Sentiment Augmented LSTM for Stock Market Trend Prediction

Aman Raj[1], Prerana Sharma[2]

[1,2]Department of Artificial Intelligence and Data Science/IIMT College of Engineering/Greater Noida, India

Abstract

In this paper, we present a real-time stock price prediction system that combines Long Short-Term Memory (LSTM) neural networks with sentiment analysis of social media data. By integrating numerical stock data with public sentiment, our system aims to improve the accuracy of short-term stock price forecasting. The system ingests live stock price data and social media posts related to specific companies, analyzes sentiment using FinBERT, and trains an LSTM model to predict future stock trends. The architecture utilizes Django, Celery, Redis, PostgreSQL, and Docker to ensure modularity, scalability, and robustness. This approach acknowledges the limitations of relying purely on quantitative indicators and emphasizes the increasing influence of behavioral economics in financial markets. As sentiment reflects the collective psychological state of investors, incorporating it alongside traditional indicators represents a more holistic view of stock market dynamics.

Keywords: Stock Market Prediction, LSTM, Sentiment Analysis, FinBERT, Real-Time Forecasting, Social Media Data, Deep Learning, Financial Modeling, Django, Docker

I. INTRODUCTION

Stock price prediction has long been a focal point of financial research and practical investment strategies. The ability to accurately predict stock market movements offers a substantial advantage to investors, traders, and analysts, providing insights into when to buy or sell assets. Traditional forecasting models, such as autoregressive integrated moving average (**ARIMA**) and exponential smoothing models, have been widely used to predict stock prices. These models rely heavily on historical price data and technical indicators, such as **moving averages**, **relative strength index (RSI)**, and **Bollinger Bands**. While these quantitative models can capture certain market trends, they often overlook or fail to account for the broader influences on the market that arise from external factors, such as market sentiment, investor behavior, and news events [1].

Over the past decade, the rise of social media platforms—especially Twitter—has drastically changed how market participants disseminate and receive information. Tweets, news articles, and social media discussions now play a significant role in shaping market sentiment, which in turn influences price movements. Public sentiment, driven by the opinions, attitudes, and emotions expressed online, has been shown to significantly impact stock market behavior. A wide body of research indicates that investor sentiment can be an important predictor of stock market trends, sometimes more so than traditional financial indicators [2].

The influence of behavioral finance on market dynamics has led to a growing recognition of the need to incorporate sentiment analysis into forecasting models. Behavioral economics, which examines how psychological factors and emotions influence market participants' decisions, plays a crucial role in understanding stock price volatility. Investors' collective emotions and psychological biases—such as fear, greed, optimism, and pessimism—can create large, short-term swings in stock prices, especially in response to news events, earnings reports, or other major announcements. Therefore, relying solely on historical price data and technical indicators is becoming increasingly inadequate for accurate short-term forecasting, especially when market sentiment can dramatically change in response to emerging news [3].

This research paper presents a novel approach to stock price prediction that integrates both **LSTM**-based time-series forecasting and sentiment analysis of social media data. By combining traditional numerical stock data (e.g., OHLC prices

and trading volume) with sentiment scores derived from real-time tweets, the proposed model seeks to provide a more holistic and accurate prediction of short-term stock trends. The system not only uses historical data to predict price movements but also incorporates market sentiment, enabling it to better reflect investor psychology and emotions that drive price fluctuations. By integrating sentiment analysis into stock price prediction, this paper aims to enhance the model's ability to adapt to changing market conditions and improve the accuracy of real-time forecasting.

II. LITERATURE SURVEY

Previous research has demonstrated that public sentiment derived from social media can significantly influence stock market performance. **Bollen et al. (2011)** explored the relationship between Twitter mood and the Dow Jones Industrial Average, finding correlations between social media sentiment and stock market movements [2]. In their study, positive moods correlated with higher stock market returns, while negative moods were linked to downward price movements. This work laid the foundation for future sentiment-based financial models.

Si et al. (2013) furthered this idea by employing sentiment analysis for predicting various aspects of financial markets, such as volatility, trading volume, and price direction. The authors used machine learning models to assess the impact of sentiment on stock movements and found that sentiment information significantly improved prediction accuracy, particularly in volatile market conditions [4]. Similarly, **Zhang et al. (2020)** demonstrated that sentiment analysis from social media data can be used to predict the movement of stock prices, thereby enhancing the forecasting ability of traditional models. They emphasized the importance of sentiment in forecasting price direction, showing that market psychology has become a key factor in predicting stock behavior.

LSTM models have emerged as a promising solution for time-series forecasting due to their ability to learn long-term dependencies in data sequences. **Hochreiter and Schmidhuber (1997)** first introduced the LSTM model, which overcomes the limitations of standard **Recurrent Neural Networks (RNNs)** by preserving important information over long sequences [5]. LSTMs have been widely applied to stock market prediction due to their ability to capture complex patterns in price movements, which are often influenced by events spanning multiple time intervals. The ability to retain information over time makes LSTM models particularly suited for financial prediction tasks, where market behavior is often influenced by both short-term events and long-term trends [6].

A hybrid approach that merges sentiment analysis with LSTM-based forecasting has shown promising results in improving prediction accuracy. Several studies have explored the combination of deep learning techniques and sentiment analysis to predict stock prices more accurately. For example, **Zhang and Yang (2020)** proposed an LSTM-based model that integrates sentiment scores derived from Twitter data and showed that it outperforms traditional LSTM models that rely solely on price data [7]. This hybrid model provides a more comprehensive approach by incorporating both numerical data and emotional factors into the prediction process.

Besides, a few scholars have aimed to improve sentiment analysis models that have been specifically created for financial markets. The **FinBERT** model developed by **Liu and Wei (2020)** was specifically optimized to be used in financial sentiment analysis. Unlike typical sentiment analysis models, FinBERT is trained with financial texts and therefore is even more appropriate when analyzing sentiment from stock-related documents and news reporting. The authors demonstrated that FinBERT could perform better in market trend prediction than general-purpose sentiment models [8]. This paper emphasizes the value of applying domain-specific models, such as FinBERT, to identify the intricate use of language in finance accurately.

Whereas earlier works have dealt with sentiment analysis and LSTM individually, not many have attempted their real-time integration in an application-ready system. This work bridges that gap by developing a complete stack system with **real-time sentiment analysis** coupled with LSTM-based predictions and delivering a scalable stock market prediction solution. The system makes use of **Django**, **Celery**, **Redis**, **PostgreSQL**, and **Docker** to provide high availability, modularity, as well as scalability, making it ready for production deployment in live trading systems [9].

III. METHODOLOGY

A. System Architecture

The system consists of multiple coordinated components:

(i)Data Ingestion: Real-time stock data is fetched using APIs like Alpha Vantage. Twitter data is scraped using snscrape, capturing tweets based on hashtags, ticker symbols, and company names.

(ii)Sentiment Analysis: Tweets are cleaned and analyzed using FinBERT, a BERT-based model fine-tuned for financial sentiment. The output is a sentiment score ranging from 1 to 10, reflecting the positive or negative sentiment of each tweet.

(iii)Data Storage: PostgreSQL is used to store the stock data, tweet content, sentiment scores, and the merged datasets for further analysis.

(iv)Task Scheduling: Celery with Redis schedules periodic tasks for data ingestion and processing.

(v)Prediction Model: An LSTM model is trained on the merged dataset and deployed using Docker containers to ensure easy scalability and deployment.

(vi)Web Interface: Django offers a user interface to visualize data and predictions, allowing users to track real-time trends.

B. Data Processing Pipeline

The system follows these steps to process data:

(i)Ingest real-time stock prices.

(ii)Scrape tweets related to selected tickers.

(iii)Clean and preprocess text (remove stopwords, punctuation, etc.).

(iv)Analyze the sentiment with FinBERT (scale: 1 to 10).

(v)Aggregate sentiment scores by ticker and time frame.

(vi)Merge sentiment and stock data by timestamp.

(vii)Store data in PostgreSQL.

(viii)Train LSTM model periodically using the merged dataset.

C. Model Design and Training

The **LSTM** model, implemented in Keras with TensorFlow, uses the following features:

(i)OHLC (Open, High, Low, Close), volume, moving averages

(ii)Aggregated sentiment statistics (mean, variance)

(iii)The model architecture includes:

(iv)Input normalization layer

(v)LSTM layer (20 units)

(vi)Dropout (0.2)

(vii)Dense + ReLU layer

(viii)Output layer with linear activation

Training is done using Adam optimizer and Mean Squared Error (MSE) loss function. Early stopping and checkpoints prevent overfitting, and the model is retrained weekly [6].

IV. RESULTS AND EVALUATION

A. Model Performance

The proposed model was tested using historical stock price data from January 2023 to March 2023. The data was divided into a training set (70%) and a test set (30%). We evaluated the performance of two models:

(i)Stock-Only Model: This model used only historical stock data (OHLC and volume) to predict future stock prices.

(ii)Sentiment-Augmented Model: This model combined both historical stock data and aggregated sentiment scores derived from real-time Twitter data using FinBERT sentiment analysis.

The Mean Squared Error (MSE) and Mean Absolute Error (MAE) were used as performance metrics for evaluation. The MSE measures the average squared difference between predicted and actual values, while MAE gives the average absolute difference. Both metrics are commonly used for regression problems, such as stock price prediction.

Model	MSE	MAE
Stock-Only Model	0.028	0.19
Sentiment-Augmented Model	0.021	0.14

B. Model Comparison

The results show that the sentiment-augmented model outperforms the stock-only model in both MSE and MAE, demonstrating the effectiveness of integrating sentiment data into stock price prediction. Specifically, the Sentiment-Augmented Model:

(i)**Reduced the MSE by 0.007**, indicating that it makes more accurate predictions in terms of squared error.

(ii)**Reduced the MAE by 0.05**, meaning that the absolute difference between predicted and actual stock prices is lower, which is crucial for financial decision-making where small price differences can have significant impacts.

In particular, the sentiment-augmented model showed a better capacity for capturing price movements during significant market events that were reflected in social media sentiment. For example, the announcement of layoffs in a major tech company triggered a sudden shift in sentiment on Twitter, which was accurately reflected in the model's predictions for the company's stock price the following day. This indicates that social media sentiment can provide valuable real-time insights that may not be captured solely through traditional financial indicators.

C. Impact of Sentiment Analysis

One of the most significant outcomes from this evaluation was the observed correlation between sentiment data and stock price fluctuations during high volatility periods. In particular:

(i)The stock-only model had difficulties predicting stock price drops or spikes during earnings season or when negative news events (e.g., scandals, layoffs, or government regulations) occurred.

(ii)The sentiment-augmented model, however, was able to detect these shifts in sentiment and provide more accurate predictions for the next day's price movements.

This demonstrates that sentiment analysis not only augments quantitative analysis but also serves as a valuable leading indicator of stock price direction. While traditional models might capture the trend over time, sentiment analysis can capture short-term shifts in market sentiment that influence immediate price changes.

D. Real-Time Testing and Scalability

The system was tested in real-time using Django and Celery for task scheduling. Tweets were scraped every 15 minutes, and sentiment scores were updated for each relevant ticker symbol. The Dockerized architecture ensured that the system was highly scalable, with the ability to handle an increasing volume of incoming tweets and stock data.

The performance in real-time testing was also in agreement with the offline test, ensuring that the system can process live data and make predictions in real-time. The system could make predictions with low latency, enabling timely decisions in trading settings.

V. FUTURE SCOPE

A. Multilingual Sentiment Analysis

The current model relies on English-language Twitter data, which limits its applicability to English-speaking markets. Given the global nature of financial markets, it would be beneficial to expand the system to support multiple languages. This could be achieved by:

(i)Fine-tuning sentiment analysis models like FinBERT for multiple languages, such as Spanish, German, Chinese, and French.

(ii)Using existing multilingual models like XLM-RoBERTa for cross-lingual sentiment analysis.

(iii)Expanding the system's capabilities to handle global sentiment would allow it to be used in markets outside the English-speaking world, thus broadening its applicability.

B. Incorporating Additional News Sources

While social media sentiment is a powerful predictor of stock market movements, it can sometimes be noisy and biased. To improve the quality of sentiment analysis and enrich the dataset, it would be beneficial to integrate additional sources

of information, such as:

(i)Financial News Articles: Integrating news scraping techniques to extract key information from financial news websites (e.g., Bloomberg, Reuters, CNBC).

(ii)Earnings Reports: Using Natural Language Processing (NLP) techniques to analyze the tone and content of quarterly earnings reports, which often drive stock price movements.

(iii)Government Announcements: Incorporating sentiment from regulatory filings and policy announcements (e.g., central bank interest rate decisions, government stimulus packages).

(iv)By integrating a broader range of news sources, the model would be able to account for events that affect stock prices but may not be immediately discussed on social media.

C. Incorporating Advanced Techniques like Reinforcement Learning

Currently, the system relies on supervised learning for stock price prediction. However, a reinforcement learning (RL)-based approach could improve the prediction model by enabling it to adapt to market changes and optimize its actions (e.g., buying, holding, selling). In a reinforcement learning model, the agent (stock prediction system) learns from its past actions and continuously improves its trading strategy by maximizing a reward function, such as profit.

Deep Q-Learning or Proximal Policy Optimization (PPO) algorithms could be applied to fine-tune decision-making for stock trading. This would allow the system to not only predict stock prices but also provide actionable recommendations based on changing market conditions, further enhancing its utility for automated trading systems.

D. Sentiment Analysis Across More Platforms

Currently, the system uses Twitter data for sentiment analysis. However, other platforms such as **Reddit**, **StockTwits**, and news aggregation websites like **Google News** also contain valuable sentiment data. Incorporating sentiment analysis from these platforms could provide a more holistic view of market sentiment. For instance, Reddit's **WallStreetBets** forum has been a significant source of discussion for stocks with high volatility, and analyzing posts from such communities could offer further insights into market movements.

E. Incorporating Technical Indicators for Hybrid Prediction

Although sentiment analysis improves model performance, it can be even more powerful when combined with traditional technical analysis indicators. Technical indicators such as Relative **Strength Index (RSI)**, **Moving Average Convergence Divergence (MACD)**, and **Bollinger Bands** provide critical information about stock price trends. Integrating these technical indicators into the existing sentiment-augmented **LSTM** model could further refine predictions by capturing both market psychology (sentiment) and market behavior (technical analysis).

F. Personalized Stock Predictions

In the future, the system could be adapted to provide personalized stock predictions based on individual investor profiles. This would involve integrating personalized data such as:

(i)Investor risk tolerance

(ii)Investment goals (short-term vs long-term)

(iii)Stock portfolio composition

By analyzing the historical performance of an investor's portfolio along with real-time sentiment and stock data, the system could provide customized predictions and recommendations, thereby enhancing the trading experience for individual investors and fund managers.

G. Improving Model Explainability and Transparency

Currently, the **LSTM** model works as a black box, making it difficult to interpret the underlying reasons behind its predictions. To increase model transparency, future research could explore explainable AI (XAI) techniques to understand how the model arrived at specific predictions. Techniques such as **SHAP values** and **LIME (Local Interpretable Model-Agnostic Explanations)** could be used to provide interpretability to the model's decision-making process, which is crucial in financial markets where understanding the rationale behind predictions is important for decision-making.

VI. CONCLUSION

The proposed sentiment-augmented LSTM model significantly improves the accuracy of stock price predictions by incorporating both quantitative stock data and sentiment analysis from social media. By combining these two data sources, the model is able to capture both market trends and investor emotions, providing a more comprehensive view of stock price behavior. The results demonstrate the power of integrating behavioral finance and deep learning to enhance stock

market forecasting, especially during volatile market conditions.

While the current system offers promising results, there are several avenues for future improvements, including multilingual sentiment analysis, incorporating additional news sources, exploring reinforcement learning, and expanding the system to multiple platforms. By refining the model and adding new features, this approach has the potential to become a valuable tool for both individual and institutional investors in the fast-paced world of stock trading.

VII. REFERENCES

[1] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science.
[2] Shiller, R. J. (2000). Irrational Exuberance. Princeton University Press.
[3] Si, L., & Yang, D. (2013). Sentiment analysis for financial prediction tasks. Financial Technology.
[4] Zhang, L., & Yang, L. (2020). Stock price prediction using sentiment analysis. Financial Innovation.
[5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation.
[6] Liu, X., & Wei, J. (2020). FinBERT: A Pretrained Language Model for Financial Text Analysis.
[7] arXiv. Zhang, H., & Lee, J. (2019). Predicting stock price using sentiment analysis. AI for Financial Systems.
[8] Dube, R., & Jadhav, A. (2020). A study on integrating sentiment analysis and LSTM for stock price prediction. Journal of Financial Data Science.
[9] Li, X., & Zhou, Y. (2021). Real-time stock prediction with hybrid models. Neural Networks.