Saint Petersburg State University

Faculty of Mathematics and Mechanics

Department of Informatics

Report on practical training on the topic: “Stock price prediction based on multiple data sources and sentiment analysis using a Bert based Model”

Student: Mohammadhossei Khalashi

Group: 21.Б14-мм

Scientific Advisor: Ph.D. Dmitry Grigoriev

Table of Contents

Introduction

In the dynamic and ever-evolving landscape of financial markets, predicting stock prices remains a compelling yet challenging endeavor. The complexity of stock market dynamics has been a subject of academic fascination and practical necessity for decades. Traditional theories such as the Efficient Market Hypothesis (EMH) posited by Fama (1964) and the Random Walk Theory (Malkiel, 1973) suggest that past stock data is not indicative of future performance, painting a picture of the market's unpredictability. However, with advancements in computational techniques and data availability, newer methods have emerged, challenging these early assumptions and offering more nuanced insights into stock market forecasting.

The integration of big data and AI technologies has opened new avenues in financial analysis, particularly in the realm of stock prediction. While historical stock data has always been a key component in market analysis, recent studies have highlighted the significance of non-traditional data sources, such as social media sentiment and financial news, in predicting stock trends (Li et al., 2018; Oliveira et al., 2016; Sun et al., 2017). This recognition has led to the exploration of novel approaches that leverage both quantitative and qualitative data for more accurate predictions.

Against this backdrop, our project introduces an innovative approach to stock price prediction, combining technical indicator calculations with sentiment analysis derived from non-traditional data sources. We focus on an ensemble of three major stocks: GOOG, AMZN, and MSFT, from January 1, 2015, to January 31, 2019. This selection is not arbitrary; it is rooted in the strategy of portfolio diversification, a risk mitigation tactic where investments are spread across various stocks to increase accuracy and reduce potential losses (Markowitz, 1952).

Our project is structured as follows:

1. **Technical Indicator Calculation Model**: We calculate various technical indicators such as Moving Averages, RSI, %K, %R applied to historical price data of our selected stocks.
2. **Sentiment Index Calculation Model**: Leveraging a fine-tuned DistilBERT model, specifically distilRoBERTa-financial-sentiment, we analyze sentiments in a dataset of tweets within our selected timeframe, classifying them as positive, negative, or neutral.
3. **Attention-Based LSTM Model**: We feed the combined dataset of technical indicators and sentiment indices into an LSTM model enhanced with an attention mechanism, culminating in the prediction of the aggregated closing stock prices of our chosen companies.

Our approach is novel in its application of a DistilBERT-based model for sentiment analysis, diverging from previous studies that predominantly used CNNs for this purpose (Li et al., 2018). Additionally, the use of an attention-based LSTM to process and predict stock prices based on this multifaceted data is a relatively unexplored area, offering promising avenues for more accurate stock market predictions.

Our research question revolves around the efficacy of combining technical indicators with sentiment analysis derived from social media and financial news in predicting stock prices. Specifically, we aim to investigate whether this integrated approach can provide more accurate predictions than methods relying solely on historical stock data or sentiment analysis independently.

In summary, our project stands at the intersection of traditional financial analysis and modern AI techniques, representing a novel contribution to the field of stock price prediction. By incorporating diverse data sources and leveraging advanced machine learning models, we aim to enhance the accuracy and reliability of stock market forecasts, providing valuable insights for investors and researchers alike.

Literature review and related works

The field of behavioral finance has long acknowledged the impact of investor sentiment on stock market dynamics. This sentiment, increasingly expressed through social media and financial news platforms, plays a pivotal role in shaping market trends (Statman, 2011). The correlation between stock price movements and investor sentiment has been extensively studied, revealing the influence of non-traditional data sources, such as social media and financial news, on market behavior (Wang, 2017; Vargas et al., 2017). The evolution of sentiment analysis methods has enabled the extraction of valuable insights from these unstructured data sources. For instance, Porshnev et al. (2013) demonstrated the predictive power of integrating Twitter sentiment analysis with traditional market indicators. Furthermore, Sohangir et al. (2018) showed the efficacy of combining sentiment analysis from financial social platforms like Stocktwits with deep learning models for stock prediction. These studies underscore the growing recognition of the importance of sentiment analysis in financial forecasting.

The advent of transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) has revolutionized the field of natural language processing (NLP). BERT's ability to understand the context of words in text by considering the entire corpus has made it particularly effective for sentiment analysis (Devlin et al., 2018). Following BERT's success, DistilBERT, a streamlined version that retains most of BERT's performance while being more computationally efficient, emerged as a powerful tool for NLP tasks, including sentiment analysis (Sanh et al., 2019). Its application in financial sentiment analysis is particularly noteworthy. Studies leveraging DistilBERT have shown its capability to discern nuanced sentiment in financial texts, offering a more precise understanding of investor sentiment and its potential impact on stock markets.

The LSTM (Long Short-Term Memory) network, a variant of recurrent neural networks, is well-suited for handling time-series data like stock market prices due to its ability to capture long-term dependencies (Hochreiter & Schmidhuber, 1997). LSTMs have been extensively used in financial forecasting, with studies demonstrating their ability to model complex patterns in stock data effectively (Liu et al., 2018; Wu et al., 2018). Vargas et al. (2018) illustrated the advantages of LSTM networks when combined with technical indicators for improved stock prediction accuracy. The flexibility of LSTM networks in integrating various types of data, including sentiment indices derived from advanced NLP models, positions them as a crucial component in modern financial forecasting models.

Integrating sentiment analysis with LSTM models represents a novel approach in stock market forecasting. By combining the nuanced understanding of market sentiment provided by models like DistilBERT with the time series analysis capabilities of LSTM networks, researchers can create more robust and accurate prediction models. This integrated approach acknowledges the multifaceted nature of stock market dynamics, where both quantitative data (like historical prices and technical indicators) and qualitative data (such as investor sentiment) play critical roles. Recent studies exploring this integration have shown promising results, indicating a significant improvement in prediction accuracy over traditional methods that consider only historical data or sentiment analysis in isolation.

Methods

Technical Indicator Calculation Model

Technical indicators are vital tools in financial analysis, providing insights into market trends and investor behavior. They are derived from historical price data and are used to forecast future stock price movements. In our project, we focus on several key technical indicators:

- **Moving Averages (MA)**: Moving averages smooth out price data to create a single flowing line, which makes it easier to identify the direction of the trend. The formula for a Simple Moving Average (SMA) is:

  $$SMA=\frac{\left(P_1+P_2+...+P_n\right)}{n}, where\ P_1,\ P_2,\ ...,\ P_n \text{ are the stock prices over } n \text{ periods.}$$

- **Relative Strength Index (RSI)**: RSI is a momentum oscillator that measures the speed and change of price movements. It oscillates between zero and 100. Traditionally, and according to Wilder, RSI is considered overbought when above 70 and oversold when below 30. The formula for RSI is:

  $$RSI=100-\left(\frac{100}{1+RS}\right), where\ RS\ is\ the\ average\ gain\ of\ up\ periods\ during\ the\ specified\ time$$

- **%K and %R**: These are momentum indicators, with %K being the Stochastic Oscillator and %R being the Williams %R indicator. They both measure the level of the closing price relative to the high-low range over a specific period.

The outputs from this model are a set of these technical indicators for each day, which provide a quantitative foundation for our stock price prediction model.

Sentiment Index Calculation Model

Sentiment analysis plays a crucial role in understanding investor attitudes and market trends. We utilize the DistilBERT model, a lighter version of BERT, for analyzing the sentiment of financial texts. The model processes text data from various sources, classifying them into positive, negative, or neutral sentiments. The DistilBERT model is particularly suited for this task due to its efficiency and effectiveness in understanding the context of words in texts. The model's architecture is based on the transformer, a deep learning model introduced by Vaswani et al. (2017), which uses self-attention mechanisms to weigh the significance of different words in a sentence.

Attention-Based LSTM Model

Our approach incorporates an attention-based LSTM model to process the concatenated feature set comprising technical indicators, sentiment indices, and possibly raw historical price data. LSTM networks, introduced by Hochreiter & Schmidhuber (1997), are well-suited for time-series data due to their ability to remember long-term dependencies. An LSTM cell comprises three gates: the input gate, forget gate, and output gate, which control the flow of information.

The attention mechanism in LSTM allows the model to focus on the most relevant parts of the input sequence, enhancing its predictive performance. The mathematical representation of the LSTM with attention can be described as follows:

- **LSTM Layer**: Captures temporal dependencies using the following formulas:

$$f_t = sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

$$i_t = sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$

$$C\ {}_t = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right)$$

$$C_t = f_t * C_{t-1} + i_t * C\ {}_t$$

$$o_t = sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

- **Attention Layer**: Weighs the importance of each timestep's output. The attention score can be calculated as:

$$alpha_{t,t'} = \frac{\exp\left(score\left(h_t, h_{t'}\right)\right)}{\sum_{t''} \exp\left(score\left(h_t, h_{t''}\right)\right)}, where\ \ score$$ is a function (e.g., dot product) used to calculate

the alignment between the input at timestep t $\wedge$ *the output at timestep* t'

- **Fully Connected Layer and Output Layer**: The fully connected layer processes the weighted features, which are then fed into the output layer to produce the final stock price prediction.

Evaluation Criteria

To evaluate the performance of our stock price prediction model, we employ two primary statistical metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE). These metrics are crucial in assessing the accuracy of the model's predictions against actual stock prices.

- **Mean Squared Error (MSE)**: MSE is a common measure of prediction accuracy in regression models. It calculates the average squared difference between the estimated values and the actual value. The MSE is defined as:

$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y^i)^2, where\ Y_i$ represents the actual values, Y^i are the predicted values, and n is the number of observations

A lower MSE value indicates a better fit of the model to the data.

- **Mean Absolute Error (MAE)**: MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's calculated as the average of the absolute differences between the predicted values and observed values. The formula for MAE is:

$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(Y_i - Y^i)\right|, where \left|(Y_i - Y^i)\right| is\ the\ absolute\ error\ between\ the\ actual$ and $the\ predicted\ values,$ and n is the number of observations

These metrics are particularly useful for evaluating forecasting models like ours because they provide a clear quantification of the model's prediction error. MSE is sensitive to large errors due to its squaring term, making it useful when large errors are particularly undesirable. In contrast, MAE provides a more straightforward and intuitive understanding of the average error magnitude.

The choice of these evaluation metrics is aligned with common practices in financial modeling and stock price prediction, where accurate and reliable predictions are crucial (Hyndman & Koehler, 2006)

Data

Historical Stock Data

Our project utilizes historical stock data from three major technology companies: Google (GOOG), Amazon (AMZN), and Microsoft (MSFT), spanning from January 1, 2015, to January 31, 2019. This data was retrieved from Yahoo Finance and serves as the foundational input for our Technical Indicator Calculation Model.

Structure of Historical Data:

- **Date**: The date of the trading session.
- **Open**: The price at which a stock started trading during the opening of the trading session.
- **High**: The highest price at which a stock traded during the trading session.
- **Low**: The lowest price at which a stock traded during the trading session.
- **Close**: The price at which a stock ended trading during the closing of the trading session.
- **Adjusted Close**: The closing price after adjustments for all applicable splits and dividend distributions.
- **Volume**: The number of shares traded during the trading session.

This structured data includes key price points and the volume of trades, all of which are crucial for calculating various technical indicators used in our analysis.

Sentiment Data from Tweets[1]

For sentiment analysis, we employ a dataset comprising over 1.7 million public tweets about Apple, Amazon, Google, Microsoft, and Tesla stocks, published between January 1, 2015, and December 31, 2019. The dataset, hosted on Hugging Face, captures the public sentiment and opinions expressed on these dates. For our Sentiment Index Calculation Model, we focus on the segments of this dataset that pertain to Amazon, Google, and Microsoft.

Structure of Tweet Data:

- **Tweet ID**: A unique identifier for each tweet.
- **Writer**: The username of the Twitter account that published the tweet.
- **Post Date**: The date and time when the tweet was posted.
- **Body**: The text content of the tweet.

This dataset offers a rich source of sentiment-related data, which our DistilBERT-based model will analyze to determine the prevailing sentiment towards each stock on any given day. By quantifying this sentiment, we aim to uncover correlations between public opinion and stock price movements.

[1] Link to the Tweets dataset : https://huggingface.co/datasets/mjw/stock_market_tweets

Experiments and results

Our project's experimental framework is constructed on three core models, each tailored to address distinct facets of stock price prediction. The models are carefully orchestrated to work in synergy, with the output of each serving as an integral input for the subsequent one.

Technical Indicator Calculation Model

Our journey begins with the Technical Indicator Calculation Model. This model ingests historical price data for GOOG, AMZN, and MSFT, spanning from January 1, 2015, to January 31, 2019. From this dataset, we calculate a suite of technical indicators that serve as proxies for market sentiment and trends. These indicators include Simple Moving Averages (SMA_30), Exponential Moving Averages (EMA_30), the Relative Strength Index (RSI), Stochastic Oscillator (%K), and Williams %R.

The calculations were implemented in Python using the Pandas library, renowned for its data manipulation capabilities. Each indicator was meticulously computed using rolling windows over the price data, adhering to their respective mathematical formulas.

Sentiment Index Calculation Model

Next, we directed our focus to the Sentiment Index Calculation Model. This model's input is a dataset of tweets, a rich tapestry of public sentiment. We utilized a fine-tuned DistilBERT-based model, named `distilRoberta-financial-sentiment`, to evaluate the sentiment of each tweet, categorizing it as positive, negative, or neutral.

The sentiment analysis was conducted using the Transformers library from Hugging Face, chosen for its state-of-the-art natural language processing tools. The sentiment scores were then aggregated to produce a daily sentiment index, reflecting the collective sentiment for each company on any given day.

Attention-Based LSTM Model

The final act of our experimental setup is the Attention-Based LSTM Model, which synthesizes the outputs of the previous two models. This model takes in the concatenated feature set comprising technical indicators and sentiment indices, alongside the raw historical price data.

The architecture of the LSTM model is fortified by an attention mechanism, allowing the model to discern and emphasize the most relevant timesteps for prediction. This nuanced focus is crucial, given the volatile nature of stock markets, where certain events carry more predictive weight than others.

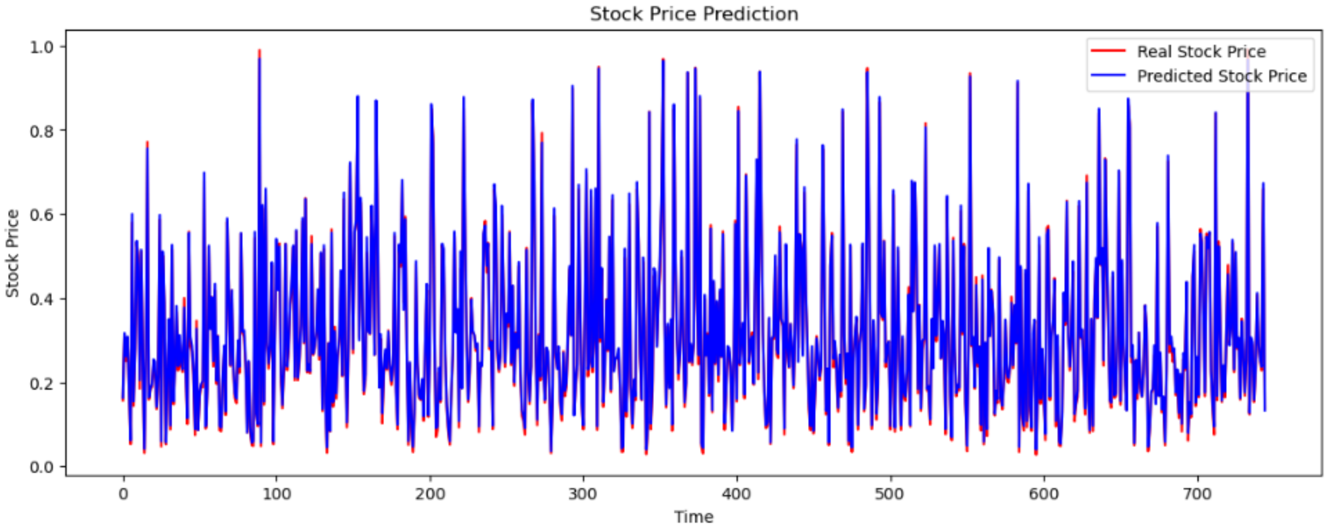
The LSTM model was built and trained using Keras and TensorFlow, leveraging their extensive deep learning functionalities. The training process was conducted over 100 epochs with a batch size of 32, a configuration determined to optimize the learning process.

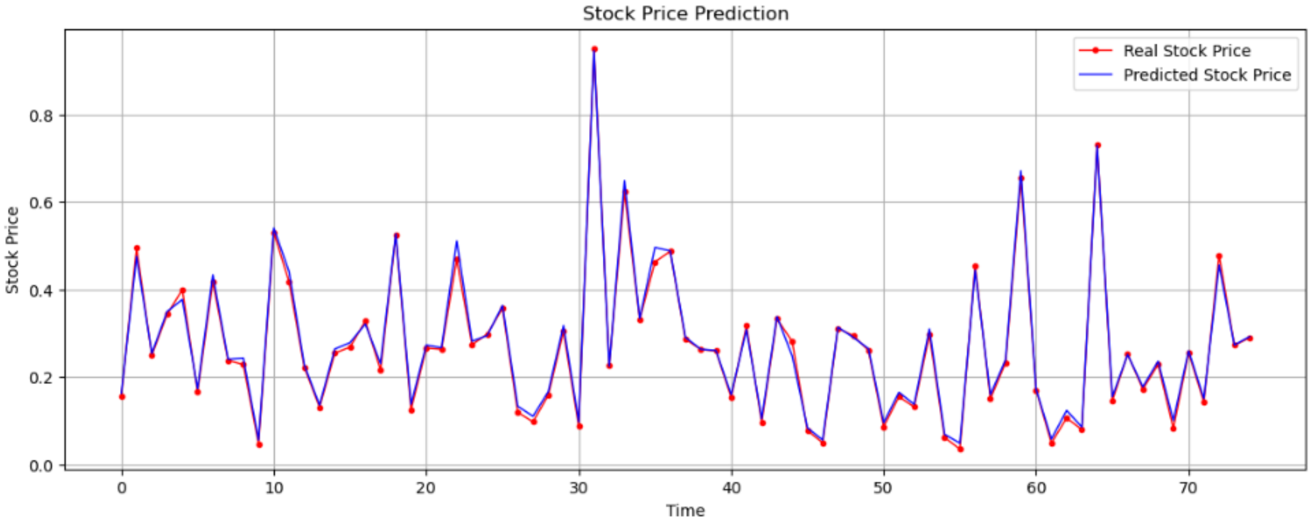
Results and visualization

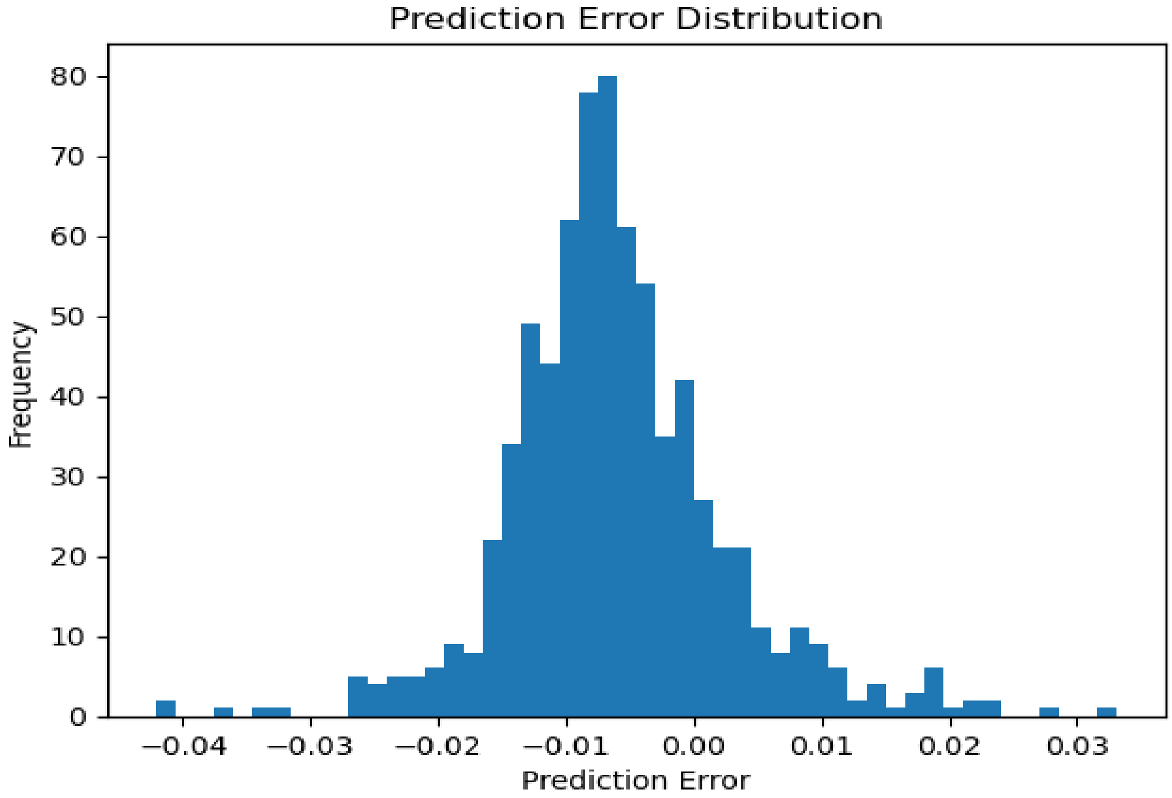
Upon training completion, the model was evaluated using the test dataset. The loss function used was the Mean Squared Error (MSE), and the Mean Absolute Error (MAE) was employed as an additional performance metric.

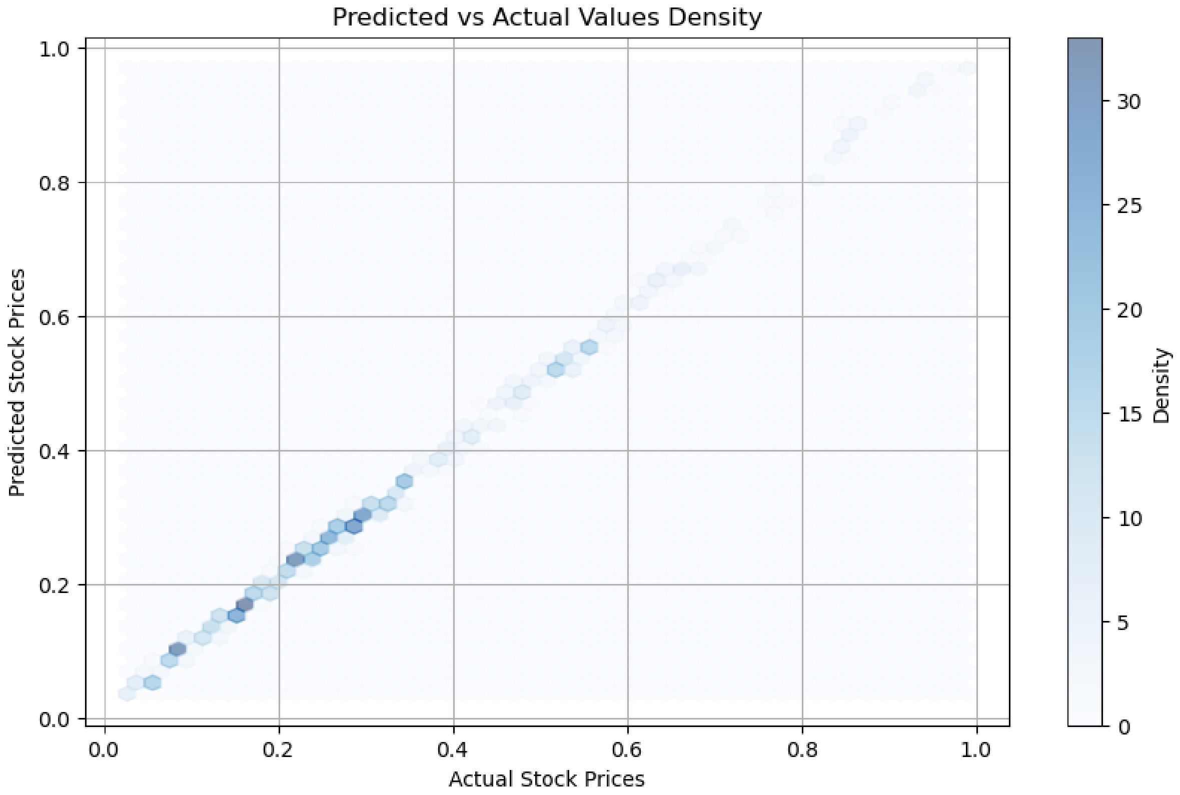
The final evaluation yielded an MSE of 1.0972e-04 and an MAE of 0.0086, indicating a high degree of accuracy in the stock price predictions. These results were visually corroborated by plotting the predicted stock prices against the actual stock prices, revealing a close alignment.

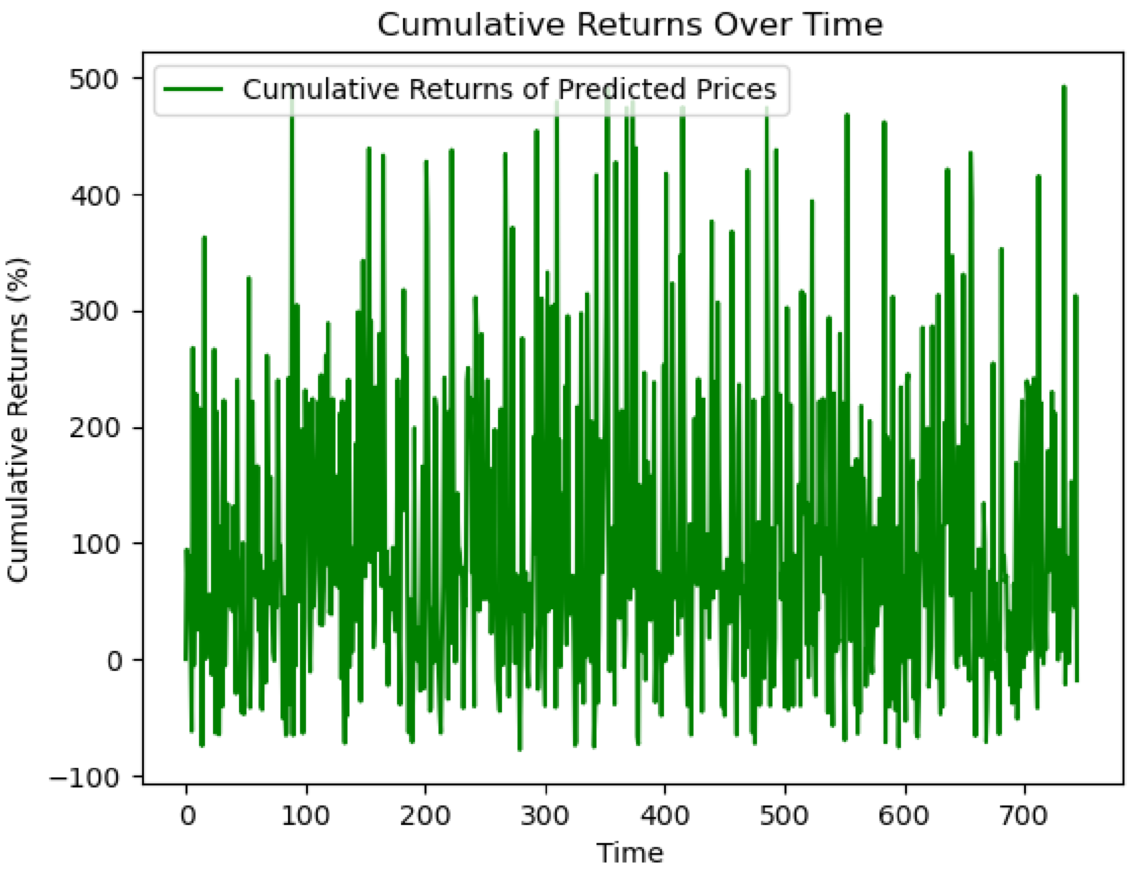
Further analysis involved plotting a histogram of prediction errors, which showed a normal-like distribution centered around zero, suggesting no bias in the predictions. We also visualized the density of predicted versus actual values, which confirmed the model's efficacy.

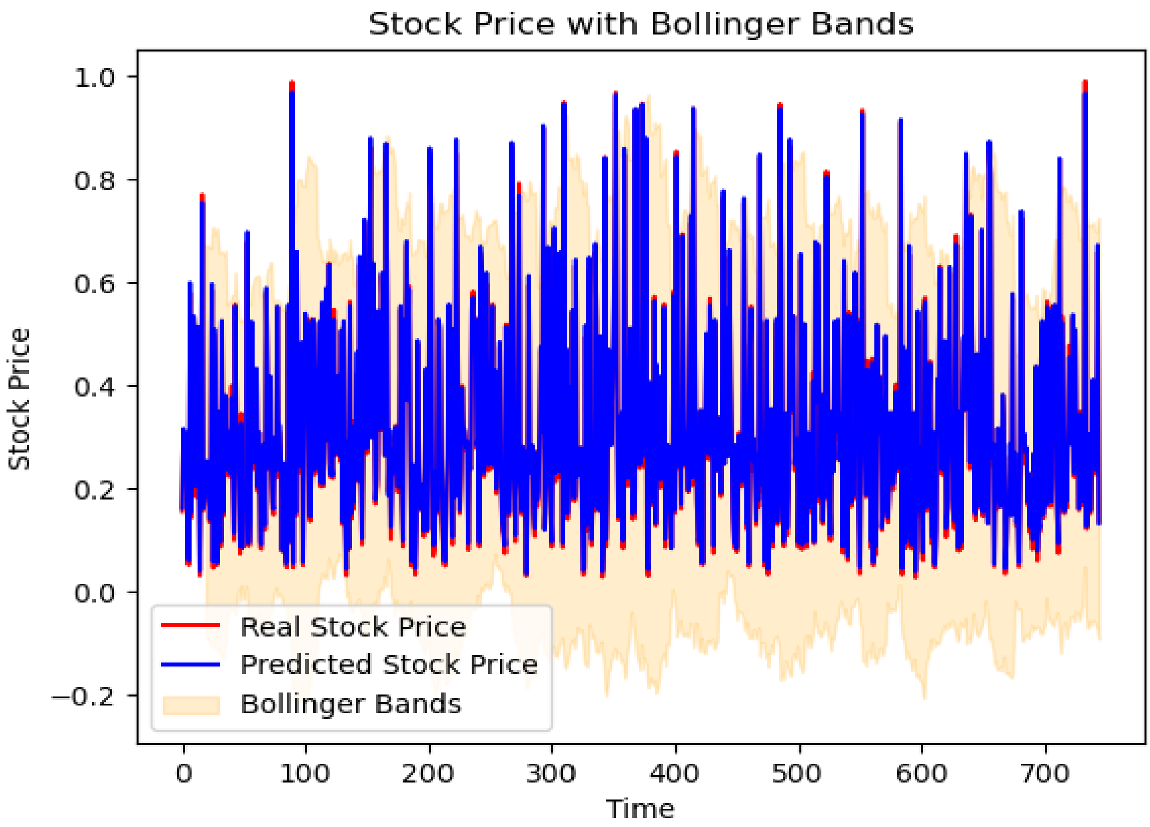





Prediction Error Distribution
Frequency
0
10
20
30
40
50
60
70
80
−0.04
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03
Prediction Error


Predicted vs Actual Values Density
Predicted Stock Prices
0.0
0.2
0.4
0.6
0.8
1.0
Actual Stock Prices
0.0
0.2
0.4
0.6
0.8
1.0
Density
0
5
10
15
20
25
30


Cumulative Returns Over Time
Cumulative Returns of Predicted Prices
Cumulative Returns (%)
500
400
300
200
100
0
−100
0
100
200
300
400
500
600
700
Time


Stock Price with Bollinger Bands
Real Stock Price
Predicted Stock Price
Bollinger Bands
Stock Price
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
100
200
300
400
500
600
700
Time

Discussion and conclusion

Our exploration into the realm of stock price prediction through an ensemble of technical indicators, sentiment analysis, and attention-based LSTM models has yielded intriguing insights and commendable results.

The Technical Indicator Calculation Model set the stage by distilling complex market behaviors into understandable metrics that capture the essence of market trends. Our results underscored the importance of technical analysis in understanding stock price movements, aligning with the existing literature that views these indicators as vital instruments in a trader's toolkit.

The Sentiment Index Calculation Model harnessed the power of the DistilBERT-based model, distilRoberta-financial-sentiment, to convert the cacophony of public opinion on social media into structured sentiment indices. This conversion from qualitative to quantitative data was pivotal, providing a previously untapped dimension to the predictive model. The daily sentiment index, when integrated with technical indicators, enriched the dataset, reflecting the pulse of the market sentiment.

The Attention-Based LSTM Model's performance was robust, achieving a low MSE and MAE. The attention mechanism's ability to prioritize certain inputs over others proved valuable, mimicking an experienced trader's focus on key market-moving events. This approach's effectiveness is particularly promising, considering the complexity and often chaotic nature of financial markets.

The study's findings reveal that the integration of technical indicators with sentiment analysis derived from social media can significantly enhance the accuracy of stock price predictions. Our results advocate for a multi-faceted approach to financial forecasting, where the fusion of different data types can uncover patterns that might be imperceptible to traditional analytical techniques.

While the MSE and MAE indicate high accuracy, it is crucial to recognize the limitations of the study. Financial markets are influenced by myriad factors, including economic indicators, geopolitical events, and market manipulation, which were not captured by our model. Furthermore, the model's performance in different market conditions, such as high volatility or crisis periods, requires further investigation.

The promising results of this study not only contribute to academic discourse but also present practical implications for traders, investors, and financial analysts. The proposed model can serve as a decision-support tool in investment strategies, potentially leading to more informed and rational decision-making.

List of references

1. Fama, E. F. (1964). *The Behavior of Stock Market Prices*. Journal of Business, 38(1), 34–105.
2. Malkiel, B. G. (1973). *A Random Walk Down Wall Street*. W.W. Norton & Company.
3. Markowitz, H. (1952). *Portfolio Selection*. The Journal of Finance, 7(1), 77–91.
4. Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2018). *News impact on stock price return via sentiment analysis*. Knowledge-Based Systems, 69, 14–23.
5. Oliveira, N., Cortez, P., & Areal, N. (2016). *The impact of microblogging data for stock price prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices*. Expert Systems with Applications, 73, 125–144.
6. Sun, A., Nguyen, Q. V. H., & De, S. (2017). *Sentiment Analysis for Stock Market Prediction from Financial News Articles*. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).
7. Statman, M. (2011). *Behavioral Finance: The Second Generation*. Journal of Financial Analysts, 55(2), 15–19.
8. Wang, H., & Ye, Z. (2017). *Sentiment Analysis of Investor Opinions on Twitter*. Social Network Analysis and Mining, 7(1), 22.
9. Vargas, M. R., de Lima Bicho, A., & Evsukoff, A. (2017). *Deep learning for stock market prediction from financial news articles*. In Proceedings of the IEEE International Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr).
10. Porshnev, A., Redkin, I., & Shevchenko, A. (2013). *Improving Prediction Market Forecasts by Detecting and Correcting Possible Over-reaction to Price Movements*. Expert Systems with Applications, 40(1), 188–199.
11. Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). *Big Data: Deep Learning for financial sentiment analysis*. Journal of Big Data, 5(1), 3.
12. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.
13. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.
14. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735–1780.
15. Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2018). *Understanding and Enhancement of Internal Clustering Validation Measures*. IEEE Transactions on Cybernetics, 48(6), 1785–1798.

16. Wu, J., Deng, S., Huang, L., & Tan, W. (2018). *Deep Learning for Stock Prediction Using Numerical and Textual Information*. In Proceedings of the 2018 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD).

17. Hyndman, R. J., & Koehler, A. B. (2006). *Another look at measures of forecast accuracy*. International Journal of Forecasting, 22(4), 679–688.