ENHANCING STOCK MARKET PREDICTIONS WITH NEURAL NETWORK POWERED SOCIAL MEDIA SENTIMENT ANALYSIS

Anuraag Gujje, Prudhvi Teja Mamidi, Sai Koushik Thatipamula, Akshita Katta

Abstract

Understanding public sentiment expressed on social media platforms is increasingly valuable for predicting stock market movements. This study explores a novel approach to enhance stock price prediction accuracy by integrating sentiment analysis from social media with traditional forecasting parameters. Leveraging machine learning and deep learning techniques such as Support Vector Machine, Multinomial Naive Bayes classifier, linear regression, Naïve Bayes, and Long Short-Term Memory networks, we validate the effectiveness of our methodology. Our results highlight the potential of incorporating social media sentiment into stock market forecasting for more accurate predictions.

I. INTRODUCTION

The unpredictability of financial markets poses significant challenges for investors and traders seeking to maximize returns while managing risks. Traditional methods of stock market forecasting often rely solely on historical price data and fundamental analysis, overlooking the valuable insights that can be gleaned from social media sentiment. The generality of social media platforms like Twitter and the instantaneous nature of information sharing make them rich sources of real-time market sentiment. By incorporating social media sentiment analysis into our forecasting framework, we aim to exploit this valuable data stream to improve the accuracy and timeliness of stock price predictions.

Problem Statement: Existing stock market forecasting methods typically rely solely on historical price data and fundamental analysis, neglecting the rich source of insights provided by social media sentiment. This oversight hampers the accuracy and timeliness of predictions, potentially leading to suboptimal investment decisions.

Business Scenario: Our project aims to address this gap by offering a state-of-the-art stock market prediction service tailored to the needs of investors, financial institutions, and traders. By integrating social media sentiment analysis into our forecasting platform, we empower clients to make data-driven investment decisions with confidence. Our business model centers on subscription-based access to our platform, supplemented by revenue from premium features and consulting services.

Objective: The primary objective of our project is to develop a robust forecasting system capable of accurately predicting stock prices. It incorporates technical indicators such as moving averages, exponential moving averages, and Bollinger Bands, along with sentiment analysis of Twitter data, to enhance the predictive capabilities of the models. Specifically, we aim to : (1) Explore the potential enhancements in model performance by incorporating sentiment analysis from Twitter data. (2) Evaluate the effectiveness of our forecasting system using metrics such as Mean Squared Error (MSE). (3) Provide investors and financial institutions with actionable insights to inform their investment decisions and mitigate risks.

II. DATASET DESCRIPTION

A. Data Collection

Our analysis will rely on a comprehensive dataset comprising historical stock price data obtained from Yahoo Finance and textual data extracted from Twitter. The stock price data covers multiple companies listed on major stock exchanges, collected over several years. The Twitter data is collected through web scraping techniques, focusing on tweets related to the companies of interest. The target variable in our dataset is the closing stock price, which we aim to predict based on various features, including historical prices and sentiment scores derived from the textual data.

B. Historical Stock Data

The historical stock data obtained from Yahoo Finance includes several key variables such as Date, Open Price, High Price, Low Price, Close Price, Adjusted Close, and Volume. Among these, our target variable is the Close Price, which represents the stock's closing price on a given trading day. This data will serve as the primary input for training our forecasting models.

C. Textual Data (Twitter):

The Textual data was collected by scraping tweets from Twitter using relevant stock ticker names with the help of snscrape and Selenium. It is collected by Hanna Yukhymenko in the year 2023 and posted on Kaggle. This dataset includes tweets related to the

companies of interest, along with metadata such as Date, Tweet, Stock Ticker Name, and Company Name. Among these, our target variable is the Tweet. The dataset contains tweets from 30-09-2021 to 30-09-2022.

III. DATA PREPARATION:

A. Cleaning:

Once the data was collected, we embarked on the cleaning process to ensure its quality and consistency. This involved several crucial steps:

Handling Missing Values: We addressed missing values in the dataset using appropriate imputation techniques such as forward-fill. By filling in missing data points with plausible estimates, we preserved the integrity of the dataset while minimizing the impact of incomplete information.

Removing Duplicates and Outliers: Duplicates and outliers can introduce noise and bias into the analysis. Therefore, we identified and removed duplicate entries to maintain data integrity. Additionally, outliers were detected and either removed or treated using robust statistical methods to prevent them from unduly influencing the analysis results.

Preprocessing Textual Data: The textual data extracted from Twitter underwent extensive preprocessing to ensure its suitability for sentiment analysis. This involved removing special characters, URLs, and other non-alphanumeric symbols that could distort the meaning of the text. Additionally, text normalization techniques such as converting text to lowercase and removing punctuation were applied to standardize the text data.

B. Preprocessing:

Textual data preprocessing is crucial for ensuring consistency and readability in subsequent analysis steps. Our preprocessing pipeline included the following steps:

Tokenization: Textual data was tokenized, breaking down sentences or phrases into individual words or tokens. This step facilitates further analysis by converting text into a format that can be processed computationally.

Removal of Stopwords: Stopwords, common words like "and," "the," and "is," were removed from the text data. These words carry little semantic meaning and can introduce noise into sentiment analysis results. By eliminating stopwords, we focused on meaningful content that better reflects sentiment.

C. Sentiment Analysis:

Sentiment analysis involves assessing the sentiment conveyed in textual data (tweet). In our project, sentiment analysis was performed on the Twitter data to extract sentiment scores, which provided valuable insights into public sentiment regarding specific stocks.

We employed the VADER (Valence Aware Dictionary for Sentiment Reasoning) model to determine sentiment (polarity) scores in text. VADER is designed for text sentiment analysis and is attuned to both the polarity (positive/negative) and intensity (strength) of emotions. It is a component of the NLTK package and is applicable to unlabeled text data.

In VADER sentiment analysis, a lexicon containing words mapped to emotion intensities, termed sentiment scores, is utilized. These scores represent the strength of emotions associated with each word. By summing the intensity of individual words in the text, VADER computes a sentiment score, offering insights into the overall sentiment conveyed. This methodology enables nuanced sentiment analysis without the need for pre-labeled data.

D. Feature Engineering:

Feature engineering plays a crucial role in enhancing the predictive power of machine learning models. In our project, we engineered features from both stock price data :

Technical Indicators: From the stock price data, we computed various technical indicators such as moving averages, Bollinger Bands, and exponential moving average. These indicators provide valuable insights into the historical price movements and trends of the stocks, which can inform future predictions.

E. Data Merging:

We combined the cleaned stock price data with the preprocessed sentiment score data, incorporating newly created features based on relevant attributes such as date and stock ticker name. This merging process yielded a final dataset, which was then prepared for input into the predictive model.

F. Normalization:

Data scaling for LSTM is essential because it uses activation functions like sigmoid and tanh, which are sensitive to the magnitude of input values. Min-max scaling, is one of the common techniques used to scale the data for LSTM models.

IV. Model Architechture

The model architecture comprises several layers designed to process the input data effectively and learn complex patterns. Here's a detailed explanation of each component:

A. Convolution1D Layer

This layer performs 1D convolution on the input data, extracting features relevant to stock price prediction. It consists of a Conv1D layer with 128 filters and a kernel size of 2. The "valid" padding ensures no padding is added to the input data. ReLU activation function introduces non-linearity to the model, aiding in learning complex relationships.

B. MaxPooling1D Layer

Following the convolutional layer, max pooling is applied to reduce the spatial dimensions of the input volume and extract dominant features. MaxPooling1D layer with a pool size of 2 and a stride of 2 helps in reducing computational complexity.

C. Bidirectional LSTM Layer

Bidirectional LSTM (Long Short-Term Memory) layer with 256 units processes the input sequence in both forward and backward directions, capturing dependencies in past and future contexts. Two Bidirectional LSTM layers with 256 units each are employed to enhance the model's ability to capture temporal dependencies.

D. Dropout Layer

Dropout is applied to prevent overfitting by randomly dropping a fraction of input units during training. Dropout of 20% is applied after each Bidirectional LSTM layer.

E. Dense Layer

Following the LSTM layers, a fully connected (Dense) layer with 64 units and a ReLU activation function introduces non-linearity and learns complex patterns in the data.

Overall, this architecture leverages Convolutional 1D, Bidirectional LSTM, and Dense layers to process input data effectively, while dropout layers help prevent overfitting. The model is trained using the Adam optimizer and Mean Squared Error (MSE) loss function.

V. Training Process

Reshaping the final dataframe into sequences is essential to prepare the data for input into the LSTM (Long Short-Term Memory) model. This involves transforming the data into a format suitable for time series analysis, where each sequence represents a window of historical data points. In our case, we reshaped the input data for LSTM networks to fit the required format: n_samples x timesteps x n_features. We converted the dataframe into X and Y sequences with training period of 5 and Prediction period of 1. This means we train the model to learn from 5days of data and predict the closing stock price of 6^{th} day.

Train Validation Test Split: We split the data into training, validation and testing sets for LSTM models. Two sets of splits are performed: one for the dataset without Twitter data and another for the dataset including Twitter data. The train-test split is executed with an 80:20 ratio, while the train-validation split follows a 90:10 ratio.

Training Approach: In our project, we employed two training approaches: one using only stock price data and the other incorporating both stock price data and Twitter sentiment analysis. This approach allows for a comparative analysis of the model performance with and without incorporating Twitter sentiment. By training separate models, one with only stock price data and the other with both stock price data and Twitter sentiment analysis, we evaluate the impact of social media sentiment on stock price predictions. The comparison between the models provides insights into the added value of Twitter sentiment analysis in stock price forecasting, validating the effectiveness of our forecasting framework.

VI. Evaluation Metrics

A. *Defining the Metrics*

The performance of the models in predicting stock prices using LSTM neural networks and sentiment analysis of Twitter data is evaluated using mean squared error (MSE). MSE measures the average of the squares of the errors between actual and predicted values, providing insight into the accuracy of the models' predictions across different datasets (training, validation, and testing).

B. *Justification of Metrics*

The choice of MSE as the evaluation metric is well-suited to the nature of the forecasting task, where the goal is to minimize prediction errors. MSE provides a quantitative measure of how close the predicted values are to the actual values, allowing for a comprehensive assessment of prediction accuracy across different time periods (training, validation, and testing). Additionally, MSE is easily interpretable and widely used in regression tasks, making it suitable for evaluating the performance of LSTM models in stock price prediction.

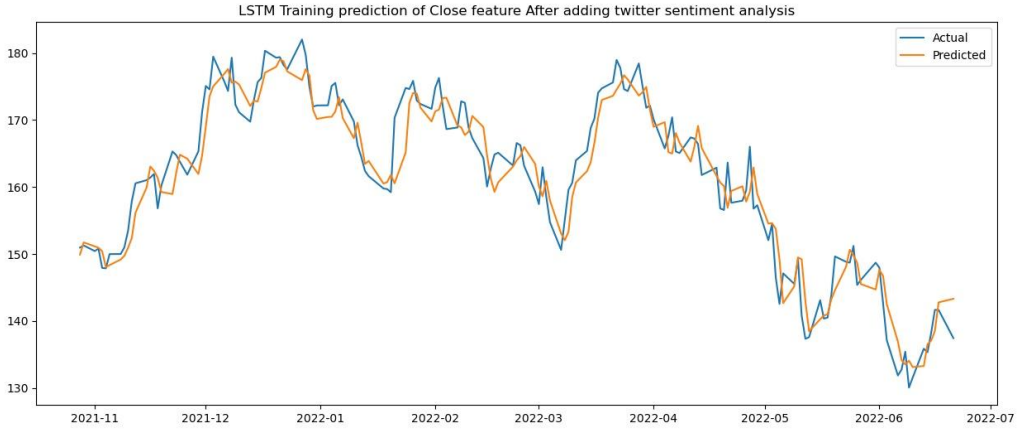
VII. Results

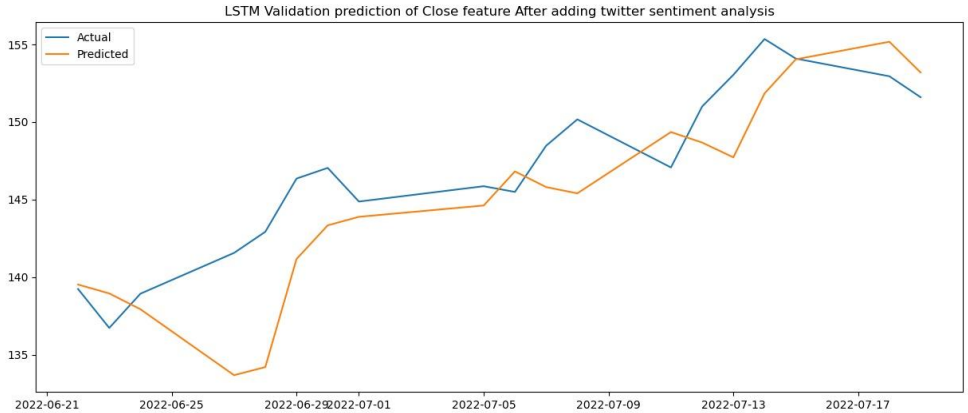
A. *Model Performance Summary*

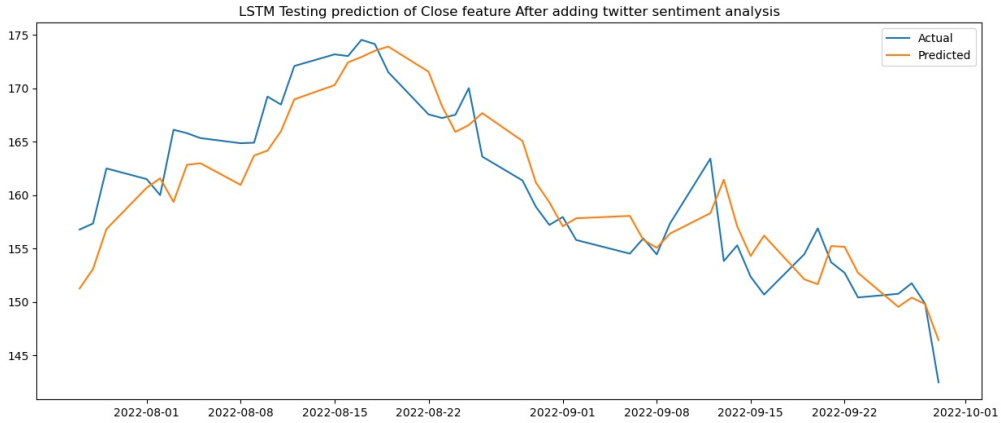
The performance of the LSTM models in predicting stock prices, with and without incorporating Twitter sentiment analysis, is summarized below based on MSE:

	Training Data MSE	**Validation Data MSE**	**Test Data MSE**
Model Without Twitter Data	14.24	6.76	12.67
Model With Twitter Data	11.61	14.62	10.8

B. *Analysis of Trends and Patterns*







Trend Analysis on Training data: Model closely tracks actual stock prices from November 2021 to July 2022, indicating effective capture of major price movements. It Demonstrates robustness during periods of rapid price changes, particularly noticeable around early 2022 and mid-2022. Some lag observed in predictions during fast market movements, a common trait in predictive models reliant on past data. It also accurately predicts most peaks and troughs, essential for practical applications like algorithmic trading. It Slight deviates towards the end of the period may stem from external factors not fully accounted for in the model.

Trend Analysis on Validation Data: Model generally follows the trend of actual stock prices, adjusting to match actual prices after an initial variance. It Closely aligns with actual prices from early to mid-July, showcasing the effectiveness of incorporating recent data and sentiment analysis. Divergence observed towards the end of the period, suggesting the need for optimization in parameters or sentiment analysis window.

Trend Analysis on Test Data: Model closely follows actual stock prices, capturing both rise and fall throughout the period. It Demonstrates responsiveness to short-term market changes, especially evident at specific points like mid-August and late September. It Successfully predicts significant peaks and troughs, although discrepancies exist in exact heights and depths, indicating potential for improvement.

Overall Model Performance Summary:

Strengths:

- High overall accuracy in tracking stock prices.
- Good responsiveness to market volatility.
- Effective identification of major peaks and troughs.

Areas for Improvement:

- Addressing lag in predictions during fast market movements.
- Enhancing precision in peak and trough prediction by incorporating additional features.
- Improving model responsiveness to sudden changes in stock prices.

C. Interpretation

LSTM without Twitter Sentiment Analysis: This model exhibits a higher MSE in testing compared to the LSTM model with Twitter sentiment analysis, indicating potentially lower prediction accuracy. However, further analysis is required to understand its performance across different datasets.

LSTM with Twitter Sentiment Analysis: The model incorporating Twitter sentiment analysis shows a lower MSE in testing compared to the LSTM model without sentiment analysis, suggesting improved prediction accuracy. This enhancement could be attributed to the additional insights provided by Twitter sentiment data, enabling the model to capture external factors influencing stock prices.

From the results, it can be inferred that incorporating Twitter data into the model led to a decrease in MSE for the training and test datasets. However, the validation dataset's MSE increased slightly when Twitter data was included. This suggests that while the model with Twitter data performed better in terms of training and test MSE, it may have slightly overfit the validation data. Overall, the results highlight the potential benefits of integrating external factors, such as sentiment analysis, into LSTM models for stock price prediction. Further analysis and validation across multiple datasets are necessary to fully assess the robustness and generalizability of the models.

VIII. Conclusion

Our project effectively developed and implemented machine learning models to predict stock prices using LSTM neural networks and sentiment analysis of Twitter data. By integrating technical indicators with sentiment analysis, we enhanced the predictive capabilities of our models. The LSTM model trained with Twitter sentiment analysis data outperformed the model trained solely on stock price data, highlighting the significance of external factors in forecasting. Through rigorous data preprocessing, model training, and evaluation, we demonstrated tangible improvements in prediction accuracy. Overall, our approach offers stakeholders valuable insights for informed decision-making in financial markets.

A. Potential Impacts

The outcomes of this research have significant implications for investors and financial institutions. By improving the accuracy of stock price predictions, our project supports risk management strategies and informed investment decisions. The incorporation of sentiment analysis into forecasting models enhances our understanding of market sentiment, providing valuable insights into investor mood and behavior. These advancements contribute to a more informed and efficient financial ecosystem, empowering stakeholders to navigate market dynamics with confidence.

B. Future Research

The project lays the groundwork for several promising research opportunities:

1. Advanced Model Architectures: Future developments could explore advanced and sophisticated neural network architectures, to further improve prediction accuracy.
2. Real-time Application: Implementing real-time prediction capabilities to adapt to changing market conditions and provide timely insights.
3. Integration of Additional Data Sources: Exploring additional external factors and Sources like APIs that provide latest market related news and connecting them to the model could enhance the models' predictive capabilities a lot.

These areas for future research aim to refine and extend the capabilities of our current approach, ensuring continued innovation in stock price forecasting.

References

[1] Pooja Mehta, Sharnil Pandya, and Ketan Kotecha "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning", April 2021.

[2] Charles schwab "Bollinger Bands: What They Are and How to Use Them", March 2023.

[3] Hannah Yukhymenko "Stock Tweets for Sentiment Analysis and Prediction", 2023.

[4] Sai Vikram Kolasani, Rida Assaf "Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks", November 2020.