SUBMITED BY: Harsh Kakkad

MSC DATA SCIENCE

Machine Learning Approaches for Predicting Stock Prices Using News Sentiment and Historical Market Data

Date of Submission :- 30 August 2024

Contents:

Table of Contents

Table of figures

Table of Tables

ABSTRACT

This Dissertation investigates the predictive capabilities of a hybrid model integrating traditional market data analysis with sentiment analysis from financial news to forecast Tesla stock prices. By leveraging a fine-tuned DistilRoBERTa model, the study successfully extracted sentiment scores from over 30,000 financial news articles, which were then incorporated into a Long Short-Term Memory (LSTM) network. The hybrid model demonstrated superior performance compared to traditional models, such as Random Forest and Gated Recurrent Unit (GRU), particularly in capturing temporal patterns and improving prediction accuracy. The optimal sequence length for data input was determined to be 5 days, highlighting the importance of recent historical data in stock price forecasting.

Despite the promising results, the study encountered several limitations. Data quality issues arose due to incomplete news summaries, which led to a reliance on headlines for sentiment extraction. Additionally, the Random Forest model experienced overfitting, reducing its generalizability, while the GRU model exhibited a lag in predictions, delaying its response to market changes. These limitations suggest that while the hybrid approach is effective, it requires further refinement to enhance its robustness and real-time applicability.

Future research should focus on expanding data sources, including additional sentiment indicators like social media, to improve sentiment analysis accuracy. Moreover, optimizing model architectures through hyperparameter tuning and exploring ensemble methods could enhance predictive performance. Finally, implementing real-time prediction frameworks and extending the analysis to other stocks and market indices could provide a broader understanding of market dynamics and improve the practical utility of the model for investors.

INTRODUCTION

In today's rapidly evolving global landscape, a nation's economic standing is closely tied to its economic stability, which in turn hinges on the performance of its businesses. One of the most significant indicators of a country's economic health is its stock market. Stock indices provide an immediate and accurate reflection of the state of the economy, responding instantly to new information. Whether it's economic data releases or quarterly earnings reports from major corporations, stock markets swiftly incorporate and reflect all available information as soon as it becomes public.

Stocks represent partial ownership in companies, and investors purchase them to provide capital to these companies in exchange for equity. In this context, the stock market wields substantial influence over economies, making stock price prediction a critical area of research and investment. Accurate stock price predictions can lead to significant profits, underscoring the importance of this endeavor in both academic and financial circles.

There are thousands of ways to predict the price of stock. Predictions can range from daily forecasts and monthly forecasts to second-by-second forecasts. In this dissertation, I will take a more day-to-day approach by considering the closing values of a specific stock, in this case Tesla, for 2516 days. After that, by taking news data related to these particular stocks, I can provide a comparative analysis of the sentiment expressed in news and stock trends of the same corresponding periods. This entire approach is highly interpretive of stock trends and would be used for prediction with sound reasoning.

This thesis aims to identify the effectiveness of combining sentiment analysis from news data with historical stock prices in predicting future movements of stock prices. By doing so, the study explores the potential to enhance the accuracy and reliability of stock price predictions by leveraging both quantitative and qualitative data sources. The integration of sentiment analysis adds a layer of understanding to traditional time series models by incorporating the market's psychological and behavioral factors.

The following are the study's specific objectives:

- Collect daily opening, closing, volume, and other relevant stock price data for Tesla over almost 7 years. This data will then be analyzed using different time-series forecasting methods to establish baseline models based solely on historical market data.

- Gather news articles from Alpaca Market news data about the selected stocks for a one-year period and analyze the sentiments expressed in these news articles. The sentiment analysis will involve determining whether the news content is positive, negative, or neutral and quantifying these sentiments numerically.

- Develop predictive models that incorporate both historical stock data and sentiment analysis results. This will involve selecting out of multiple models from the time series, comparing their results, and integrating the most effective models into a comprehensive hybrid predictive model.

- Evaluate the performance of the integrated models in predicting stock prices relative to traditional methods that rely solely on historical market data. This evaluation will help determine the added value of incorporating sentiment analysis into stock price prediction.

The quality and depth of this research enable it to significantly enhance the accuracy of predicting stock prices by utilizing not only quantitative historical data but also qualitative sentiment data. This approach aims to present a more reliable and holistic understanding of stock market dynamics, which will ultimately assist investors in making sound decisions. Additionally, it explores the various factors that influence stock prices, providing a comprehensive analysis that can be valuable for future economic and financial studies.

Integrating sentiment analysis is particularly significant because it captures the psychological and behavioral aspects of market participants that traditional quantitative models often overlook. News sentiment reflects public opinion, market mood, and potential events that may not yet be evident in historical price data. By combining these qualitative insights with robust time series analysis, this research aims to develop a hybrid model that not only predicts stock prices with higher accuracy but also offers deeper insights into market mechanisms.

Moreover, this research acknowledges the substantial impact that news, readily available to all investors, has on stock prices. Studies that consider the sentiment of news articles and its influence on stock trends are likely to yield better results compared to analyses based solely on historical stock data. This type of comprehensive study provides a more nuanced understanding of the stock market, facilitating more informed investment decisions.

In conclusion, the integration of sentiment analysis with traditional time series models represents a significant advancement in the field of stock price prediction. This research aims to bridge the gap between quantitative data and qualitative sentiment, offering a comprehensive tool for investors and analysts to navigate the complexities of the stock market. By doing so, it contributes to a more robust and reliable framework for understanding and predicting stock price movements, ultimately enhancing the decision-making process for investors and financial analysts alike.

Literature Review

When discussing the dynamic realm of stock markets, numerous exchanges capture global attention, from the NSE and BSE in India to the NYSE and NASDAQ in the USA, alongside the London Stock Exchange in the UK. According to Investopedia (2024), the year 2023 witnessed approximately 80 major stock exchanges worldwide, dispersed across continents and nations. This global landscape reflects a thriving environment where entrepreneurs establish companies, drawing investments from individuals seeking profitable returns. Notably, Investopedia reports that about 55,214 companies are listed on these exchanges, highlighting both the diversity and the inherent risks investors face (Investopedia, 2024) .

The sheer volume of companies listed across global exchanges underscores the critical need for meticulous market analysis. Investors must navigate through intricate financial landscapes to make informed decisions. This involves analysing diverse factors influencing stock prices: economic indicators, company performance metrics, market trends, geopolitical developments, and public sentiment (Fama, 1970) .

Investors employ various strategies for stock market analysis, leveraging fundamental, technical, and quantitative approaches. Fundamental analysis scrutinizes a company's financial health—such as revenue, earnings, and growth prospects—to assess its intrinsic value. Conversely, technical analysis focuses on statistical patterns derived from trading activities, like price movements and trading volumes, to predict future market trends. Quantitative analysis uses mathematical algorithms and models to process financial data, potential trading opportunities.

In recent years, machine learning techniques have revolutionized stock market analysis by processing extensive historical data and identifying intricate patterns that traditional methods might overlook. This approach enhances investors' understanding of market dynamics and augments prediction accuracy (Fischer and Krauss, 2018a).

Moreover, sentiment analysis has introduced a qualitative dimension to stock market assessments. By evaluating sentiments expressed in news articles, social media posts, and other textual sources, analysts gauge public sentiment toward companies or the broader market. This sentiment can significantly influence stock prices, underscoring its relevance in investment decisions (Tetlock, 2007a). Integrating sentiment analysis with historical market data provides a comprehensive view, empowering investors to make well-informed decisions (Bollen, Mao and Zeng, 2011a).

In conclusion, the diversity of global stock exchanges and the multitude of listed companies present both opportunities and risks for investors worldwide. The advancement of analytical methodologies—such as machine learning and sentiment analysis—has bolstered the ability to predict stock market trends and manage investment risks effectively. By joining quantitative data with qualitative insights, investors can gain a much better understanding of market dynamics, enabling more informed investment strategies.

Quantitative Approach

For this research, the aim is to develop 2 models using a Quantitative approach and a Qualitative approach. The first one is Quantitative approach where the market data such as stock opening and closing values, trading volumes, and other relevant metrics are used to predict the future values of the stocks by applying different methods. This approach is well-researched and widely adopted in the field of financial analysis. Numerous studies have employed this method, leveraging historical market data to predict stock prices and trends.

(Chen, Leung and Daouk, 2003) conducted a comprehensive analysis of stock market prediction using market data. In their study, the researchers utilized artificial neural networks (ANN) to forecast stock prices. They found that ANNs are effective in capturing complex patterns in financial time series data, which traditional linear models might miss. Their model showed significant potential in improving the prediction accuracy of stock prices, especially in emerging markets. This study focused on the Taiwan Stock Index and applied neural networks to forecast stock prices. The researchers found that neural networks could effectively model the non-linear dynamics of stock prices, providing a valuable tool for investors in emerging markets. Their approach demonstrated that neural networks could outperform traditional linear models by capturing complex interactions in the data.

(Patel *et al.*, 2015) compared various machine learning techniques for stock price prediction in their study, they explored models including support vector machines (SVM), random forest (RF), and artificial neural networks (ANN). Their research highlighted that different machine learning algorithms have varying strengths, with SVM and RF showing superior performance in trend prediction, while ANNs were effective in capturing non-linear relationships within the data. They used a trend deterministic data preparation method to enhance the performance of these models. The results indicated that while all models had their merits, SVM and RF were particularly effective in predicting stock trends due to their robustness in handling diverse datasets, whereas ANNs excelled in recognizing intricate patterns.

(Fischer and Krauss, 2018a) delved into the use of deep learning techniques for financial market predictions in their paper. They employed long short-term memory (LSTM) networks, a type of recurrent neural network (RNN), to model stock price movements. LSTM networks are particularly suited for time series data due to their ability to retain information over long periods, which is crucial for understanding stock market trends. The study compared the performance of LSTM networks with traditional ARIMA models, revealing that LSTMs provided superior prediction accuracy. The authors highlighted the importance of LSTM's memory capabilities in capturing long-term dependencies, crucial for accurate stock market forecasting.

Qualitative Approach

This approach, specifically focusing on sentiment analysis of news articles related to stock performance involves analyzing the sentiment expressed in news articles or other texts to gauge public opinion and its impact on stock prices. Not as much as before but there are still many studies that have explored this methodology, demonstrating the effectiveness of sentiment analysis in predicting stock market trends.

(Bollen, Mao and Zeng, 2011b) conducted a seminal study which analyzed the sentiment of tweets to predict stock market movements. By examining the collective mood on Twitter, they identified a correlation between public sentiment and market performance. Their research demonstrated that shifts in Twitter sentiment could be used as a predictor for stock market trends, offering a novel approach to financial forecasting. The researchers used a mood tracking tool to measure the collective sentiment on Twitter and found that significant changes in public mood correlated with stock market trends. Their findings indicated that Twitter sentiment could serve as a valuable predictor for financial forecasting, providing a real-time gauge of public opinion and its impact on the market.

(Tetlock, 2007b) in his study explored how the tone of news articles affects stock prices. He utilized sentiment analysis to assess the impact of media coverage on investor sentiment and subsequent market reactions. This study highlighted the significant influence of media sentiment on market dynamics. his research investigated the influence of media sentiment on stock prices by analyzing the tone of news articles. He used sentiment analysis techniques to categorize news articles as positive, negative, or neutral and examined their effect on investor sentiment and market reactions. The study concluded that media sentiment plays a crucial role in shaping investor behavior and market dynamics, with negative news often leading to stock price declines and positive news driving increases.

(Nguyen, Shirai and Velcin, 2015) examined the integration of news sentiment analysis with financial indicators in their study, this study combined sentiment analysis of news articles with traditional financial data to predict stock price movements. The researchers developed a hybrid model that integrated sentiment scores from news articles with financial indicators such as trading volume and stock prices. Their results showed that the combined model outperformed those relying solely on either sentiment analysis or financial data, demonstrating the enhanced predictive power of incorporating qualitative information. This is the research approach that we have followed and discussed in later sections.

(Li *et al.*, 2011) explored the use of textual analysis on financial reports in his paper. Li's research focused on extracting sentiment from corporate disclosures and financial statements to predict stock price movements. By analyzing the tone and language used in these documents, he found that textual sentiment could provide insights into future stock performance, particularly when combined with quantitative financial metrics. Li's research focused on the textual analysis of corporate disclosures and financial statements to predict stock prices. Li's study highlighted the importance of qualitative data in financial

analysis, particularly when combined with quantitative metrics to form a comprehensive prediction model.

Hybrid approach

All the approaches we have examined so far are distinct from one another, but they share the same objective: predicting the future value of specific stocks. The approach used in this research is much more similar to the one used by (Nguyen, Shirai and Velcin, 2015), where a hybrid model was developed. However, there are several differences from their approach, which will be highlighted in the results section. Developing a hybrid model requires considerably more trial and error than other types of models because it is often unpredictable how two different models from each approach will interact and potentially outperform the individually best models from each respective approach.

Creating a successful hybrid model involves integrating quantitative and qualitative data to leverage the strengths of both methodologies. The quantitative approach focuses on analyzing historical market data, such as opening and closing values, trading volumes, and other financial indicators. This method benefits from well-established mathematical and statistical techniques, allowing for the identification of patterns and trends based on past performance.

On the other hand, the qualitative approach employs sentiment analysis to assess the emotional tone and opinions expressed in news articles, social media posts, and other textual sources. This method captures the market's psychological and behavioral aspects, providing insights that are not reflected in numerical data alone. By understanding public sentiment, investors can gain an edge in predicting how market perceptions may influence stock prices.

One significant challenge in developing a hybrid model is the need to find the right balance between the two approaches. It involves experimenting with combinations of quantitative and qualitative data inputs, as well as different machine learning algorithms, to identify the most effective blend. The goal is to create a model that capitalizes on the predictive strengths of both approaches while minimizing their individual limitations.

Methodology

In this section, we will go through the comprehensive methodology undertaken to develop a hybrid model for stock price prediction. This approach involves a systematic and iterative process that integrates both quantitative market data and qualitative sentiment analysis from news sources. The objective is to harness the predictive power of time series analysis while incorporating the nuanced insights gleaned from public sentiment.

In the initial phase of the methodology, the primary task was to gather the necessary data to build the models. Reliable market data is crucial for accuracy, and while there are numerous sources available, not all meet the required standards of reliability. To ensure data integrity, I opted to download a CSV file directly from the official NASDAQ website. For the news data, it required a dependable source that provided consistent coverage over a significant period. To fulfil this need, I utilized the Alpaca API, which supplied with approximately one year of reliable and consistent news articles.

In the model-building phase, the focus was on constructing a robust model using historical market data alone. It involves gathering extensive stock price data and applying various time series analysis techniques. The goal is to establish a baseline predictive model that relies solely on past price movements and trading volumes. I explored multiple model options, including traditional statistical methods and advanced machine learning algorithms, to determine the most effective approach for capturing the inherent patterns and trends in stock prices.

Subsequently, I used a large language model that analyses the sentiment of news articles related to the stocks under consideration within the specific time period. This involves collecting and preprocessing a large corpus of text data, which is then subjected to sentiment analysis to derive numerical sentiment scores. These scores quantify the positive, negative, or neutral tone of the news content, providing a valuable additional feature for our predictive model.

The final phase involves integrating the sentiment scores with the market data model to create a hybrid model. This integration aims to enhance the predictive accuracy by incorporating real-time sentiment analysis alongside historical price data. By applying this hybrid model to multiple stocks, we can compare the results and evaluate the added value of sentiment analysis in stock price prediction. This comprehensive approach not only improves the robustness of our predictions but also offers a deeper understanding of how market sentiment influences stock performance.

Data Acquisition

Data acquisition is a critical step in the development of predictive models, particularly in the context of financial markets where the accuracy and reliability of data can significantly impact model performance. For this project, the primary data sources required were historical market data and news articles related to the stock in question. The process of acquiring this data involved exploring multiple avenues to ensure that the information was both comprehensive and trustworthy.

The historical market data, including stock prices, trading volumes, and other relevant indicators, was sourced directly from the official NASDAQ website. Given the importance of data reliability, this choice was made after considering several alternatives. While various online platforms provide stock market data, many of them lack the accuracy or timeliness needed for effective modelling. The NASDAQ's official data, delivered in a downloadable CSV format, provided a solid foundation for constructing the baseline predictive models. This dataset was essential for the time series analysis and the exploration of various modelling techniques.

In parallel with acquiring market data, it was also necessary to obtain consistent and reliable news data to incorporate sentiment analysis into the models. To achieve this, I utilized the Alpaca API, which offered access to nearly one year of continuous news coverage. This API was chosen for its reliability and the breadth of data it provided, ensuring that the sentiment analysis could be based on a robust dataset. However, the quest for news data also led to attempts at web scraping from several financial news websites such as Money control and investing.com. While web scraping presented an opportunity to collect a vast amount of data, it posed challenges in terms of data consistency, legality, and maintenance, which ultimately made it a less viable option compared to the structured and reliable data offered by Alpaca.

A hybrid model was then developed, integrating both sentiment scores derived from the news articles and the historical market data. This approach aimed to enhance the predictive power of the model by combining quantitative data from stock prices and trading volumes with qualitative data from market sentiment. The sentiment scores provided an additional layer of information, capturing market psychology and potential investor behaviour, which are often not reflected in historical price data alone. By using both types of data, the hybrid model sought to achieve a more nuanced and accurate prediction of stock price movements.

Developing the Market Data Model

The first step in our methodology involves creating a predictive model using only market data, focusing specifically on time series analysis. This phase begins with data collection, where we gather historical stock prices from reliable source, in this case, NASDAQ official website data is considered. This data includes daily opening, closing, high, and low prices, as well as trading volumes. Once collected, we preprocess this data to handle missing values, outliers, and inconsistencies. Normalization of the data ensures all features contribute equally to the model, enhancing its reliability.

Next, we perform Exploratory Data Analysis (EDA) to gain insights into the data's structure and patterns. Trend analysis involves visualizing the data through line plots to understand overall trends in stock prices, while identifying any cyclical patterns or seasonality. Statistical analysis helps calculate key metrics such as mean, median, standard deviation, and correlation among different stock attributes.

In model selection and training, we evaluate 2 time series forecasting models: LSTM (Long Short-Term Memory) networks and GRU (Gated Recurrent Unit). These models are chosen based on their suitability for the given data characteristics. And to also compare base machine learning technique, I used random forest Regressor, Where I split the data into training and testing sets, and the selected models are trained on the training data.

Model evaluation involves assessing the trained models using performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). Cross-validation is performed to ensure the model's robustness. Finally, the model is validated on the testing set, comparing its predictions against actual values. The graph is then analysed to check for any patterns indicating model shortcomings.

Developing the NLP Model for Sentiment Analysis

The second phase of our methodology focuses on developing an NLP model to analyse news sentiments related to stocks and convert these sentiments into numerical scores. We began by collecting news articles and headlines using the Alpaca Market API, which provides real-time, stock-specific news data. This ensures that our sentiment analysis is closely aligned with the stocks under consideration.

For sentiment analysis, we employed the DistilRoBERTa-financial-sentiment model, a fine-tuned version of the distil Roberta-base model specifically trained on financial news data available on HuggingFace. This model was chosen for its superior performance and relevance to financial markets. The model was fine-tuned on the Financial Phrase Bank dataset, which contains 4,840 sentences from English-language financial news, categorized by sentiment. Training occurred over five epochs, achieving a final validation accuracy of 98.23% with a loss of 0.1116. The dataset was annotated by 5-8 annotators to ensure high agreement on sentiment labels. This high accuracy underscores the model's effectiveness in capturing the nuanced sentiments present in financial news.

Once the sentiment scores were generated using this model, they were aggregated daily to match the granularity of our market data. The sentiment scores were then normalized to ensure consistency with other financial indicators such as stock prices and trading volumes. This process enabled the creation of a comprehensive time series dataset that integrates both the quantitative aspects of stock performance and the qualitative insights from market sentiment, forming the foundation for our hybrid predictive models.

Integrating Market Data Model and Sentiment Analysis

The final phase involves integrating the market data model and the sentiment analysis model to create a comprehensive hybrid model. This begins with feature engineering, where the time series of market data is combined with sentiment scores based on the date. The data from both sources is meticulously aligned by date to ensure accurate integration. New features that capture the interaction between market indicators and sentiment scores, such as the moving average of sentiment scores, are created. Lagged features of both market data and sentiment scores are included to capture temporal dependencies and delayed effects.

In model development, we select Long Short-Term Memory (LSTM) as the appropriate model, given that previous results indicated its favorable performance, the results of the LSTM and other models are compared in the results section. LSTM is well-suited for handling both numerical and time series data, making it an ideal choice for the hybrid model.

Model evaluation involves using the same performance metrics employed for the market data model: Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and graph representation of predictions. The performance of the hybrid LSTM model is compared against the baseline market data model to assess the added value of incorporating sentiment analysis.

Comparative Analysis and Conclusion

The final step in the methodology involves applying our hybrid model to predict the stock price of Tesla and comparing the results to draw meaningful conclusions. The model, which integrates both market data and sentiment analysis, is applied to Tesla's stock, and performance metrics are recorded for evaluation.

In the comparative analysis, we focus on the performance of the hybrid model in predicting Tesla's stock price. Specifically, we compare the Mean Absolute Percentage Error (MAPE) of the hybrid model against that of the baseline market data model to assess the added value of incorporating sentiment analysis.

Finally, the results are interpreted to extract key insights about the effectiveness of integrating sentiment analysis with market data for Tesla. The analysis highlights how public sentiment influences Tesla's stock performance and offers potential implications for investors and market analysts. Any limitations of the current methodology and potential sources of bias are also identified, along with suggestions for future improvements to the hybrid model.

In conclusion, this methodology outlines a detailed approach to developing a hybrid model that leverages both market data and sentiment analysis for Tesla's stock price prediction. By following these steps, we aim to create a robust and insightful model that enhances traditional market data analysis with valuable sentiment insights. This hybrid approach not only improves prediction accuracy but also provides a deeper understanding of how market sentiment impacts Tesla's stock dynamics.

This concludes the methodology section and leads us into the results section, where the performance of the hybrid model will be analysed and compared in detail.

Result Discussion and Analysis

When interpreting the results of stock price prediction models, it's crucial to understand that the success of these models is not directly correlated with profit earned. The distinction between training phase performance and real-world application is significant. During training, models can often show promising results by learning from historical data; however, real-life scenarios are far more complex and unpredictable. Factors such as market volatility, unexpected news, and broader economic conditions can all influence stock prices in ways that are difficult to capture in a controlled training environment. For instance, research by (Zhang *et al.*, 2018) highlights that models trained on historical data may fail to generalize well when faced with unseen market conditions due to overfitting to past trends. Consequently, while models may perform well during testing, their real-world effectiveness may vary significantly.

One critical aspect often overlooked in stock prediction models is the choice of data. Many researchers use daily open, high, low, and close prices to train their models. However, this approach can be flawed because the high and low prices for a given day are only known after the market closes, making them impractical for real-time prediction. Studies such as those by (Bao, Yue and Rao, 2017) have shown that relying on daily data can introduce lag in predictions, reducing their accuracy for intraday trading. Additionally, using the opening price to predict the closing price of the same day requires a more granular approach, such as using hourly or minute-level data, rather than relying on daily aggregates.

Evaluating these models goes beyond accuracy metrics; it requires a deep understanding of the data used and the model's robustness in real-time scenarios. (Fischer and Krauss, 2018b) emphasize the importance of testing models under different market conditions to ensure they perform reliably. A thorough analysis of each model's performance is necessary to truly gauge their effectiveness in stock market prediction.

Market Data Analysis

For this analysis, Tesla stock data from NASDAQ was utilized, sourced directly from the NASDAQ website. The dataset is clean, with no missing values, and comprises five columns: Date, Close/Last, Volume, Open, High, and Low. To prepare the data for further processing, the 'Date' column was first converted to a datetime format and set as the index, ensuring proper alignment for time series analysis. Dollar signs were then removed from the price-related columns to standardize the numerical values. Although min-max scaling was applied to normalize the data, it did not significantly impact subsequent modelling steps, and thus, the analysis proceeded without this transformation.

Figure 1 - Closing price before scaling

The graph in figure 1 represents Tesla's closing prices over 2,516 trading days, covering nearly nine years. Notably, significant price spikes are observed beginning in 2020, which coincides with the release of the Tesla Model Y. The Model Y, which has since become one of the world's best-selling vehicles, marked a pivotal moment in Tesla's growth trajectory. Following its release, increased investor attention and market interest contributed to a marked rise in Tesla's stock price.

Figure 2 - Closing prices after Min-Max Scaling

Figure 2 presents the same Tesla stock price data after applying min-max scaling. When compared to the unscaled data in Figure 1, the scaled graph exhibits no significant visual differences. This suggests that in this particular case, min-max scaling does not substantially alter the data's representation, indicating that scaling may not be critical for this analysis.

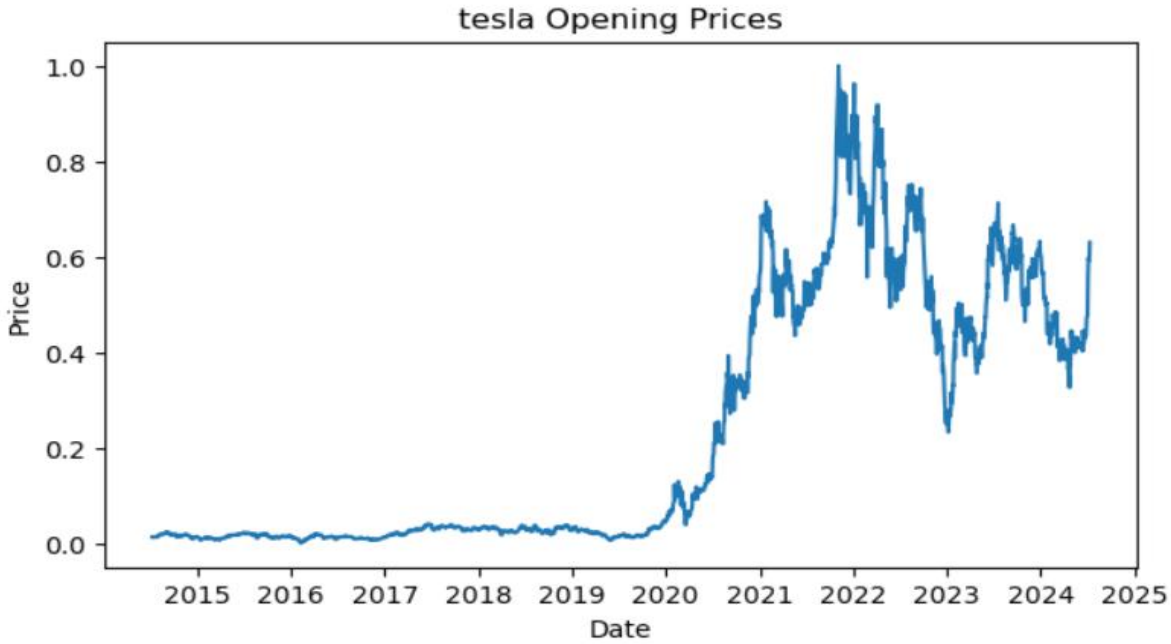

Figure 3 - Opening Prices

Figure 3 displays the opening prices of Tesla stock. When compared with the closing prices shown in Figure 1, the opening and closing values appear to be quite similar. This observation is further validated by the correlation plot presented later in the analysis. Additionally, using opening prices to predict closing prices may not be effective, as the stock prices typically do not undergo significant changes within a single day when viewed in the context of the entire dataset.

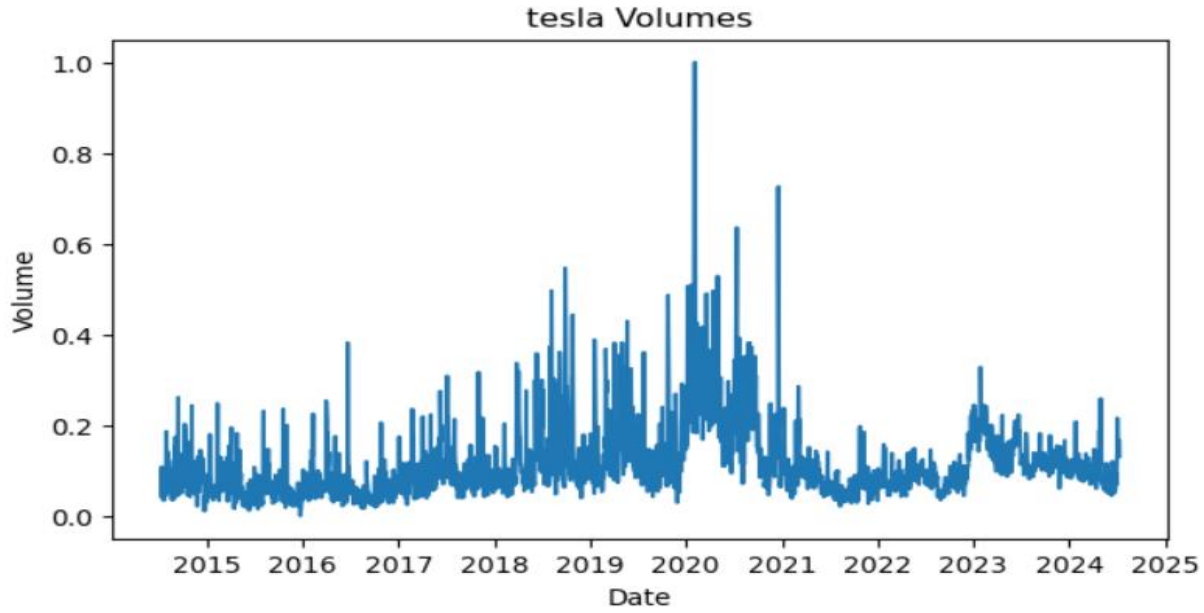

Figure 4 - Volume

Figure 4 illustrates the trading volume of Tesla stock in the market. While trading volume is often considered a positive indicator of market activity and investor interest, it can also have negative implications depending on the context. A deeper analysis is required to determine whether the trading volume reflects positive momentum or indicates potential market concerns.

	Close/Last	**Volume**	**Open**	**High**	**Low**
Close/Last	1.000000	-0.049430	0.999043	0.999444	0.999614
Volume	-0.049430	1.000000	-0.050789	-0.045559	-0.055307
Open	0.999043	-0.050789	1.000000	0.999538	0.999543
High	0.999444	-0.045559	0.999538	1.000000	0.999403
Low	0.999614	-0.055307	0.999543	0.999403	1.000000

Table 1 - Correlation Matrix

This step involves determining the next course of action by analysing the available data. When comparing Figure 4 (Volume) with Figure 1 (Closing prices), there is little apparent correlation. To clarify these relationships, a correlation matrix of all the variables has been generated, as shown in Table 1. Here, the Close/Last value serves as the target variable, and the Open, High, and Low values exhibit a high correlation with the target. This strong correlation is expected, as stock prices typically do not fluctuate significantly from opening to closing within a single day, and the range from high to low remains relatively narrow.

However, predicting stock prices based solely on daily closing values often lacks precision for short-term forecasts and is less valuable for long-term predictions. Short-term predictions tend to have limited accuracy and are heavily influenced by minute-by-minute fluctuations, which would require more granular data, such as hourly or minute-level information. Relying on daily data, especially when Open, High, and Low values are closely related to the Close/Last value, can lead to overfitting, making the model less effective in predicting long-term trends in real-world scenarios. Additionally, the Volume variable does not show a significant correlation with the target variable, making it less relevant for inclusion in the prediction model. Therefore, only the closing prices are used for prediction in this analysis.

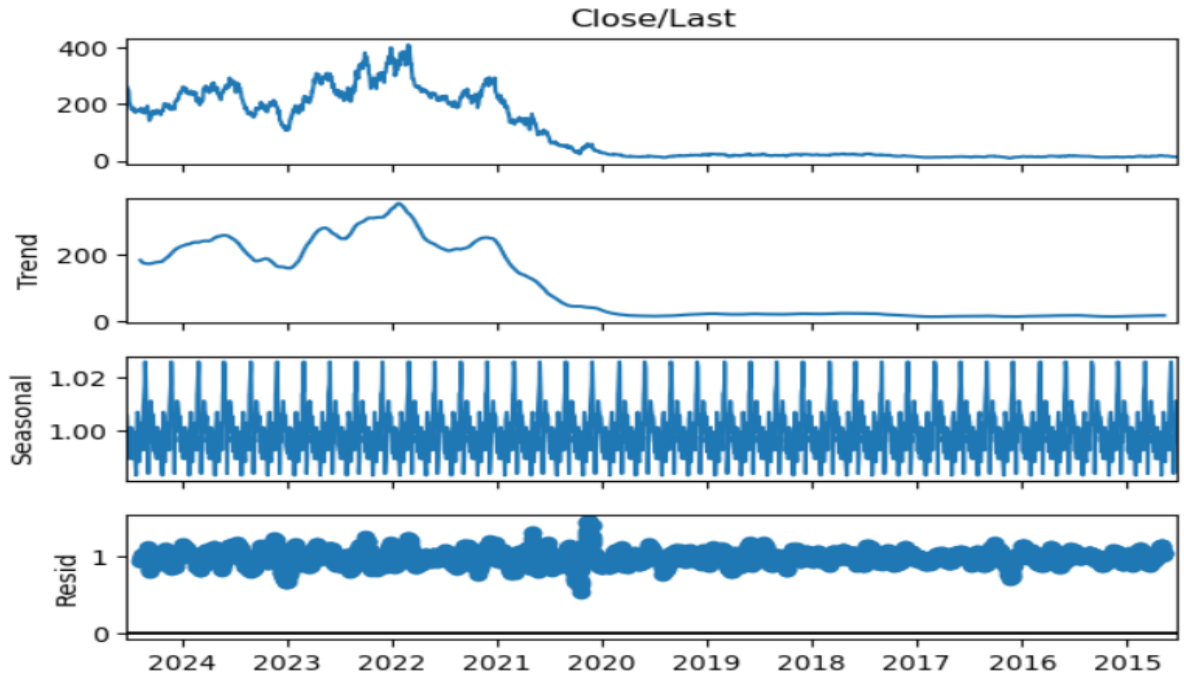

Figure 5 - Decomposition of Closing prices

Within the scope of closing prices, there are various approaches to time series analysis. One such method involves decomposing the time series into three components: Trend, Seasonality, and Residual, as illustrated in Figure 5. A period of 63 was selected for this decomposition, which may seem unconventional but is justified by the fact that stock prices tend to undergo significant changes around the release of quarterly financial results. Excluding weekends, this period equates to approximately 63 trading days. While decomposition could be effective, the high residual component in the plot indicates that it may not be the best method for this analysis.

News Data Analysis

For this project, I utilized news data from the Alpaca API, which included over 30,000 article summaries, headlines, and additional features. Among these features, the summary was identified as the most crucial, as it provides clear insights into the sentiment of each news article. However, upon analysis, it was discovered that none of the articles contained full content, and nearly half lacked summaries altogether.

To optimize sentiment analysis, a logical approach was implemented: a function was created to extract sentiment scores from the summaries of articles that included them. For those articles without summaries, sentiment scores were derived from the headlines instead. Additionally, the date information was extracted from the "created_at" column and converted into a DateTime format to facilitate further analysis.

Model Building

To focus specifically on predicting closing prices, I excluded all columns from the dataset except for the closing prices. The data was then divided into a training set, comprising about 80% of the data, with the remaining 20% reserved for testing across all three models utilized in the analysis.

Given the sequential nature of time series data, I selected the LSTM (Long Short-Term Memory) model, which is particularly adept at capturing long-term dependencies within the data. Additionally, I employed a comparable recurrent neural network (RNN) model, the GRU (Gated Recurrent Unit), which is also effective for time series forecasting. In contrast to these RNN approaches, I included a regression technique, specifically the Random Forest Regressor, to facilitate a comparative analysis.

This section provides a brief overview of the models utilized for prediction, emphasizing their suitability for the task at hand.

LSTM market data model

The Long Short-Term Memory (LSTM) model was applied to the historical market data of Tesla Stock, specifically focusing on predicting stock prices based solely on past price movements. LSTM, a type of recurrent neural network (RNN), is particularly well-suited for time series forecasting due to its ability to retain long-term dependencies and patterns in sequential data, making them ideal for stock price prediction. In preparation for the LSTM model, the data was divided into sequences of 63-time steps, as previously discussed. This sequence length is strategically chosen to capture quarterly trends in the stock data, aligning with the typical financial reporting periods. After this transformation, the training dataset was reduced to 1,953 sequences, down from the original 2,016 values.

The LSTM model architecture was carefully designed with the following specifications:

- Layers: The model consists of two LSTM layers, each with 50 nodes. These layers are responsible for learning the temporal patterns in the stock price data.
- Dense Layers: Two fully connected dense layers were added following the LSTM layers to further process the extracted features.
- Optimizer: The Adam optimizer was used to minimize the loss function. Adam is a popular choice for training deep learning models due to its efficient handling of sparse gradients and adaptive learning rates.
- Loss Function: Mean Squared Error (MSE) was selected as the loss function, which is standard for regression tasks like stock price prediction.
- Training: The model was trained for 10 epochs, which is a moderate number to allow the model to learn effectively while minimizing the risk of overfitting.
- Dropout Layers: To mitigate overfitting, a Dropout layer was added after each LSTM layer, with a dropout rate of 0.2, when making the model second time to compare the results of both the models.

This technique randomly sets a fraction of input units to zero during training, promoting robustness and improving generalization by preventing the model from becoming too reliant on specific nodes.

This architecture was designed to capture the intricate relationships within the historical price data, providing a robust framework for predicting future stock prices. The choice of sequence length, the configuration of the LSTM layers, and the selection of the optimizer and loss function were all tailored to optimize the model's performance on the Tesla stock dataset.

GRU market data model

The Gated Recurrent Unit (GRU) model was applied to the historical market data of Tesla stock, with a focus on predicting stock prices based solely on past price movements. Like LSTM, GRU is a type of recurrent neural network (RNN) designed for time series forecasting, but it simplifies the architecture by combining the forget and input gates into a single update gate. This allows GRU to effectively retain long-term dependencies and patterns in sequential data, making it suitable for stock price prediction.

In preparation for the GRU model, the data was similarly divided into sequences of 63-time steps, as discussed previously. This sequence length was strategically chosen to capture quarterly trends in the stock data, aligning with typical financial reporting periods. After this transformation, the training dataset consisted of 1,953 sequences, mirroring the reduction observed in the LSTM approach, which started with 2,016 values.

The architecture of the GRU model was carefully designed with the following specifications:

- Layers: The model consists of two GRU layers, each with 50 nodes. These layers are responsible for learning the temporal patterns in the stock price data, similar to the two LSTM layers used in the LSTM model. However, GRUs are often faster to train due to their simpler structure.
- Dropout Layers: To mitigate overfitting, a Dropout layer was added after each GRU layer, with a dropout rate of 0.2. This technique randomly sets a fraction of input units to zero during training, promoting robustness and improving generalization by preventing the model from becoming too reliant on specific nodes.
- Dense Layers: Two fully connected dense layers were added following the GRU layers to further process the extracted features, just as in the LSTM architecture.
- Optimizer: The Adam optimizer was employed to minimize the loss function, consistent with the LSTM model.
- Loss Function: Mean Squared Error (MSE) was selected as the loss function, which is standard for regression tasks like stock price prediction. This choice remains the same as in the LSTM model to ensure comparability.

- Training: The model was trained for 50 epochs, different from the LSTM model. This is done as the dropout layer reduces the time and complexity of our model and hence for better prediction 50 epochs are used.

Random forest Regressor for market data

The Random Forest Regressor was employed to predict the historical market data of Tesla stock, focusing specifically on forecasting stock prices based solely on past price movements and related features. Unlike the LSTM and GRU models, which are deep learning techniques that rely on sequential data, Random Forest is an ensemble learning method based on decision trees. This model aggregates predictions from multiple decision trees to improve accuracy and control overfitting, making it suitable for regression tasks.

The architecture of the Random Forest Regressor is characterized by the following specifications:

- Ensemble Method: The model consists of a collection of decision trees, for this case 100 trees were used. Each tree is trained on a random subset of the training data, ensuring diversity in predictions and enhancing overall model robustness.
- Feature Selection: During the construction of each tree, a random subset of features is selected for splitting at each node. This approach helps reduce correlation among the trees and prevents overfitting, which can be a common issue in single decision trees.
- Loss Function: The model uses Mean Squared Error (MSE) as the loss function for regression tasks, enabling straightforward comparison with the LSTM and GRU models.
- Training: The model was trained using the same training set as the LSTM and GRU models, which comprised 80% of the data. Given the nature of Random Forest, training typically occurs in parallel, making it computationally efficient.

The Random Forest Regressor was selected for its ability to complement the predictions made by the LSTM and GRU models. Its ensemble approach offers a different perspective on the data, enabling a comprehensive evaluation of the predictive capabilities when combined with time series analysis. The choice of hyperparameters, along with the standard loss function of MSE, was tailored to ensure optimal performance on the Tesla stock dataset, facilitating direct comparisons with the deep learning models.

DistilRoBERTa model for sentiment analysis

For this research, the “mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis” model, a fine-tuned version of DistilRoBERTa provided by HuggingFace, was employed to perform sentiment analysis on financial news articles. DistilRoBERTa is a compact variant of RoBERTa, known for its efficiency and near-comparable performance to the full model (Sanh *et al.*, 2019) . This model, specifically fine-tuned for the financial domain, is adept at classifying sentiment in financial news as positive, negative, or neutral.

DistilRoBERTa was chosen for its efficiency, retaining 97% of RoBERTa's performance while being smaller and faster (Sanh *et al.*, 2019) . The fine-tuning was conducted on the Financial Phrase Bank dataset, which includes 4,840 financial news sentences. This fine-tuning process is crucial as it enables the model to accurately interpret the nuanced sentiment in financial news, which is critical for predicting stock movements.

While it was an option to develop a custom sentiment analysis model using various NLP techniques, such an approach would likely fall short in accuracy compared to the advanced capabilities of the pre-trained Large Language Model used here. The inherent complexity of financial language, with its domain-specific expressions, demands a model that has been rigorously trained on a vast and relevant dataset. By leveraging a fine-tuned model like the one used here, which has already been optimized for financial contexts, the analysis benefits from a higher level of precision and reliability. This pre-trained model not only saves significant time and resources that would otherwise be spent on developing and training a custom model but also ensures that the sentiment analysis is robust enough to support accurate stock price predictions.

The model in this case analysed over 30,000 news articles related to Tesla, focusing on the summaries for sentiment extraction. In cases where summaries were absent, headlines were used. The daily sentiment scores were averaged to create a sentiment index, which was then integrated with historical market data to enhance the predictive model.

Model Evaluations

In this section, I will evaluate and compare the three models used for market data analysis, aiming to identify the most suitable approach for integration into a hybrid model that will also incorporate sentiment scores. The goal of this comparative analysis is to determine which model most effectively captures the underlying patterns in market data, providing a robust foundation for enhancing predictive accuracy with sentiment analysis.

When evaluating a model, it is crucial to consider the primary objective: selecting the most appropriate model for the hybrid approach. This means that a model's performance cannot be judged solely based on metrics like accuracy or mean squared error (MSE). Instead, it is essential to understand the model's underlying methodology and determine whether its approach is likely to yield the best results when combined with sentiment analysis.

LSTM market data model

The first model to be evaluated is the Long Short-Term Memory (LSTM) model. We ran this model for 10 epochs, with the training process taking approximately 10 minutes to complete. The training was performed using 2,016 closing price values, which were reduced to 1,953 values after applying sequencing. Testing commenced from October 12, 2022, using all subsequent values. In Figure 7 and 8, there is a noticeable gap where the training phase ends, and the testing phase begins. This gap is due to the sequencing process, which spans 63 days. This sequencing step is essential for the LSTM model to effectively learn the temporal dependencies in the market data.

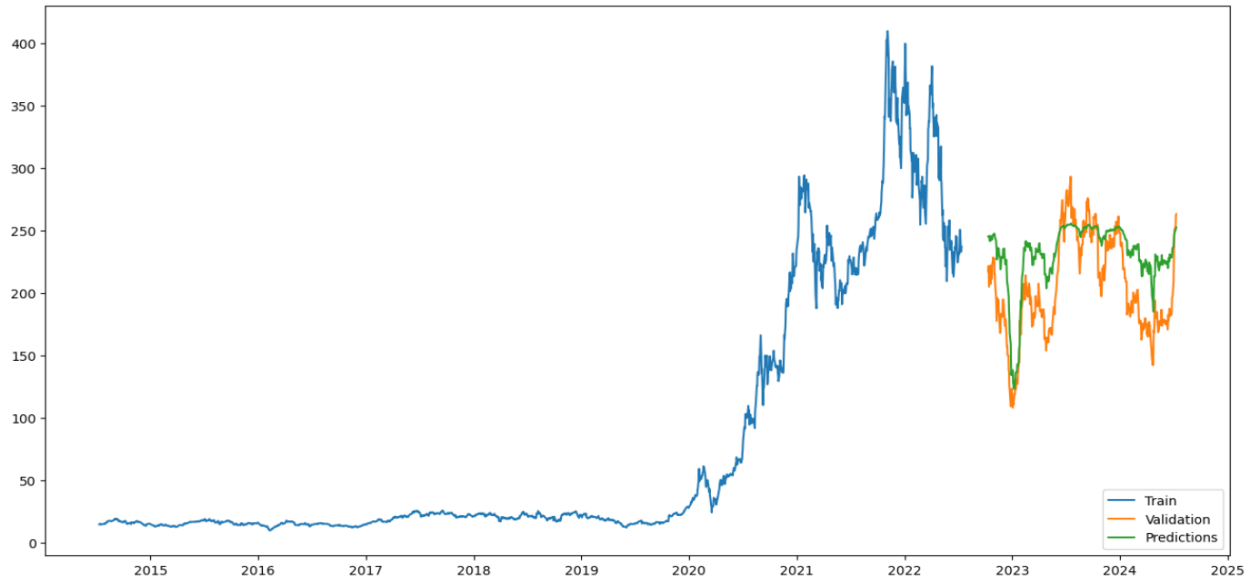

Figure 6 - LSTM market data and predictions vs actual values without dropout

When constructing the LSTM model, a key decision was whether to include dropout layers. Figure 7 represents the model without a dropout layer, while Figure 8 shows the model with a dropout layer. Although the overall differences between the two results may appear subtle, a closer examination reveals that Figure 7 is more effective at predicting market drops, whereas Figure 8 is better at identifying peaks. This suggests that while the LSTM model is actively attempting to make predictions, albeit with some inaccuracies, it demonstrates considerable potential for further refinement and analysis.

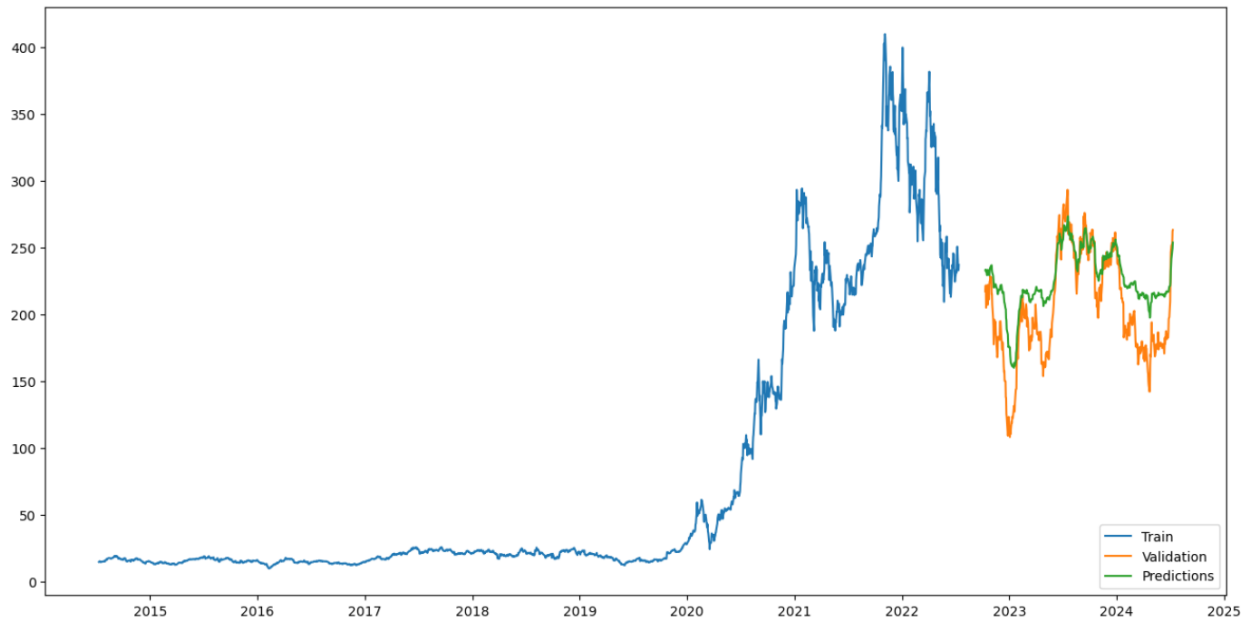

Figure 7 - LSTM market data and predictions vs actual values with dropout layer

However, graphical analysis alone is not sufficient for a comprehensive evaluation. To quantify the model's performance, we calculated the Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) for both versions of the model, as shown in Table 2. Comparing these error metrics reveals a more significant difference between the two models: all three types of errors are notably reduced when the dropout layer is included, indicating that the dropout-enhanced model is better suited for our use case.

LSTM model	MSE	MAE	MAPE
With Dropout Layer	1231.81	30.30	25.13%
Without Dropout Layer	854.60	24.05	22.79%

Table 2 - LSTM error Table

It's important to note, however, that the choice of the best model can vary depending on the specific goal. For example, if the primary objective were to predict market drops, the first model (without dropout) might be more appropriate. Additionally, the MAPE value, even in the best case, is 22.79%, which is relatively high for stock price prediction, indicating room for improvement.

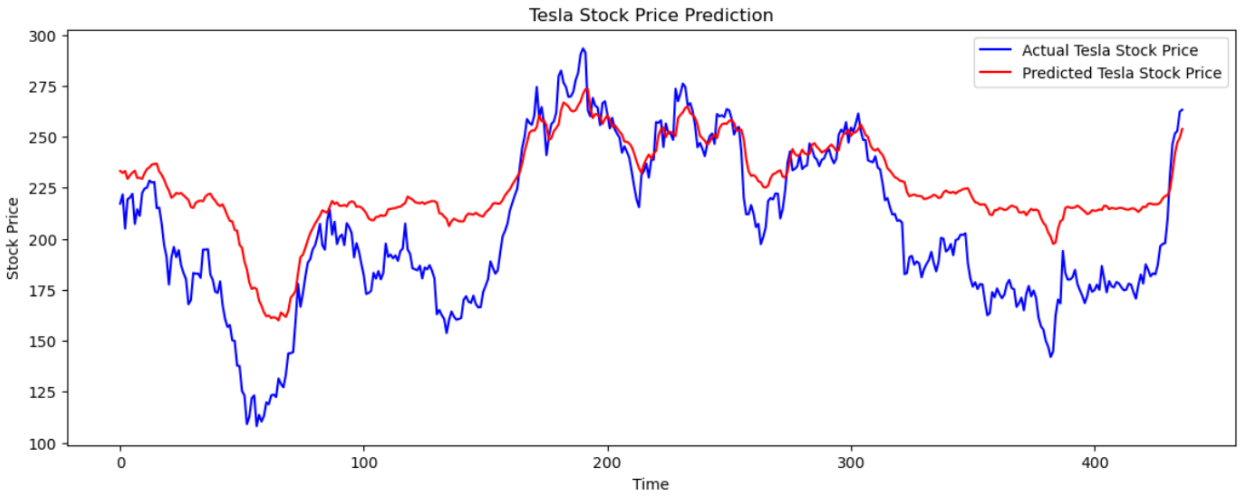

Figure 8 - LSTM predictions vs actual prices

Figure 9 presents a graph that compares the predicted prices with the actual prices during the testing phase. One critical aspect to observe is the alignment between the prediction line and the actual prices line specifically, that there is no noticeable lag in the prediction line. This alignment is crucial for the model's effectiveness, as a lag would indicate a delay in the model's responsiveness to market changes, potentially reducing its predictive accuracy and real-world applicability. The significance of this alignment and its implications for the model's performance are explored in greater detail in the next section.

GRU market data model

After evaluating the LSTM model, we turned our attention to a similar approach: the Gated Recurrent Unit (GRU) model. GRU and LSTM models share similarities, but their effectiveness can vary depending on the specific data and circumstances. While LSTM may outperform GRU in some cases, GRU can yield better results in others.

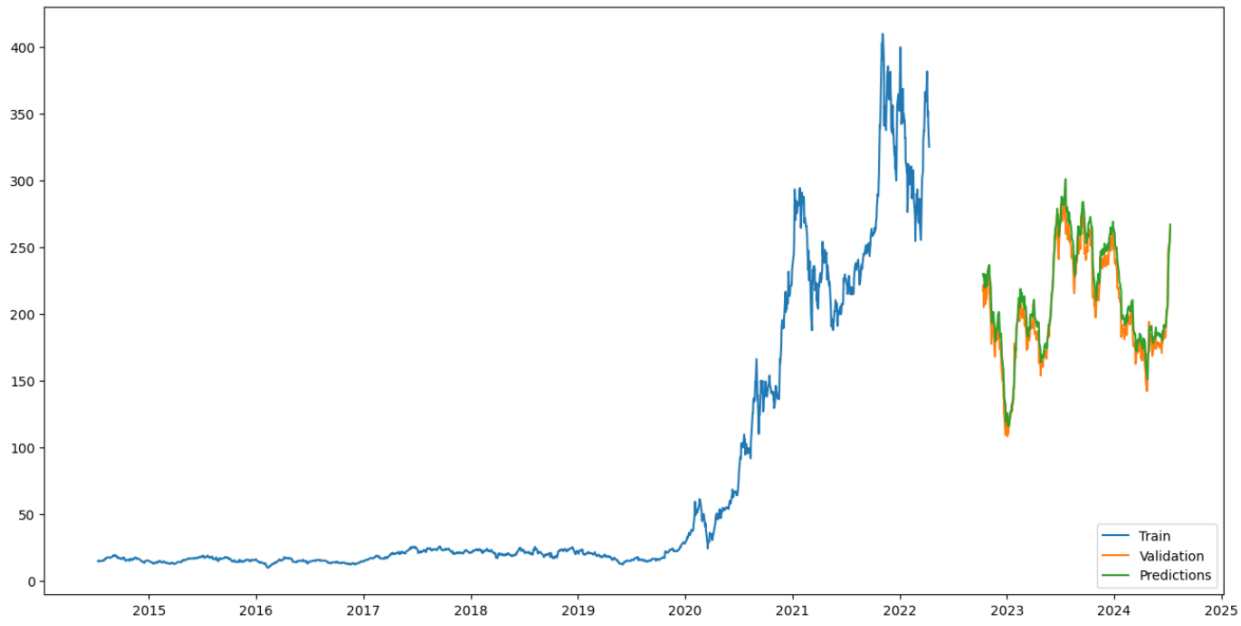

Figure 9 – GRU market data and predictions vs actual values

GRU model errors	MSE	MAE	MAPE
Values	130.72	9.49	4.78%

Table 3 - Errors in GRU model

In contrast to the LSTM model, which took 10 minutes to run 10 epochs, the GRU model was significantly faster, taking only 5 minutes to complete 50 epochs, with a dropout rate of 0.2. As illustrated in Figure 10, the predicted values (green line) closely align with the actual values (yellow line), suggesting very low error rates. This observation is supported by the data in Table 3, where the Mean Absolute Percentage Error (MAPE) is just 4.78%, and other error metrics are also much lower than those for the LSTM model. At first glance, this would suggest that the GRU model is the best option.

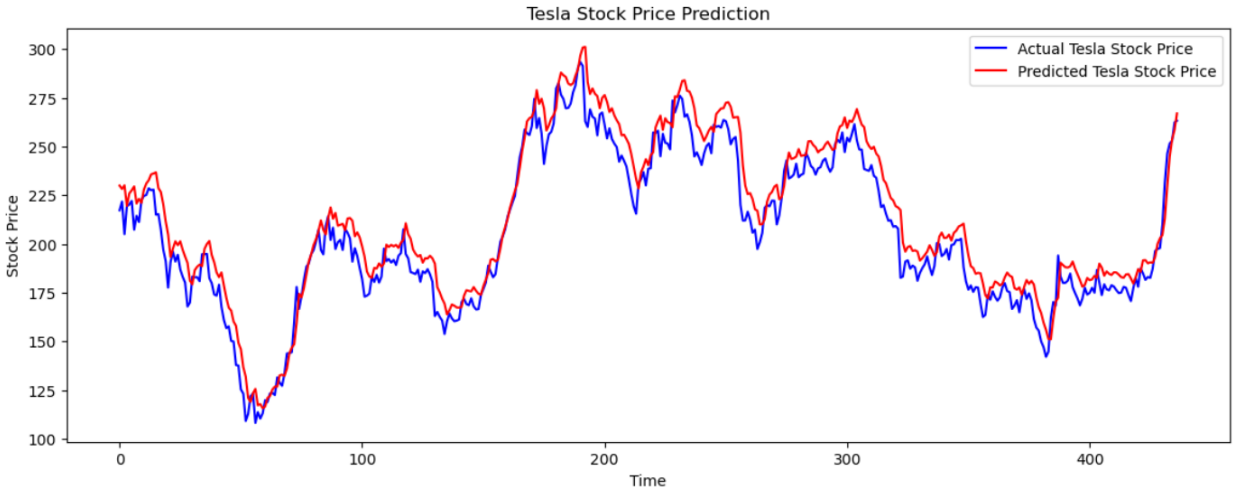

Figure 10 - GRU test data predictions vs actual

However, there is a critical issue to consider, as highlighted in Figure 11. A closer examination of this figure reveals a noticeable lag in the predicted values compared to the actual values. Although this lag does not technically increase the model's error metrics, it presents a significant problem in practical applications. The lag indicates that the GRU model places too much emphasis on the most recent data point, failing to adequately learn from previous trends. This makes the GRU model less suitable for our use case, where timely and trend-sensitive predictions are essential.

Random Forest Regressor model

After analyzing the two RNN models (LSTM and GRU) to understand temporal patterns, it is logical to explore a more traditional machine learning approach. For this, I employed a Random Forest Regression model, using Tesla stock's closing prices and volume data to predict future values. Unlike the RNN models, which took 5-10 minutes to train, the Random Forest model ran in just seconds. On paper, the results appear

to be the best among the three models, as illustrated in Figure 12. Additionally, when comparing error metrics in Table 4, the Random Forest model seems to outperform the RNN models.

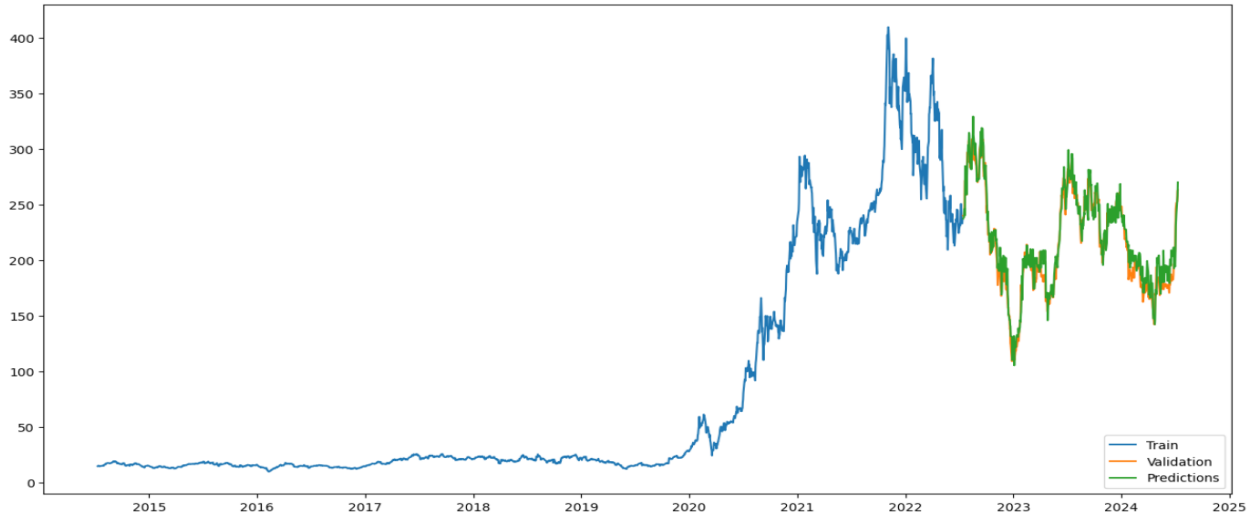

Figure 11 - Random Forest Regression market data and predictions vs actual values

Random Forest Regression model errors	MSE	MAE	MAPE
Values	134.94	9.04	4.53%

Table 4 - Random Forest Regression Errors

However, despite these promising results, the situation is "too good to be true." Unlike the GRU model, the Random Forest model does not exhibit significant lag, as observed in Figure 13. But this apparent accuracy is misleading. The model fails to understand the temporal patterns in the data and instead produces extreme values where possible, without effectively learning from past trends.

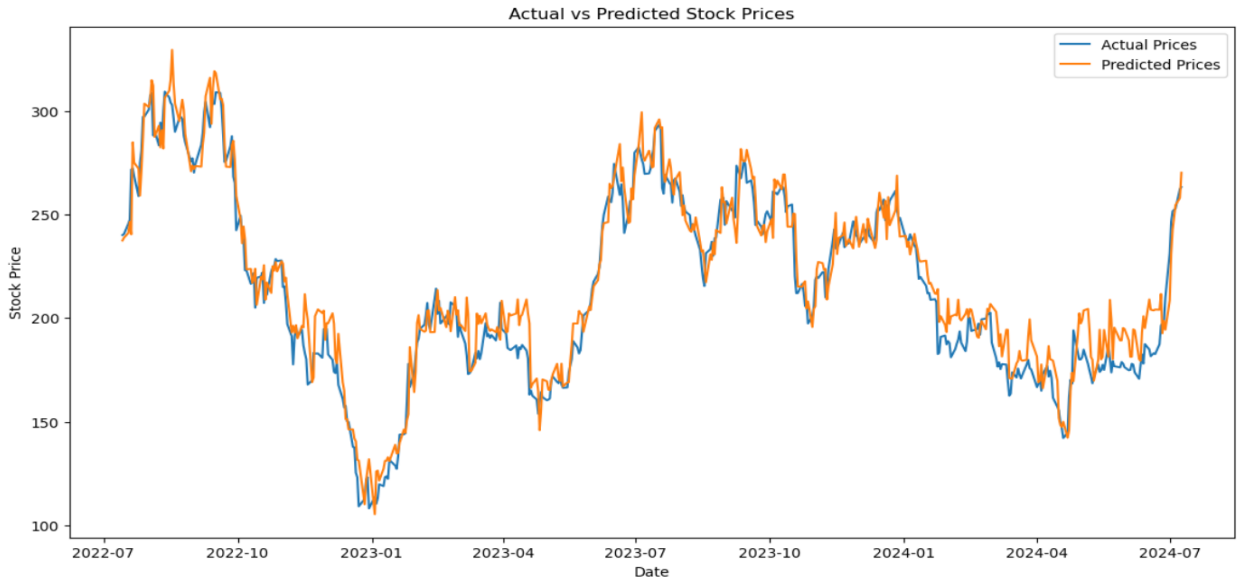

Figure 12 - Random Forest Regression actual vs predicted values

This is evident in Figure 13, where the model fails to make accurate predictions throughout the entire 500-day period. Moreover, the model is unable to predict smaller fluctuations, and larger changes are often lagged by a day. This behavior indicates that the model is overfitting to the data, focusing too much on certain features while ignoring the temporal context. As a result, the Random Forest model is not a suitable choice for a hybrid approach, as it does not adequately capture the nuances necessary for accurate stock price prediction.

Distil RoBERTa model

Hugging Face is an open-source community offering a variety of models tailored for different purposes. For this use case, I utilized a DistilRoBERTa model that had been fine-tuned on financial news, which delivered excellent results for sentiment analysis.

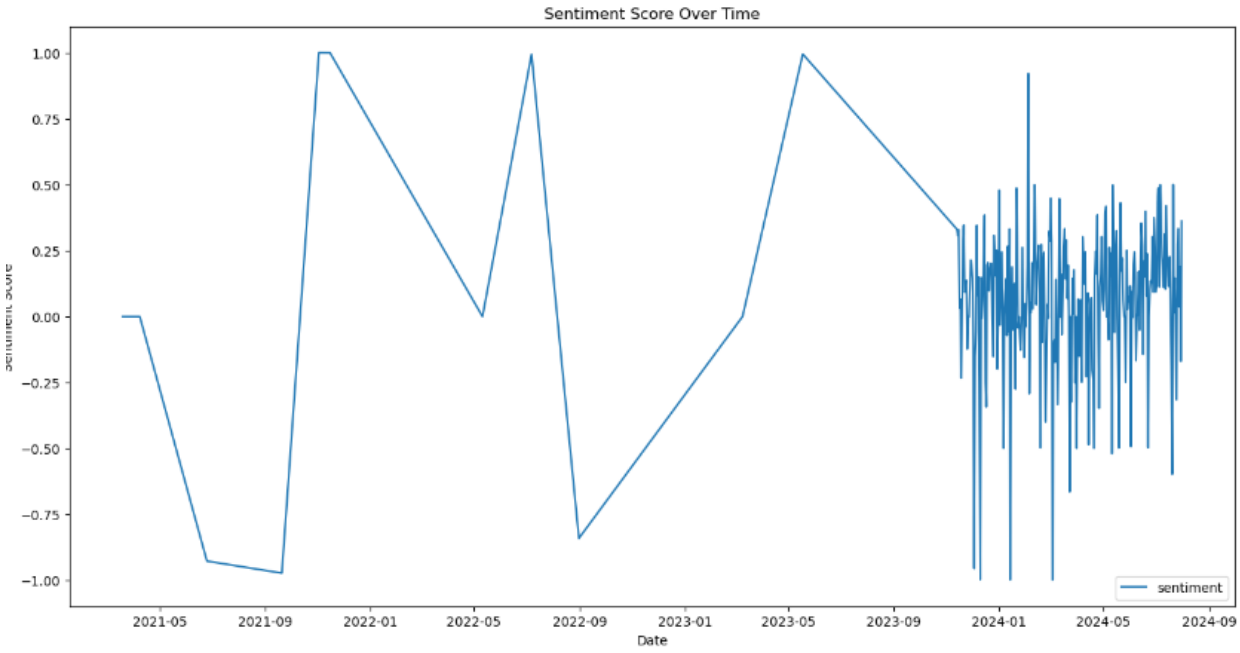

Figure 13 - Sentiment scores over time for Tesla

After obtaining sentiment scores using the model discussed in the previous section, I observed that there were multiple articles published each day. To simplify the analysis, I averaged the sentiment scores for each day. The resulting dataset provided a clearer picture for subsequent steps, as illustrated in Figure 13. However, I noted that the number of sentiment scores was quite limited over an extended period. Consequently, I decided to remove the first 11 sentiment scores to maintain the viability of the analysis. Including every data point could have hindered the results and compromised the overall quality of the analysis. As a result, the final dataset consists of sentiment scores spanning 259 days.

Hybrid Model

After understanding the evaluation of all the 3 market data models and getting the sentiment scores, I made a hybrid model using LSTM approach as it was the best approach for our use case as we have discussed previously. So, to make that I got the sentiment scores of all 259 days and applied inner join with the market data, to obtain 179 rows of data as market is closed on Saturdays and Sundays. I even scaled the data to get better LSTM results.

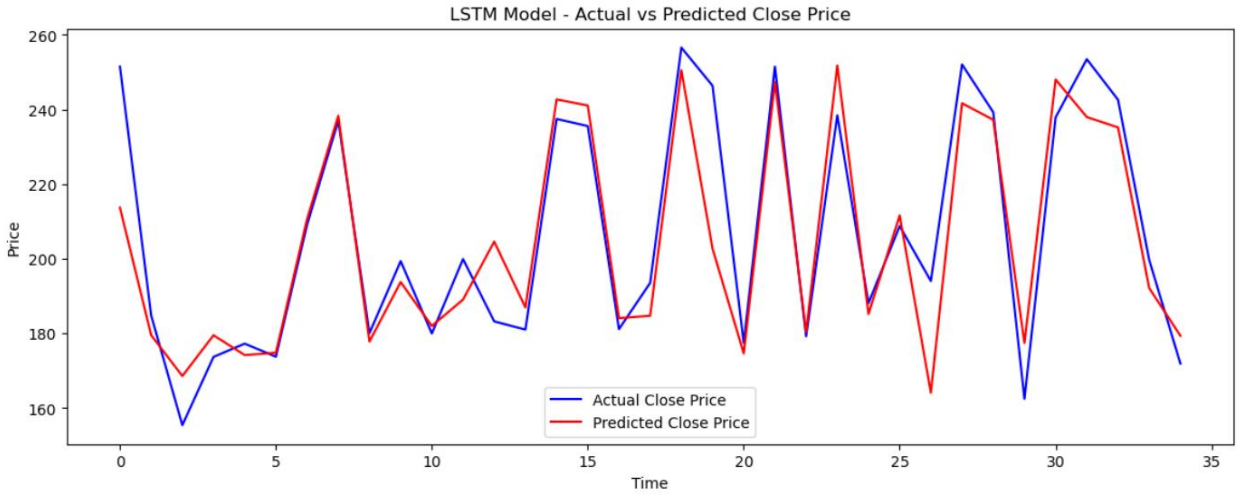

Figure 14 - Actual closing price vs Predictions of Hybrid model with 5 sequence length

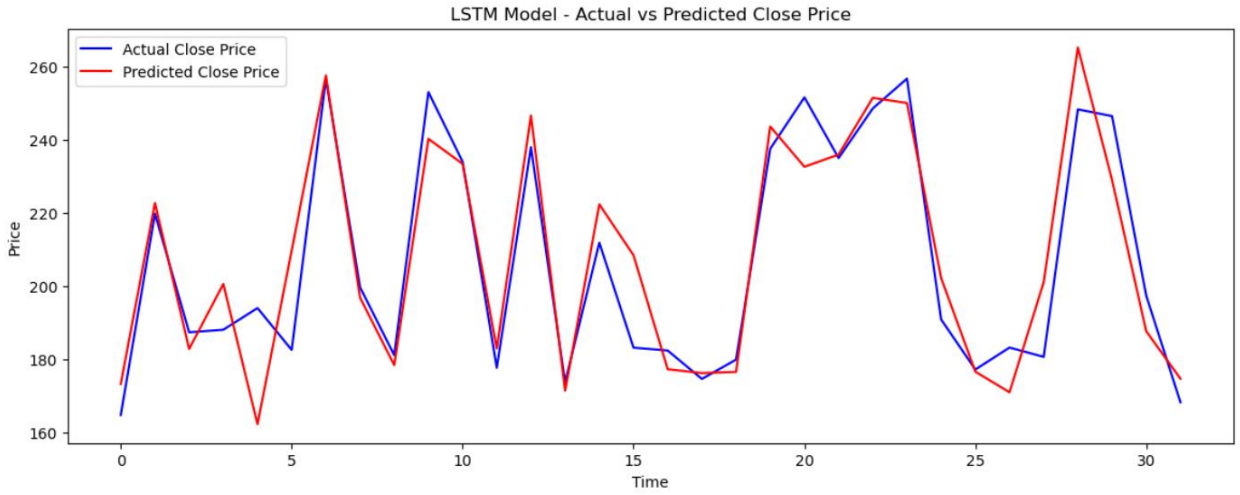

Figure 15 - Actual closing prices vs predictions of hybrid model with 20 sequence length

LSTM Hybrid Model	MSE	MAE	MAPE
With sequence length 5	174.47	9.04	4.31%
With sequence length 20	152.66	9.47	4.71%

Table 5 - Hybrid model errors

Next, I applied the LSTM model on Closing prices, Volume and Sentiment Scores where this time I kept sequence length as 5 as there is smaller data and the working week for market is of 5 days, also I then tried with 20 days as well, which is approximate number of days market is open in a month. Both the results are shown in figure 14 and figure 15. The results look a bit similar but the sequence length 5 has more test size

than sequence length 20 which would explain the result of errors given in table 5, where there is low MAE and MAPE for figure 14 but higher MSE as there are more results to add into the calculation. Basically, sequence length 5 is giving better results than the sequence length 20.

Comparative Analysis of Model Performance

In this section, I compare the performance of the three market data models: LSTM, GRU, and Random Forest Regression alongside the hybrid model that integrates sentiment analysis. The goal is to identify the most effective model for predicting stock prices in our hybrid approach.

- The LSTM model with dropout layers outperformed the version without dropout in terms of error metrics, especially in reducing MSE, MAE, and MAPE. However, the non-dropout model was better at capturing market drops, while the dropout model excelled in predicting peaks. Despite this, the MAPE remained relatively high (22.79% at best), indicating room for improvement. While the LSTM model shows promise, particularly with dropout layers, the relatively high MAPE suggests that it may require further tuning for better predictive accuracy.
- The GRU model produced lower error metrics across the board compared to the LSTM model, with a notably lower MAPE of 4.78%. However, a critical issue with the GRU model is the noticeable lag in predictions, which, while not affecting error metrics, suggests that the model is overly reliant on the most recent data point and fails to capture longer-term trends effectively. The GRU model offers strong predictive accuracy on paper, but its practical applicability is limited by the lag in predictions, making it less suitable for applications requiring timely and trend-sensitive forecasts.
- The Random Forest model yielded the lowest error metrics among the three models, with a MAPE of 4.53%. However, this accuracy is misleading as the model fails to capture temporal patterns and often predicts extreme values, resulting in poor prediction quality over the entire testing period. Despite its low error rates, the Random Forest model's inability to effectively learn from past trends and its tendency to overfit make it unsuitable for use in the hybrid model.
- The hybrid model with a sequence length of 5 showed better results in terms of MAE and MAPE but had a higher MSE due to the larger test size. The sequence length of 20, while slightly better at reducing MSE, had higher MAE and MAPE, indicating less accuracy in individual predictions. The hybrid LSTM model with a sequence length of 5 days appears to be the most suitable approach for integrating sentiment analysis, offering a balanced trade-off between prediction accuracy and capturing temporal patterns effectively.

Models	**MSE**	**MAE**	**MAPE**
LSTM (No Dropout)	854.60	24.05	22.79%
LSTM (With Dropout)	1,231.81	30.30	25.13%
GRU	130.72	9.49	4.78%

Random Forest	134.94	9.04	4.53%
Hybrid LSTM (Seq 5)	174.47	9.04	4.31%
Hybrid LSTM (Seq 20)	152.66	9.47	4.71%

Table 6 - Error comparison for every model

This comparative analysis helps in determining that while the Random Forest model appears optimal in terms of error metrics, it fails to capture the nuances required for reliable stock price prediction. The hybrid LSTM model, particularly with a sequence length of 5, strikes the best balance between accuracy and capturing market patterns, making it the preferred choice for our use case.

Conclusion

This research focused on developing a predictive model for Tesla stock prices by integrating market data analysis with sentiment analysis derived from financial news articles. The key findings are as follows:

- Data Collection: Market data was sourced from the official NASDAQ website to ensure high quality, while news data was collected using the Alpaca API.
- Market Data Predictions: Three approaches were applied for time series predictions. The Long Short-Term Memory (LSTM) model yielded promising results, whereas the Gated Recurrent Unit

(GRU) and Random Forest models produced flawed outcomes, which have been addressed accordingly.

- Effective Sentiment Analysis: A fine-tuned DistilRoBERTa model from Hugging Face was employed for sentiment analysis, successfully processing over 30,000 articles. This provided valuable sentiment scores that correlated with stock price movements.
- Hybrid Model Performance: The hybrid LSTM model, which integrated sentiment analysis, outperformed traditional models like Random Forest and GRU, demonstrating enhanced predictive accuracy and a superior ability to capture temporal patterns in stock prices.

In conclusion, this dissertation has successfully demonstrated the effectiveness of developing a hybrid model for predicting stock market prices using sentiment analysis, despite encountering some operational challenges.

Limitations

Several limitations were encountered during this study:

- Data Collection: Collecting news data proved more challenging than anticipated. Many sources are paid, and free options often lack reliability. I attempted to web scrape data from websites such as Money Control and Investing.com, but ultimately, the Alpaca API emerged as the best solution, providing nearly a year's worth of news data, although many articles lacked summaries.
- Model Selection: While LSTM and GRU were straightforward choices, I initially considered using the ARIMA model. However, its performance was subpar, prompting me to choose Random Forest regression instead.
- Practical Considerations: The GRU model exhibited a lag in predictions, resulting in delayed responses to market changes, which could be a significant drawback for real-time trading scenarios. Additionally, the Random Forest model demonstrated signs of overfitting, posing another critical limitation.

Recommendations for Future Research

To enhance the findings and address current limitations, future research should consider the following areas:

- Enhanced Data Acquisition: Expanding the dataset to include more comprehensive and diverse news sources, along with utilizing additional features (e.g., social media sentiment) could improve the robustness of sentiment analysis.
- Model Optimization: Further experimentation with hyperparameter tuning and model architectures, including exploring ensemble methods that combine the strengths of multiple models, could yield better predictive capabilities.

- Real-Time Predictions: Investigating real-time prediction frameworks that integrate live data feeds could help mitigate the lag issues observed in the GRU model and improve the responsiveness of predictions.
- Broader Market Analysis: Extending the analysis to include other stocks and indices could provide a more generalized understanding of market behavior and the influence of sentiment across different sectors.

References:

Bao, W., Yue, J. and Rao, Y. (2017) 'A deep learning framework for financial time series using stacked autoencoders and long-short term memory', *PLoS ONE*, 12(7). Available at: https://doi.org/10.1371/journal.pone.0180944.

Bollen, J., Mao, H. and Zeng, X. (2011a) 'Twitter mood predicts the stock market', *Journal of Computational Science*, 2(1), pp. 1–8. Available at: https://doi.org/10.1016/j.jocs.2010.12.007.

Bollen, J., Mao, H. and Zeng, X. (2011b) 'Twitter mood predicts the stock market', *Journal of Computational Science*, 2(1), pp. 1–8. Available at: https://doi.org/10.1016/j.jocs.2010.12.007.

Chen, A.-S., Leung, M.T. and Daouk, H. (2003) *Application of neural networks to an emerging ÿnancial market: forecasting and trading the Taiwan Stock Index*, *Computers & Operations Research*. Available at: www.elsevier.com/locate/dsw.

Fama, E.F. (1970) *Efficient Capital Markets: A Review of Theory and Empirical Work*, *Source: The Journal of Finance*.

Fischer, T.; and Krauss, C. (2018a) *Deep learning with long short-term memory networks for financial market predictions Standard-Nutzungsbedingungen*. Available at: https://www.iwf.rw.fau.de/research/iwf-discussion-paper-series/.

Fischer, T.; and Krauss, C. (2018b) *Deep learning with long short-term memory networks for financial market predictions Standard-Nutzungsbedingungen*. Available at: https://www.iwf.rw.fau.de/research/iwf-discussion-paper-series/.

Investopedia (2024) *Stock Exchanges Around the World*, *https://www.investopedia.com/financial-edge/1212/stock-exchanges-around-the-world.aspx*.

Li, F., Thank, I., Chen, J., Huang, Y., Merkley, K., Shroff, N. and Tucker, J. (2011) *Textual Analysis of Corporate Disclosures: A Survey of the Literature Forthcoming in the Journal of Accounting Literature*.

Nguyen, T.H., Shirai, K. and Velcin, J. (2015) 'Sentiment analysis on social media for stock movement prediction', *Expert Systems with Applications*, 42(24), pp. 9603–9611. Available at: https://doi.org/10.1016/j.eswa.2015.07.052.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015) 'Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques', *Expert Systems with Applications*, 42(1), pp. 259–268. Available at: https://doi.org/10.1016/j.eswa.2014.07.040.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter'. Available at: http://arxiv.org/abs/1910.01108.

Tetlock, P.C. (2007a) 'Giving content to investor sentiment: The role of media in the stock market', *Journal of Finance*, 62(3), pp. 1139–1168. Available at: https://doi.org/10.1111/j.1540-6261.2007.01232.x.

Tetlock, P.C. (2007b) 'Giving content to investor sentiment: The role of media in the stock market', *Journal of Finance*, 62(3), pp. 1139–1168. Available at: https://doi.org/10.1111/j.1540-6261.2007.01232.x.

Zhang, X., Qu, S., Huang, J., Fang, B. and Yu, P. (2018) 'Stock Market Prediction via Multi-Source Multiple Instance Learning', *IEEE Access*, 6, pp. 50720–50728. Available at: https://doi.org/10.1109/ACCESS.2018.2869735.