Enhancing Stock Market Prediction through Historical Prices and Sentiment Analysis: A Deep Learning Approach

L Kamatchi Priya
Dept. of Computer Science and Engineering
PES University, Bangalore, India
priyal@pes.edu

P Niveditha
Dept. of Computer Science and Engineering
PES University, Bangalore, India
nive129persona@gmail.com

Nihal T M
Dept. of Computer Science and Engineering
PES University, Bangalore, India
nihaltm2002@gmail.com

Melvin J Joseph
Dept. of Computer Science and Engineering
PES University, Bangalore, India
melvinjjoseph2002@gmail.com

Mayank P M
Dept. of Computer Science and Engineering
PES University, Bangalore, India
mayankpm285@gmail.com

***Abstract*—This paper presents a novel approach for stock market prediction leveraging historical price data and sentiment analysis of news headlines. We propose a model that integrates Long Short-Term Memory (LSTM) neural networks with sentiment analysis to forecast stock price movements. Our methodology involves scraping relevant news articles, extracting sentiment features, and combining them with historical price data for stocks and training the LSTM model. Results indicate promising predictive performance, demonstrating the efficiency of our approach in capturing the relationship between news sentiment and stock price dynamics.**

***Index Terms*—LSTM, sentiment analysis, stock prediction**

I. Introduction

Predicting stock market movements is a challenging endeavor, primarily due to the dynamic nature of financial markets. Traditional methods heavily depend on historical price data but often fail to capture the influence of external factors such as news sentiment, reflecting current global trends. Deep learning techniques have emerged as a promising course for improving prediction accuracy in recent years. By leveraging neural networks to analyze sequential data, these methods offer a potential solution to the limitations of traditional approaches.

In this paper, we propose a deep learning-based approach that combines LSTM networks with sentiment analysis to enhance stock market prediction. This novel integration aims to harness the strengths of both methodologies: the temporal understanding of LSTM networks and the insights derived from sentiment analysis. By synergizing these two techniques, we seek to unlock new opportunities for more accurate and nuanced predictions in the realm of stock market forecasting.

Our research endeavors to explore the transformative potential of deep learning in financial markets. Through a rigorous examination of our proposed methodology and empirical validation of its effectiveness, we aim to contribute valuable insights to the field of predictive analytics. By presenting our findings in a clear and concise manner, we strive to provide a solid foundation for further advancements in stock market prediction, ultimately facilitating more informed decision-making in the financial domain.

Identify applicable funding agency here. If none, delete this.

II. Related Work

Several studies have explored the use of sentiment analysis and deep learning for stock market prediction, each offering unique insights and methodologies. Gupta and Chen [1] employed sentiment analysis techniques such as Naive-Bayes and Logistic Regression, combined with featurization methods like TF-IDF, achieving accuracy levels ranging from 75% to 85%. This approach highlights the importance of leveraging both sentiment analysis and featurization techniques to enhance prediction accuracy.

Building upon this foundation, Bharti and Gupta [2] introduced a novel approach by utilizing the VADER lexicon and implementing a custom Convolutional Neural Network (CNN) for sentiment analysis. Their methodology, supplemented by ensemble learning techniques for trend prediction, showcased the versatility of deep learning in capturing complex market dynamics. This underscores the significance of incorporating

advanced deep learning architectures for robust prediction models.

In a similar vein, Selvin et al. [3] conducted a comprehensive comparison of deep learning models, including LSTM, RNN, and CNN, for stock price prediction. Their findings highlighted the superiority of CNNs, particularly in capturing short-term trends in stock prices. This emphasizes the importance of selecting appropriate model architectures tailored to the specific nuances of financial data.

Additionally, Xu and Keselj [4] contributed to the discourse by demonstrating the effectiveness of deep learning models over traditional methods. Their study showcased the superiority of CNNs in capturing dynamic changes in stock prices, outperforming both RNNs and LSTMs. This reinforces the notion that deep learning techniques offer promising avenues for enhancing prediction accuracy and capturing the inherent complexities of financial markets.

In summary, the literature review provides compelling evidence of the growing interest in leveraging deep learning and sentiment analysis for stock market prediction. While traditional methodologies remain prevalent, the advent of deep learning techniques presents a paradigm shift in the field. By incorporating unstructured data such as news sentiment, deep learning models offer promising avenues for advancing the state-of-the-art in stock market prediction.

III. Proposed Approach

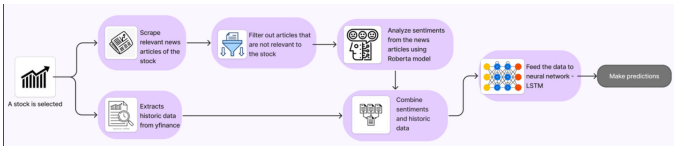

Fig. 1. The proposed architecture

Our proposed methodology encompasses a comprehensive workflow designed to leverage the synergy between historical price data and news sentiment for stock market prediction. At the core of our approach lies a meticulously orchestrated series of steps aimed at harnessing the collective power of data integration, sentiment analysis, and deep learning techniques.

To initiate the process, we begin by sourcing relevant news articles through web scraping techniques, scouring reputable sources for insights pertinent to the target stock. Leveraging the Yahoo Finance API, we concurrently retrieve historical stock prices, spanning a period conducive to meaningful analysis. This dual-pronged data acquisition strategy ensures a robust foundation for subsequent analysis, encompassing both temporal trends and contemporary market sentiments.

With the raw data in hand, we embark on the task of preprocessing and refining the news headlines to distill key insights while filtering out extraneous noise. This critical step involves the application of advanced natural language processing techniques to extract sentiment features from the textual corpus. Employing state-of-the-art models such as Roberta, we delve deep into the semantic nuances of the headlines, discerning underlying sentiments with precision and accuracy.

The culmination of these preparatory steps sets the stage for the integration of sentiment features with historical price data, a pivotal moment in our methodology. Through careful alignment and harmonization of these disparate data streams, we create a unified dataset primed for predictive analysis. This cohesive dataset, comprising a rich tapestry of temporal and sentiment-based signals, serves as the cornerstone of our predictive model's training regimen.

At the heart of our methodology lies the LSTM neural network, a versatile and powerful tool for sequential data analysis. Through iterative training and refinement, the LSTM network learns to decipher the complex interplay between historical price movements and contemporaneous market sentiments. Armed with this newfound understanding, the model endeavors to extrapolate future stock price trends with unprecedented accuracy and foresight.

In essence, our proposed methodology represents a holistic approach to stock market prediction, integrating cutting-edge technologies and advanced analytical techniques to unlock new frontiers in predictive analytics. By seamlessly blending historical data with real-time sentiment analysis, we aspire to empower investors and financial analysts with actionable insights, enabling more informed decision-making in today's dynamic and ever-evolving markets.

A. Dataset creation

Central to our research is the development of a robust dataset adapted to our objectives. Our dataset focuses on historical prices of Tata Motors Ltd Stock, identified by the ticker symbol "TATAMOTORS.NS", spanning five years from March 13, 2019, to March 13, 2024, comprising 1639 meticulously curated data points. Creating this dataset involved several meticulous steps. Initially, historical stock prices were gathered using the Yfinance API. Simultaneously, news headlines relevant to Tata Motors were carefully scraped from reputable sources. These articles were then filtered to ensure accuracy and relevance, with keywords extracted to refine the selection process. By combining historical prices with filtered news headlines, our dataset was structured into a CSV format, providing a solid foundation for our predictive model's development.

This meticulously curated dataset underpins our research, providing the necessary framework for accurate analysis and prediction.

B. LSTM architecture

The LSTM architecture consists of multiple LSTM layers stacked on top of each other, allowing the network to capture complex temporal dependencies in the input time series data. Each LSTM layer comprises several memory cells, input gates, output gates, and forget gates, which collectively enable the model to retain long-term information and selectively update its internal state. The mathematical expressions governing the behavior of an LSTM unit are as follows:

C. Equations

$$
\begin{aligned}
f_t &= \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \\
i_t &= \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \\
\tilde{C}_t &= \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
o_t &= \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \\
h_t &= o_t * \tanh\left(C_t\right)
\end{aligned}
$$

Where: - f_t is the forget gate - i_t is the input gate - $\tilde{C}_t$ is the candidate cell state - C_t is the cell state - o_t is the output gate - h_t is the hidden state - σ represents the sigmoid activation function - W_f, W_i, W_C, W_o are weight matrices - b_f, b_i, b_C, b_o are bias vectors

D. Training Algorithm

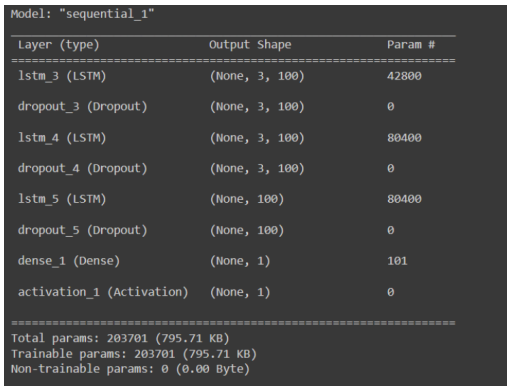

Fig. 2. The LSTM model

The training process involves optimizing the model parameters to minimize the prediction error between the predicted and actual stock prices. We employ the Adam optimizer with a learning rate of alpha to update the network weights iteratively. The loss function used for training is Mean Squared Error (MSE), which measures the discrepancy between the predicted and actual stock prices over the training dataset. The algorithm for training the LSTM model is as follows:

- Initialize the LSTM model with random weights and biases.
- Split our dataset into training and validation sets.
- Iterate over the training set in mini-batches.
- Forward propagate the input data through the LSTM layers to compute the predicted stock prices.
- Calculate the loss between the predicted and actual prices using MSE.
- Backpropagate the error through the network to update the weights using the Adam optimizer.
- Repeat steps 4-6 until convergence or a predefined number of epochs.
- Evaluate the model performance on the validation set to monitor for overfitting.
- The training algorithm aims to optimize the LSTM model parameters to learn the underlying patterns and relationships in the data, enabling accurate prediction of future stock prices.

IV. Result

Based on the result obtained from the model, it is evident that our approach has yielded promising outcomes in predicting stock price movements. The Root Mean Square Error (RMSE) values for both the training and testing datasets, 14.71 and 16.10 respectively, indicate relatively low levels of prediction error. This suggests that, on average, the model's predictions deviate from the actual values by a reasonable margin. Additionally, the high training and testing accuracies, 94.91% and 96.19% respectively, imply that the model's predictions closely align with the actual values, demonstrating its effectiveness in capturing the underlying patterns and relationships in the data.

V. Conclusion and Future Work

In conclusion, our study introduces an innovative methodology for enhancing stock market prediction by synergizing LSTM neural networks with sentiment analysis of news headlines. Through meticulous experimentation and analysis, we have demonstrated the potential of deep learning techniques in elucidating the intricate interplay between news sentiment and stock price dynamics.

Our preliminary findings reveal encouraging performance metrics, reaffirming the effectiveness of our proposed approach in capturing the nuanced relationships inherent in financial markets. By leveraging LSTM networks to model temporal dependencies and sentiment analysis to gauge market sentiment, we have laid the groundwork for a robust predictive framework capable of providing valuable insights into stock price movements.

Looking ahead, several avenues for future research beckon. One promising direction involves exploring alternative deep learning architectures to further enhance prediction accuracy and robustness. Experimenting with variations of LSTM networks, such as bidirectional LSTMs or attention mechanisms, could offer valuable insights into optimizing model performance.

Additionally, there is a pressing need to refine sentiment analysis methods to better capture the subtleties of market sentiment. Investigating advanced natural language processing techniques and sentiment lexicons tailored specifically for financial markets could yield more nuanced sentiment features, thereby enriching our predictive model's capabilities.

Furthermore, extending our framework to incorporate additional features beyond historical price data and news sentiment holds promise for improving prediction accuracy. Incorporating macroeconomic indicators, industry-specific trends, and geopolitical factors could provide a more comprehensive understanding of market dynamics, thereby enhancing the predictive power of our model.

In essence, our research lays the basic foundation for future advancements in stock market prediction, offering a roadmap for leveraging deep learning and sentiment analysis to navigate the complexities of financial markets with precision and insight. By continuing to explore different deep learning techniques and refining our methodologies, we aim to contribute to the ongoing evolution of predictive analytics in the realm of finance, allowing investors and financial analysts with the tools they need to make informed decisions in an increasingly complex but budding landscape.

REFERENCES

[1] R. Gupta and M. Chen, "Sentiment Analysis for Stock Price Prediction," 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China, 2020, pp. 213-218, doi: 10.1109/MIPR49039.2020.00051.

[2] S. K. Bharti and R. K. Gupta, "Stock Market Price Prediction via Sentiment Analysis & Ensemble Learning," 2022 IEEE Conference on Sustainable Energy, Gunupur, Odisha, India, 2022, pp. 1-5, doi: 10.1109/iSSSC56467.2022.10051623.

[3] Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 2017, pp. 1643-1647, doi: 10.1109/ICACCI.2017.8126078.

[4] Y. Xu and V. Keselj, "Stock Prediction using Deep Learning and Sentiment Analysis," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 5573-5580, doi: 10.1109/BigData47090.2019.9006342.