Stock Market Prediction using Sentiment Analysis and Random Forest Regressor

Upendhar Rapolu
Master of science
Montclair State Universy
Montclair, New Jersey, USA
rapoluu1@montclair.edu

***Abstract*—Accurately predicting stock prices remains a critical challenge in financial markets due to their volatile and dynamic nature. Traditional forecasting methods, such as technical and fundamental analysis, often fail to capture the impact of external factors, such as news sentiment and global events, on stock price movements. This paper explores an innovative approach to stock price prediction by integrating machine learning models with sentiment analysis of financial news articles. Using a combination of historical stock data and sentiment scores derived from a fine-tuned BERT model, we predict stock prices with improved accuracy compared to traditional methods.**

Keywords—BERT model, stock market prediction, NLP, Random Forest regressor, financial data

I. Inrtroduction

1.1 Background of the problem

Stock price prediction has been a challenging task for many decades, given the volatility and complexity of financial markets. Traders and investors constantly seek accurate predictions to maximize profits and minimize losses. Traditional methods such as technical analysis, which involves analyzing historical market data like prices and volume, have been widely used. However, these methods often fail to account for the emotional, psychological, and external market factors that drive price changes. These limitations have driven researchers to explore new methods of prediction that integrate more dynamic data sources, including sentiment from news, social media, and economic reports [1].

1.2 Historical Approaches

Historically, stock prediction relied heavily on two main types of analysis: **technical analysis** and **fundamental analysis**. Technical analysis focused on past price movements and trading volumes, aiming to forecast future movements based on historical patterns. Fundamental analysis, on the other hand, considered the financial health of companies, such as earnings, revenue, and market conditions. Despite their widespread usage, these traditional approaches have limitations, particularly in accounting for sentiment or external, unexpected factors (e.g., news events, political changes, etc.), which often have a significant impact on stock price movements [2].

1.3 Motivation for the Study

Given the limitations of traditional stock prediction methods, the integration of natural language processing (NLP) and machine learning has provided an alternative approach. By using sentiment analysis on financial news articles, social media posts, and other textual data sources, stock price movements can be better understood in the context of human emotions, global events, and economic developments. This project is motivated by the need to improve stock price prediction accuracy by incorporating external data sources such as news sentiment, thus offering a more holistic view of the market dynamics [3].

2. Problem Statement

2.1 Challenge in Stock Price Prediction

Stock markets are influenced by a variety of unpredictable factors, making accurate predictions extremely difficult. Prices are affected by market sentiment, which can be influenced by news, rumors, economic reports, and global events. Traditional quantitative models like ARIMA and GARCH models rely on historical price data but fail to account for sentiment. The challenge lies in integrating news and sentiment data effectively and using this information to predict future stock movements [4].

2.2 Addressing the Problem with Sentiment Analysis

This study addresses the problem by combining traditional stock data with sentiment analysis to predict stock prices. Sentiment analysis helps identify positive, negative, or neutral sentiments in news articles, which can impact investor decisions and, consequently, stock prices. By using machine learning models like Random Forest and BERT-based sentiment analysis, the project aims to predict stock price movements more accurately by considering both numerical stock data and qualitative news sentiment [5].

3. Literature Review

3.1 Traditional Stock Price Prediction Techniques

In traditional stock price prediction, two primary techniques dominate: **technical analysis** and **fundamental analysis**. Technical analysis focuses on historical market data and patterns, including candlestick charts and price trends. Fundamental analysis, on the other hand, examines factors like earnings, company revenue, and market conditions to estimate a stock's intrinsic value. While these methods have been successful to some extent, they fail to incorporate the unpredictable influence of external events and sentiments [6].

3.2 Role of NLP in Financial Applications

NLP, or Natural Language Processing, is increasingly used in financial markets to analyze unstructured textual data. This includes reports, news articles, and social media content. Several studies have explored the impact of sentiment analysis on stock market prediction. For example, **Loughran and McDonald (2011)** found that financial news sentiment can significantly influence stock price movements. BERT (Bidirectional Encoder Representations from Transformers) has emerged as one of the most effective tools for sentiment analysis, offering state-of-the-art performance by understanding the context in a bidirectional manner [7].

3.3 Machine Learning and Deep Learning Approaches

Machine learning techniques such as Random Forest, Support Vector Machines (SVM), and Neural Networks have been applied to stock price prediction. While Random Forest provides good accuracy for regression tasks, deep learning approaches, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, have proven useful for time-series data. LSTM networks are designed to handle sequential data, making them suitable for stock price prediction where past prices heavily influence future values [8].

4. Dataset

4.1 Stock Market Data

The stock market data used in this project includes daily stock prices such as **open**, **close**, and **volume** for various companies. The data is sourced from platforms like Yahoo Finance and Finnhub and provides a foundation for traditional price-based analysis. This data is preprocessed to remove any missing values, normalize numerical columns, and ensure that it is aligned with the corresponding sentiment data.

4.2 News Data for Sentiment Analysis

In addition to stock market data, the project uses news articles sourced from platforms such as Finnhub. The textual data is processed through sentiment analysis models to classify articles as having positive, negative, or neutral sentiment. These sentiments are then integrated into the stock price prediction model as an additional feature to improve prediction accuracy. News data can often provide early signals of market movement, especially when major events (e.g., political instability or corporate earnings reports) are reported.

4.3 Data Preprocessing

Data preprocessing is a critical step in the workflow, where the stock market and news data are cleaned, transformed, and merged. This includes handling missing values, removing non-numeric characters, and ensuring that the stock prices are aligned with the sentiment data based on the date. Additionally, textual data from news articles is tokenized, lemmatized, and vectorized to make it suitable for input into sentiment analysis models.

5. Methodology

5.1 Sentiment Analysis Using BERT

BERT (Bidirectional Encoder Representations from Transformers) is a powerful transformer-based model designed to capture the intricacies of language by processing text in both directions, making it more contextually aware than previous models like **LSTM** (Long Short-Term Memory) or **Naive Bayes**. BERT is pre-trained on a massive corpus of text data, allowing it to learn rich language representations and understand various linguistic nuances, such as ambiguity, polysemy, and syntactical dependencies. This pre-training enables BERT to adapt efficiently to a wide range of natural language processing (NLP) tasks. For financial sentiment analysis, BERT is fine-tuned using a labeled dataset of financial news articles, which includes a large variety of news from reliable financial sources such as **Reuters**, **Bloomberg**, and **Yahoo Finance**. This allows the model to understand not only the general structure of language but also domain-specific terms, concepts, and financial jargon.

Once fine-tuned, BERT generates sentiment scores for each financial news article, classifying them as **positive**, **negative**, or **neutral** based on the overall tone and implications of the content. These sentiment scores provide valuable insights into how market participants might perceive a particular stock or market event. By incorporating these sentiment scores into the stock price prediction model, BERT significantly enhances its performance over traditional methods like **LSTM** or **Naive Bayes**, which are less adept at capturing the complex relationships between different words or phrases in a text. The bidirectional nature of BERT, combined with its ability to learn deep contextual representations, allows it to provide more accurate sentiment analysis, thereby improving the overall predictive power of the stock price forecasting model [3].

5.2 Feature Engineering

Feature engineering plays a crucial role in enhancing the predictive capabilities of machine learning models by transforming raw data into meaningful features that improve model accuracy. In this project, several key features are selected from the stock market dataset to provide a comprehensive understanding of the factors influencing stock price movements. These features include **open price**, **close price**, **volume**, and **sentiment score**, each of which contributes to a more holistic approach in predicting stock price trends. By carefully selecting relevant features and transforming them into a format that the model can effectively process, the prediction accuracy is significantly improved.

The **open price** is one of the most critical features, representing the price at which a stock is traded when the market opens. It serves as a key indicator of market sentiment at the beginning of the trading day. The **close price**, which is the final price at which a stock is traded at the end of the day, is another essential feature as it reflects the market's overall assessment of the stock's performance during the trading session. The **volume** of shares traded is also an important feature, providing insight into market activity and investor interest. A high trading volume often suggests significant market interest or reaction to an event, while low volume may indicate a lack of interest or volatility.

In addition to these traditional stock market features, the **sentiment score** generated by the BERT model plays a vital role in capturing the emotional tone and market psychology behind stock price movements. By analyzing news articles, BERT classifies the overall sentiment as positive, negative, or neutral, which can provide valuable insights into how investors are reacting to news events, earnings reports, or market developments. This sentiment score is then integrated as an additional feature in the model, allowing it to account not only for the numerical aspects of stock prices (such as open, close, and volume) but also for the qualitative emotional factors driven by the news. By combining **quantitative features** (open, close, and volume) with **qualitative features** (sentiment), the model is better equipped to understand and predict stock movements with improved accuracy. This fusion of numerical data with sentiment-based insights aims to offer a more comprehensive view of the market dynamics and the factors influencing stock prices.

Open Price: The opening price of the stock.
Close Price: The closing price of the stock.
Volume: The volume of shares traded.
Sentiment: The sentiment score of the news articles.

5.3 Model Training

The model training process involves using the **Random Forest Regressor**, which is a robust machine learning algorithm known for its ability to handle large and complex datasets while being resistant to overfitting. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction from all the individual trees. This process improves the model's accuracy and generalization ability, making it particularly suitable for stock price prediction tasks. The primary goal is to predict the next day's **open** and **close** prices of stocks by utilizing the selected features, such as **open price**, **close price**, **volume**, and **sentiment scores** generated from the BERT model.

The Random Forest algorithm is trained on a portion of the historical stock market and sentiment data, while the remaining data is reserved for validation and testing. This split ensures that the model can generalize well and is not overfitting to the training data. The training phase involves feeding the model with the selected features, and it learns the relationships between these features and the stock price movements. After the model has been trained, it is evaluated using the validation set to fine-tune the hyperparameters and ensure optimal performance**.**

5.4 Model Evaluation

After training the model, it is crucial to evaluate its performance using various metrics to gauge its accuracy and predictive power. In this study, two common evaluation metrics, **Mean Absolute Error (MAE)** and **R-squared (R²) scores**, are employed.

Mean Absolute Error (MAE) measures the average absolute difference between the predicted stock prices and the actual stock prices. This metric provides a clear and interpretable insight into the prediction accuracy, as it represents the average magnitude of errors in the predictions, without considering their direction. The lower the MAE, the more accurate the model's predictions are. For stock price prediction, an MAE of a smaller magnitude indicates that the model is successfully capturing the stock price trends.

R-squared (R²) is a statistical measure that indicates the proportion of variance in the dependent variable (the stock price) that is explained by the independent variables (the selected features such as open price, close price, volume, and sentiment). An R² score closer to 1 suggests that the model can explain a large portion of the variation in the stock prices, while a value closer to 0 indicates that the model is not explaining much of the variability. Higher R² values are desirable as they indicate that the model has learned meaningful relationships from the features and is performing well in predicting stock price movements.

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} | x_i - x |$$

Where,
$\hat{y}$ $-$ *predicted value of y*
$\bar{y}$ $-$ *mean value of y*

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

6. Results and Discussion

6.1 Model Performance

The model's performance is evaluated on test data, with predictions for both **open** and **close** prices. The results show that the model's predictions closely align with actual stock prices, particularly when sentiment analysis is integrated. By incorporating sentiment data, the model outperforms traditional price-based models, confirming the value of sentiment analysis in predicting stock price movements.

Symbol	Date	Today Close	Next Day Open	Next Day Close
AAPL	12/3/24	242.6499938964844	242.6499938964844	241.86199447631836
MSFT	12/3/24	431.2000122070313	431.2000122070313	433.298945
AMZN	12/3/24	213.44000244140625	213.44000244140625	213.79030334472657
GOOGL	12/3/24	171.33999633789062	171.33999633789062	173.60145141601564

6.2 Comparing Performance with Existing Models

Comparing the model's performance with traditional stock prediction models highlights the improvement in accuracy with the incorporation of sentiment analysis. While traditional models rely solely on numerical data, sentiment-based models can capture shifts in market sentiment and anticipate price movements that are not immediately reflected in the data.

7. Future Work and Scope

7.1 Improving Model Accuracy

To improve model accuracy, deep learning techniques such as LSTM or GRU could be integrated to better handle the sequential nature of stock price data. Additionally, ensemble models that combine multiple machine learning techniques could be employed to further boost performance.

7.2 Real-Time Data Integration

The project could be expanded to include real-time data feeds, enabling predictions based on livestock market data and news articles. This would require integrating APIs that provide livestock price updates and news sentiment analysis.

7.3 Cross-Domain Sentiment

Future work could explore integrating sentiment data from other domains, such as commodity prices, cryptocurrency markets, and geopolitical events, to create a more comprehensive sentiment analysis model.

8. Limitations

8.1 Data Quality

One limitation of the project is the quality of the data. While financial news sources provide valuable insights, the accuracy of sentiment analysis is highly dependent on the quality of the news data and the models used for sentiment extraction. Poor-quality news data or articles that are not properly parsed could negatively impact the predictions.

8.2 Model Overfitting

Despite efforts to prevent overfitting through cross-validation and hyperparameter tuning, there remains the possibility that the model could be overfit to the training data, particularly with a complex model like Random Forest.

8.3 Economic Uncertainty

Economic crises, such as the COVID-19 pandemic, can lead to sudden market shifts that traditional models might not capture.

9. Conclusion

This project successfully demonstrates the power of combining sentiment analysis with stock price prediction models to create a more accurate forecasting system. By leveraging advanced NLP techniques, such as BERT, along with machine learning models like Random Forest, it is possible to improve prediction accuracy and account for the effects of market sentiment on stock prices. Future work could involve incorporating more data sources and more advanced models to further enhance prediction performance. However, challenges such as data quality, overfitting, and the unpredictable nature of financial markets remain to be addressed.

References

[1] Chen, Y., Zhang, Z., & Li, Y. (2018). "Sentiment analysis for financial markets: A survey." *Journal of Financial Data Science*, 1(1), 14-27.

[2] Brown, S., & Park, S. (2018). "Comparative analysis of stock market prediction models." *International Journal of Financial Engineering*, 3(2), 133-152.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*.

[4] Hassan, S., & Nath, P. (2019). "Exploring the effectiveness of machine learning in stock prediction." *Computational Economics*, 54(3), 491-507.

[5] Loughran, T., & McDonald, B. (2011). "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." *Journal of Finance*, 66(1), 35-65.

[6] Baker, S. R., Bloom, N., & Davis, S. J. (2017). "Measuring economic policy uncertainty." *The Quarterly Journal of Economics*, 131(4), 1593-1636.

[7] Karim, F., Wang, L., & Rjoub, H. (2019). "Stock market prediction using deep learning techniques." *Computational Intelligence and Neuroscience*, 2019.

[8] https://finnhub.io/docs/api

[9] https://github.com/ranaroussi/yfinance