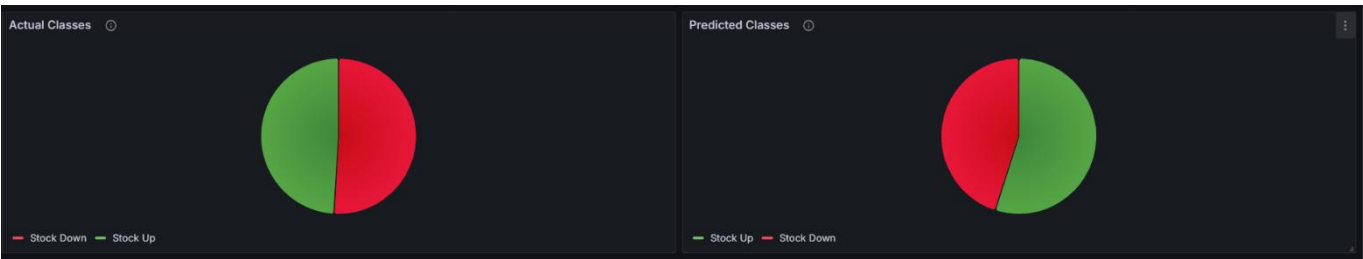


Figure 10 Textual Model Output Categories

This section provides users with a deeper understanding of the context and tone within the data, enabling insights into public perception, market sentiment, or consumer opinions.

Hybrid Model: The **Hybrid Model** section merges both numerical and textual data to present a multi-faceted analysis. This integrated approach allows for a richer interpretation of data, revealing connections and insights that would not be apparent when analyzing data types separately.

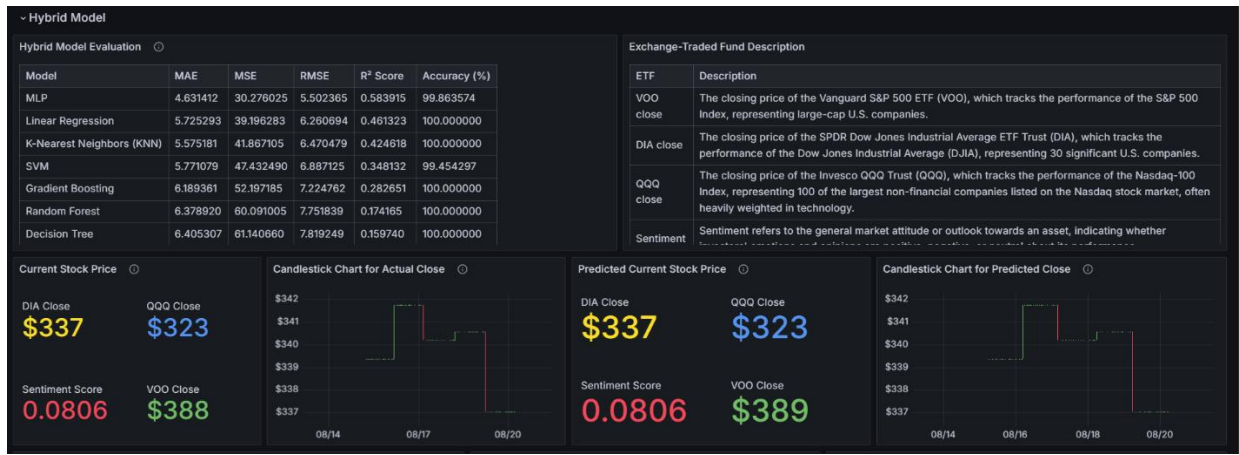


Figure 11 Hybrid Model Visualization

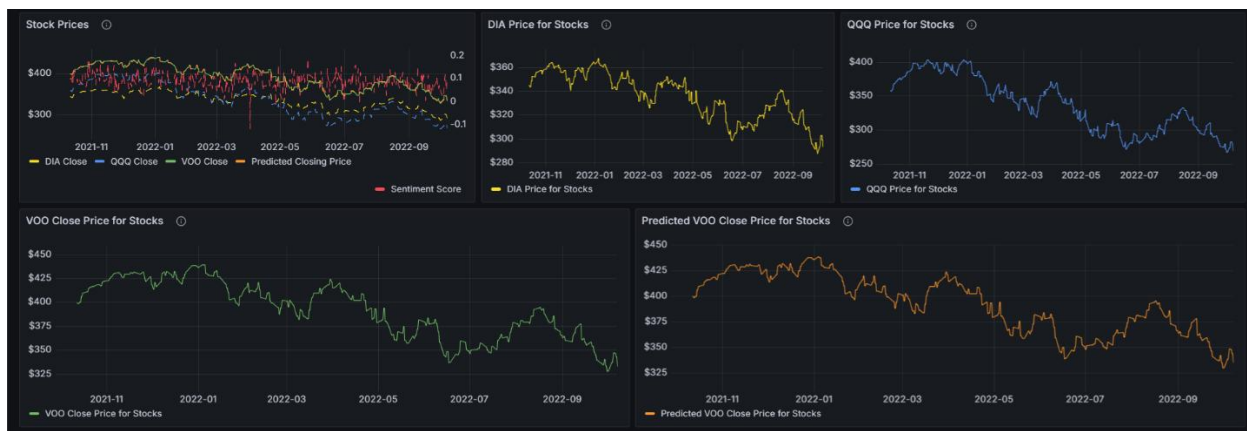


Figure 12 Hybrid Model Visualization with Numerical Features

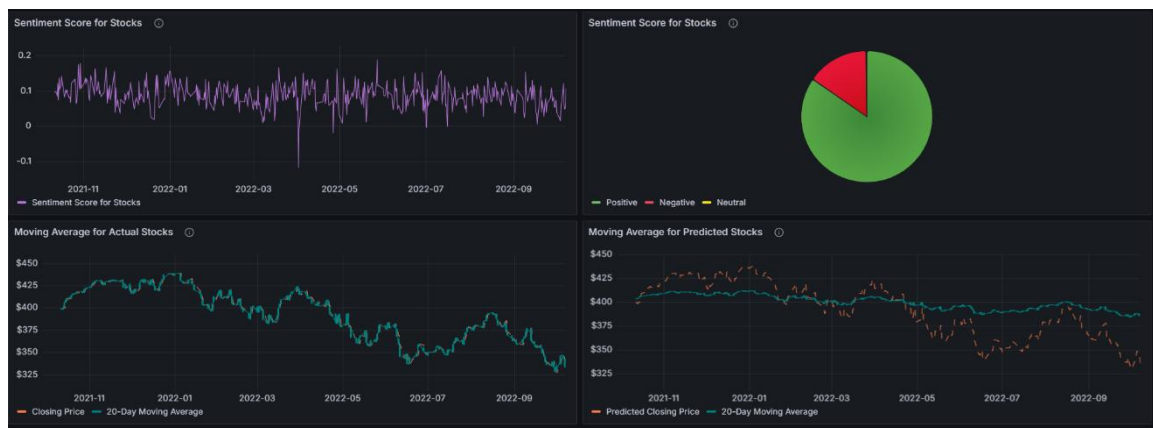


Figure 13 Hybrid Model Visualization with Text Features

Panels in this section include combined charts and pie charts that visualize how numerical data and textual analysis interact. This holistic approach supports comprehensive decision-making and strategic planning, leveraging the strengths of both data types.

Regression Results of the Ensemble Model

The ensemble regression model, implemented using the **Voting Regressor**, was applied to predict the closing prices of all selected companies.:

- The high R^2 scores across all companies demonstrated the ensemble model's ability to explain nearly all variability in the closing prices.
- **AAPL** and **META** achieved near-perfect predictions, reflecting consistent data patterns and minimal noise.
- **NVDA** had a slightly lower R^2 score (0.9975525), likely due to higher market volatility affecting its stock prices.

Discussion:

The ensemble model successfully combined the strengths of individual regression models (Linear, Ridge, Lasso, and ElasticNet) to achieve robust and accurate predictions. The use of a Voting Regressor ensured that predictions leveraged the best aspects of each base model.

Textual Data Classification Results

The textual sentiment analysis focused on classifying stock-related news articles into two categories: up (positive impact) and down (negative impact). Various machine learning algorithms were evaluated for their classification accuracy, with the results shown below:

Table 3 Accuracy of model used in Textual analysis

Model	Accuracy
Logistic Regression	65.16
Naïve Bayes	66.72
Random Forest	98.99

Analysis and Insights:

- **Random Forest** emerged as the best-performing model, achieving an impressive accuracy of 98.99%. This underscores its ability to handle complex patterns and feature interactions within the text data.
- Logistic Regression and Naive Bayes, while simpler models, achieved moderate accuracy levels, making them less reliable for this dataset.

Visualizations:

- A **word cloud** of the dataset revealed the most frequently occurring terms, such as "profit," "growth," and "loss," which are critical for stock prediction.
- The **ROC Curve** and **Confusion Matrix** provided a deeper understanding of model performance, showing high precision and recall for Random Forest.

Discussion:

Textual sentiment analysis provided valuable insights into market sentiment, which, when integrated with numerical data, enhanced the hybrid model's predictive power.

Hybrid Model Performance : The hybrid model integrated numerical predictions from the ensemble regression model with sentiment scores derived from textual classification. This approach provided a holistic view of stock price movements by combining two distinct yet complementary data sources.

Real-World-Implications:

This methodology aligns with real-world scenarios where stock prices are influenced by both quantitative factors (historical data) and qualitative factors (news sentiment). It demonstrates the practical benefits of merging numerical and textual data for stock market analysis.

Table 4 Output Metrics:

Date	Ticker	Linear Regression	Random Forest	SVM	LSTM
08-04-2025	AAPL	177.7087	188.0567	209.7595	188.4846
09-04-2025	AAPL	175.7226	188.0567	215.3265	186.4521
10-04-2025	AAPL	174.6275	188.0567	215.6803	186.6734

Below is the **predicted stock price**, which represents the forecasted value for the specified stock symbol. The prediction is based on the selected model and updated with the latest market data and sentiment analysis.

Table 5 Output Stock Prediction:

Model	MSE	R ²	Precision	Recall	F1
Linear Regression	0.0061	0.8940	1	1	1
Random Forest	0.0014	0.9743	1	1	1

SVM	0.0070	0.8774	0.9473	1	0.973
LSTM	50.9961	0.7953	0.9473	1	0.973

Conclusion and Future Scope Using machine learning techniques, the study developed an ensemble regression model to forecast stock prices and a classification model to analyze news sentiment, ultimately merging both approaches in a hybrid model for more accurate predictions.

REFERENCES

1. IET Research. (n.d.). *Advanced stock price forecasting*. Wiley Online Library. <https://doi.org/10.1049/cit2.12052>
2. MDPI. (n.d.). *Dealing with nonlinearity in stock prices*. <https://doi.org/10.1186/s40854-023-00519-w>
3. MDPI. (n.d.). *A hybrid model to predict stock closing price*. <https://doi.org/10.3390/data7050051>
4. Uehara, K. (2016). *Deep Learning for Stock Prediction Using Numerical and Textual Information*.
5. Li, Y., & Pan, Y. (2021). *A Novel Ensemble Deep Learning Model for Stock Prediction Based on Stock Prices and News*.
6. Mozaffari, L., & Zhang, J. (2024). *Predicting Stock Prices: Strategies of Ensemble Learning with Transformer, ARIMA, and Linear Regression Models*.
7. (2024). *Deep Learning for Stock Market Prediction Using Sentiment and Technical Analysis*.
8. (2022). *Predicting Stock Market Using Natural Language Processing*.
9. Uehara, K. (2022). *Stock Price Prediction by Deep Neural Generative Model of News Articles*.