2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)

2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE) | 979-8-3503-2820-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC-ETITE58242.2024.10493402

Sentiment Analysis on Reddit Trading Data

Govinda K
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
Tamil Nadu, India
kgovinda@vit.ac.in

Akhil Chintalapati
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
Tamil Nadu, India
akhil.chintalapati01@gmail.com

Aparajita Senapati
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
Tamil Nadu, India
aparajita1603@gmail.com

Khashbhat Enkhbat
Economics, Digital Business and Innovation Department
Tokyo International University
Saitama, Japan
khashbat.enkh@gmail.com

***Abstract*— In the volatile world of trading, success frequently depends on the ability to devise and implement effective methods. Understanding market sentiment has been increasingly important for stabilizing revenue and separating strategic trading from simply speculation since the advent of internet trading forums such as those on Reddit. This research offers a novel machine learning model that uses sentiment analysis to evaluate Reddit trading data, with the goal of providing traders with a better-informed decision-making base. This model tries to decipher the frequently complicated and trend-driven conversation within trading groups by merging techniques from machine learning, deep learning, and mathematical transformations—including VADER Analysis and Fourier Transforms, complemented by Long-Short Term Memory networks. The idea is to turn what often appears to be gambling into a more systematic approach to trading led by data-driven insights into market emotion.**

Keywords— Sentiment Analysis, Strategy Development, VADER Analysis, Fourier Transforms, Long-Short Term Memory

I. INTRODUCTION

Trading has evolved as both a popular route for investment and a subject of significant scholarly study in the ever-changing world of financial markets. The introduction of internet trading platforms, as well as the growth of social media forums such as Reddit, have drastically altered the trade dynamics. These platforms not only democratise access to financial markets, but they also act as crucial exchange points for information, opinions, and plans. This paper delves into sentiment analysis as it applies to Reddit trading data, with the goal of uncovering the underlying feelings that drive market patterns and investor behaviour.

Trading is a tough yet fascinating topic of study due to its unpredictable nature, which is influenced by a plethora of factors ranging from global economic movements to individual investor psychology. Traditional financial research has frequently failed to capture the market's swiftly changing sentiments, a gap that sentiment analysis seeks to remedy. Sentiment analysis, a subfield of artificial intelligence, aims to recognise, extract, measure, and investigate emotive states and subjective information in a systematic manner. This study aims to provide unique insights into the collective mood of the trading community by applying this technique to online trading discussions, particularly those on Reddit, which can have major repercussions for market movements.

With its wide and active user base, Reddit has become a hub for traders ranging from newbies to seasoned pros. The importance of forums like 'r/wallstreetbets' in key market events like the GameStop short squeeze has highlighted the platform's influence on trading decisions. The open and informal character of Reddit communication provides a rich, albeit complex, dataset for sentiment analysis. The difficulty stems from the intricacies of online discourse—slang, sarcasm, and frequently shifting topics—which necessitate sophisticated models to effectively comprehend.

The research provides a comprehensive machine learning model that incorporates advanced techniques such as VADER Analysis, Fourier Transforms, and Long-Short Term Memory networks to handle this difficulty. These methods were chosen for their capacity to capture both short-term and long-term sentiment trends in complicated and huge datasets, such as those from Reddit. This model is more than a technical exercise; it has real-world consequences for traders and investors, providing a more data-driven approach to interpreting market mood and assisting in strategy building.

To summarise, the purpose of this work is to connect the qualitative nature of online trading discussions and the quantitative rigour of financial analysis. It aspires to enlighten the often opaque world of trading by leveraging the power of sentiment analysis and machine learning, delivering insights that could lead to more educated and strategic trading decisions.

II. LITERATURE REVIEW

In recent years, there has been substantial progress in the field of sentiment analysis in financial markets, particularly through the use of social media and online forums. Previous research has set a solid basis by investigating numerous approaches and data sources to better understand market sentiments and their impact on trade. Our work, on the other hand, stands out due to its specialized emphasis, techniques, and novel application.

Ching-Ru Ko and Hsien-Tsung C [1] pioneered the analysis of information from news and numerous forums, laying the groundwork for the use of diverse data sources. While based on the concept of multi-source data analysis, our approach narrows its attention especially to Reddit, providing a more in-depth and platform-specific insight.

Masoud M, et al [2], as well as Venkata S P, et al [3], established the utility of sentiment analysis in forecasting stock

market fluctuations. Our research expands on these findings by including a more advanced machine learning model that incorporates VADER Analysis, Fourier Transforms, and LSTM networks, hence improving forecast accuracy and applicability in real-time trading settings.

Franco V et al. [4] and Siavesh K et al. [5] focused on sentiment analysis in the Bitcoin and larger stock markets, respectively. Our study adds to this body of knowledge by focusing on the Reddit trading community, which has emerged as a prominent influencer in recent market events.

Wenbin Z, Steven S [6] and Khurshid A, et al [7] investigated the effect of media sentiment on stock volumes and market dynamics. We expand on this by combining sentiment analysis with advanced machine learning techniques to develop a more comprehensive model.

E. Naresh, B. J. Ananda, K. S. Keerthi, and M. R. Tejonidhi [10], as well as Q. Wang [12], predicted stock prices using social media and internet terms. Our approach is unique in that it uses not only social media data but also a novel combination of analytical approaches to extract more nuanced sentiment insights from Reddit discussions.

In a study by Guo [8], the integration of sentiment analysis derived from news articles into the stock price prediction model was investigated. By employing Long Short-Term Memory (LSTM) neural networks and incorporating sentiment scores from the New York Times articles, Guo aimed to enhance the prediction accuracy of stock close prices and returns. The study specifically focused on comparing models that included sentiment analysis with those that solely relied on historical stock data. The results indicated a notable improvement in prediction accuracy when sentiment data were included, highlighting the potential of sentiment analysis in financial forecasting.

Furthering this line of research, Sidogi, Mbuvha, and Marwala [9] examined the impact of financial news sentiment on stock price predictability using LSTM networks. Their approach utilized FinBERT, a version of the BERT model fine-tuned for financial contexts, to analyze sentiment in company-related news headlines. By comparing models with and without sentiment data from FinBERT, they found that incorporating domain-specific sentiment analysis significantly improved the models' predictive performance in terms of RMSE and MAE. This study underscored the importance of domain-specific fine-tuning in enhancing the effectiveness of sentiment analysis for stock price prediction.

On another front, Q. Wang [12] approached the prediction of the Chinese stock market using a different type of alternative data: internet keyword popularity. Through statistical time series regression analysis, Wang assessed the impact of the popularity of specific keywords on social media platforms on stock prices. The study utilized the Granger causality test to determine the relationship between keyword popularity and stock market performance. The findings suggested that while keyword popularity did influence stock prices, its effect was largely consistent with overall market influences, indicating an efficient market. However, the study also noted that the nature and audience of the keywords could affect their impact on the market, with broader appeal leading to a more significant influence.

Building on the momentum of integrating advanced analytics into financial forecasting, Rekha G., Bhanu Sravanthi D., Ramasubbareddy S., and Govinda K. [11] delve into the application of neural network strategies, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for predicting stock market trends. Their study focuses on comparing these non-linear models to determine which better aligns with actual market movements, aiming to optimize investment returns. By evaluating the predictive accuracy of CNNs versus RNNs, this research sheds light on the most effective neural network approach for stock market forecasting, offering valuable insights for investors and financial analysts in strategy selection.

Mohbey, K. & Khan, Mohammad & Indian, Ajay [13], Pandey, Arvind & Shukla, Shipra & Mohbey, K. [14], and Meena, Gaurav & Mohbey, K. & Indian, Ajay [15] expanded the application of machine learning and sentiment analysis in various financial and online contexts. Our study adds to this growing body of work by applying these techniques to a new and quickly evolving data source—Reddit trading forums—and constructing a model customised to the platform's specific characteristics.

Finally, while earlier research has set the framework for sentiment analysis and its use in financial markets, our study presents a fresh approach customised to the particular environment of Reddit.

Our study offers new insights and tools for traders and investors by integrating powerful machine learning techniques with a detailed investigation of Reddit trading data, significantly expanding the field of sentiment analysis in finance.

III. Methodology

Our research employs a three-phase architecture to correlate public sentiment with stock price volatility, as illustrated in Fig. 1. This methodology is designed to rigorously analyse sentiment data from Reddit and apply it to develop trading strategies.

A. Sentiment Analysis

The initial phase focuses on sentiment analysis, comprising data collection and preprocessing, followed by sentiment scoring.

1. Positive-Negative Sentiment Analysis: Post preprocessing, the Valence Aware Dictionary and Sentiment Reasoner (VADER) is utilized. VADER scores the data in terms of positive and negative values. This scoring is based on a valence-specific dictionary, allowing us to align sentiment scores with specific stock prices on corresponding dates. The mathematical model for VADER sentiment scoring is given in (1).

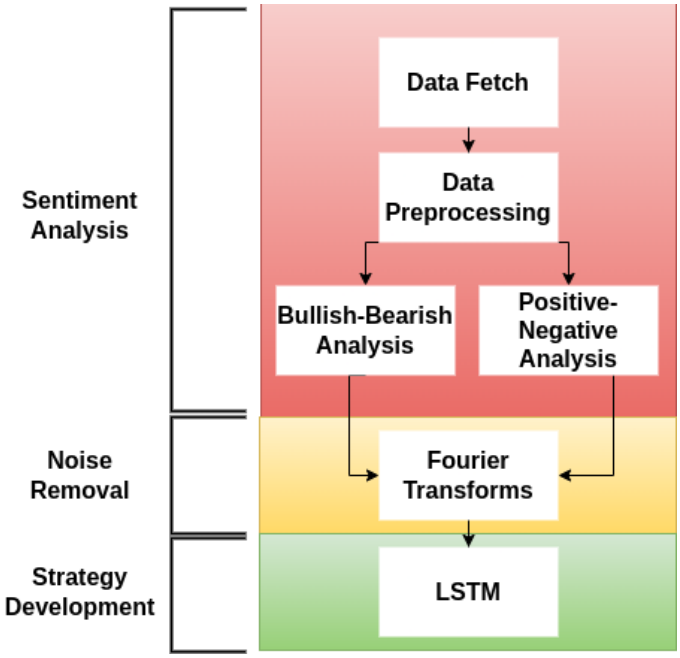

Fig. 1. Proposed Architeture.

In (1), the valence function represents the valence score of each word in the post, and N is the total number of words.

$$\text{Sentiment Score} = \sum_{i=1}^{N} \text{Valence}(Wi) \quad (1)$$

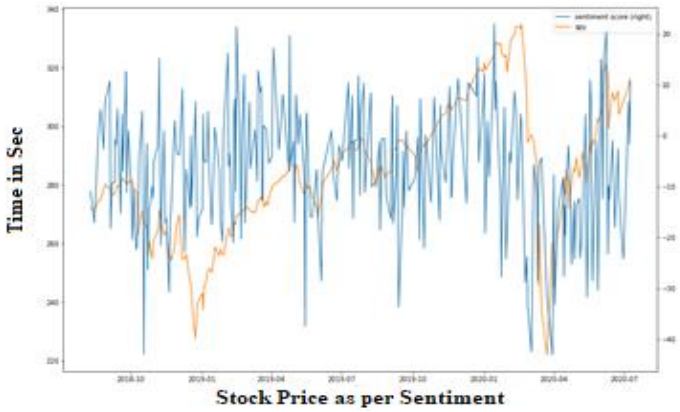

Fig. 2. Positive-Negative Sentiment Analysis, where X-Axis denotes the date and Y-Axis denotes the stock price and the sentiment score.

2. Bullish-Bearish Sentiment Analysis: After preprocessing, the data undergoes analysis using two custom dictionaries tailored for 'bullish' and 'bearish' sentiments. These dictionaries comprise terms frequently used to express market confidence or pessimism, enabling us to score the sentiment related to specific stock prices. The sentiment score is computed as shown in (2).

$$\frac{\text{Bullish}}{\text{Bearish}} \text{ Score} = \frac{\text{Count of Bullish/Bearish words}}{\text{Total words in post}} \quad (2)$$

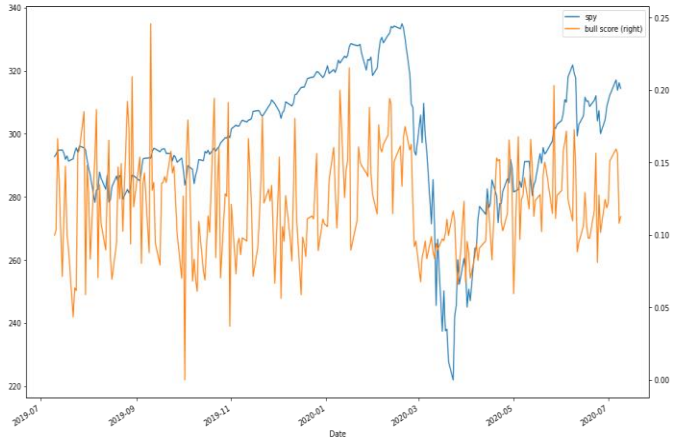

Fig. 3. Bullish Sentiment Analysis, where X-Axis denotes the date and Y-Axis denotes the stock price and the sentiment score.

Fig. 2, 3, and 4 are critical in depicting the relationship between market sentiment and stock prices. The Positive-Negative Sentiment Analysis is depicted in Fig. 2, where the X-Axis depicts dates and the Y-Axis displays both the stock price and the sentiment score. This graph is essential for comprehending how general market sentiment, both good and negative, correlates with company price changes. Fig. 3 depicts the Bullish Sentiment Analysis, with the X-Axis representing the date and the Y-Axis representing the stock price and sentiment score. This graphic representation is critical for understanding the market's reaction to optimistic sentiment.

B. Noise Removal

The second phase entails noise elimination, which is an important step in refining the sentiment analysis results. This procedure deals with probable outliers and negative numbers that could distort the study.

We employ the Fourier Transform method for its effectiveness in minimizing amplitude variations caused by rapid fluctuations in data. As demonstrated in Fig. 5, 7, and 9, multiple cycles of this process significantly reduce noise, eventually rendering it negligible. This results in a more discernible linear trend and normalization of amplitude closeness, which are crucial for accurate strategy development.

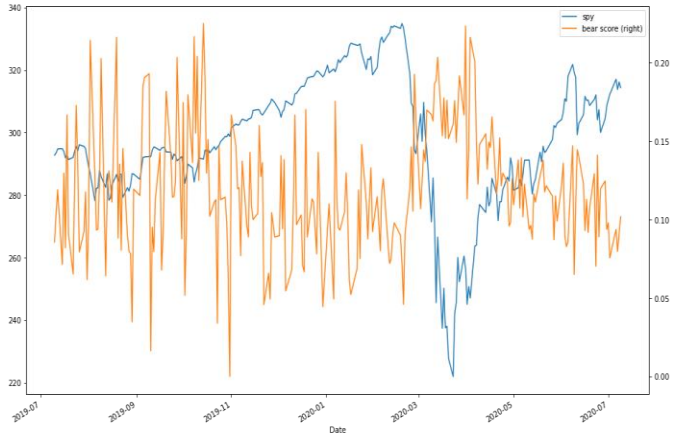

Fig. 4. Bearish Sentiment Analysis, where X-Axis denotes the date and Y-Axis denotes the stock price and the sentiment score.

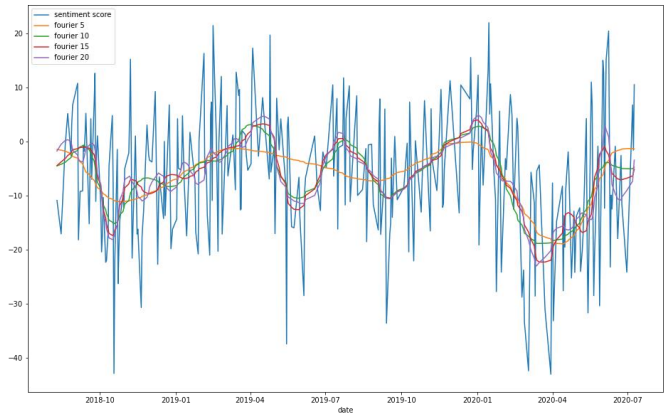

Fig. 5. Noise Removal at different clock/Fourier cycles for Positive and Negative Sentiment Analysis, where X-Axis denotes the date and Y-Axis denotes the sentiment score.

Fig. 5 depicts noise removal at various clock/Fourier cycles for Positive and Negative Sentiment Analysis, with the X-Axis representing the date and the Y-Axis representing the sentiment score.

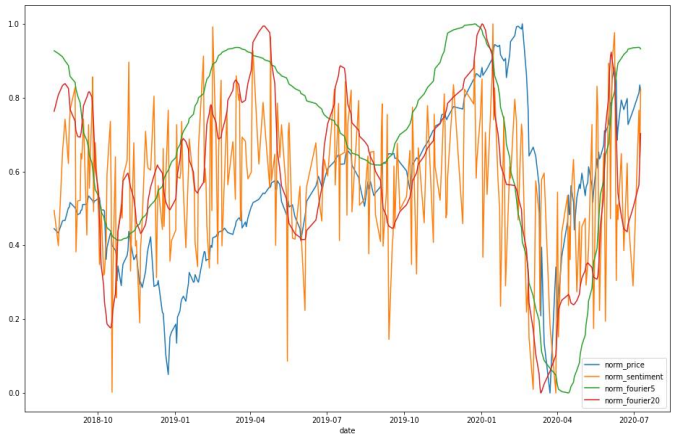

Fig. 6. Normalized Positive-Negative Sentiment Scores after Noise Removal, where X-Axis denotes the date and Y-Axis denotes the normalized sentiment score.

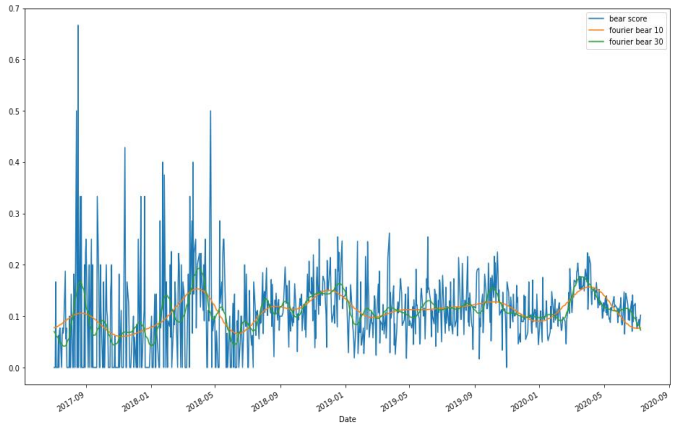

Fig. 7. Noise Removal at different clock/Fourier cycles for Bearish Sentiment Analysis, where X-Axis denotes the date and Y-Axis denotes the sentiment score.

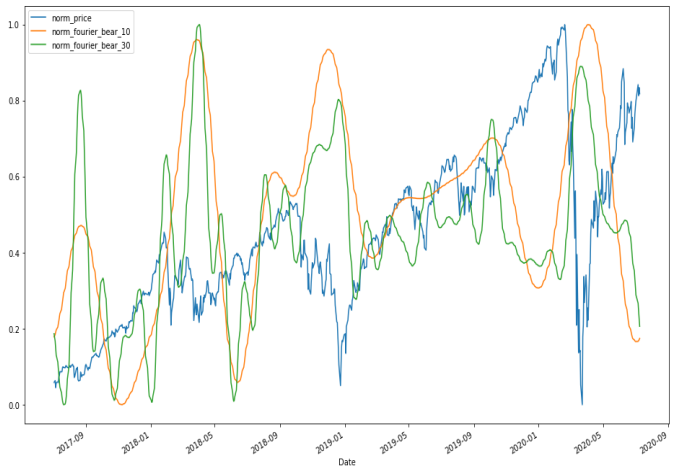

Fig. 8. Normalized Bearish Sentiment Scores after Noise Removal, where X-Axis denotes the date and Y-Axis denotes the normalized sentiment score.

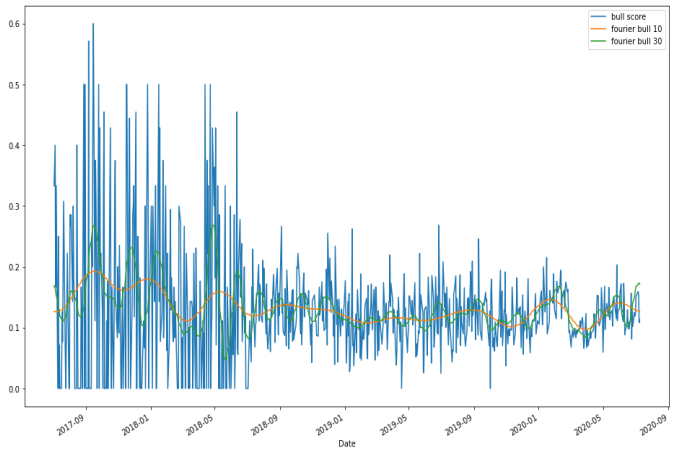

Fig. 9. Noise Removal at different clock/Fourier cycles for Bullish Sentiment Analysis, where X-Axis denotes the date and Y-Axis denotes the sentiment score.

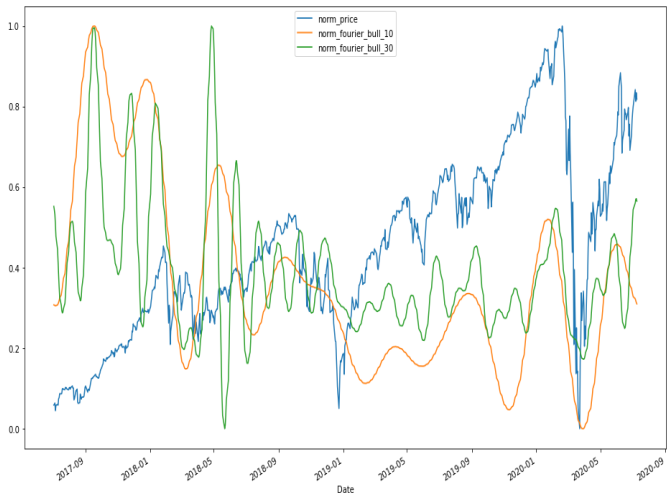

Fig. 10. Normalized Bullish Sentiment Scores after Noise Removal, where X-Axis denotes the date and Y-Axis denotes the normalized sentiment score.

This graphical representation is essential for comprehending the effect of Fourier cycles on sentiment data. Fig. 6 shows the Normalized Positive-Negative Sentiment Scores after noise removal, providing insights into the resulting data quality and preparedness for further analysis. Fig. 7 shows the noise removal procedure for Bearish Sentiment Analysis, with the X-Axis representing the date and the Y-Axis representing the sentiment score. Fig. 8 depicts the efficacy of the noise removal procedure by plotting the Normalised Bearish Sentiment Scores.

Fig. 9 shows the noise removal across multiple cycles for Bullish Sentiment Analysis, and Fig. 10 shows the Normalized Bullish Sentiment Scores following this process. These numbers show the efficacy of our noise removal method and its vital role in refining sentiment scores for future strategic development.

C. *Strategy Development*

The final phase leverages the cleaned and analyzed sentiment data for prediction and strategy development.

By using a combination of rolling correlation and Long-Short Term Memory (LSTM) Recurrent Neural Networks, we establish a model that dynamically adapts to new data. This approach not only helps in creating accurate predictive models but also allows for the development of robust trading strategies. The model checkpoint feature in LSTM assists in tracking previous states, offering valuable insights for future strategy adjustments and enhancing prediction accuracy.

Fig. 11 in our study is crucial because it depicts the dynamic relationship between sentiment data and stock prices using the notion of rolling correlation. The X-Axis in this graph depicts time, while the Y-Axis represents both the rolling correlation coefficient and the normalised stock price. The graph's red line depicts the mean-standard deviation coefficient, which effectively captures the fluctuation and strength of the correlation between market sentiment and stock prices across time. Meanwhile, the black line indicates the mean normalised stock price, which serves as a clear benchmark against which to assess sentiment swings.

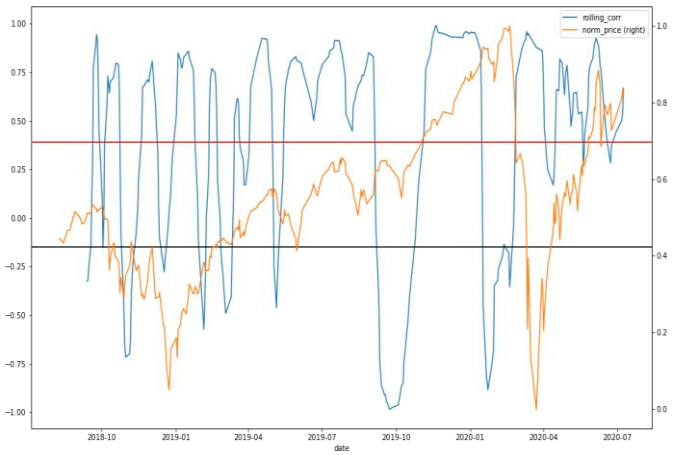

Fig. 11. Rolling Correlation between data and price where the red line depicts the mean-standard deviation coefficient and black line depicts the mean normalized stock price.

This visualisation is very useful since it shows how closely sentiment data trends correspond to actual stock price fluctuations, highlighting the predictive power of our sentiment analysis algorithm. The rolling correlation approach, as seen in this figure, allows us to study variations in correlation over time, providing a more detailed picture of how emotion drives stock prices under various market scenarios.

IV. Results

The findings of our study indicate the robustness and efficacy of the suggested sentiment analysis model in the formulation of stock trading strategies. Fig. 12 and 13 demonstrate the model's performance in terms of loss during training and validation, as well as its forecast accuracy in terms of stock price fluctuations.

1. Training and Validation Loss: Understanding the model's efficiency and dependability requires an examination of training and validation loss. Fig. 12 illustrating 'Effective Losses in Strategy Development' depicts the epochs on the X-Axis and the equivalent loss on the Y-Axis. We noticed a little loss throughout the training phase, indicating that the model efficiently understood the fundamental patterns in the sentiment data. The validation step, which is critical for validating the model's performance on unseen data, revealed a 4% loss. This minimal validation loss indicates that the model generalises successfully and is not overfit to the training data. The consistency of training and validation loss is an excellent predictor of model robustness.

2. Predictive Accuracy of Stock Prices: One of the most important components of our research is the model's ability to reliably anticipate stock price fluctuations. Fig. 13, headed 'Actual Stock Price vs Predicted Stock Price,' compares actual stock prices to our model's predictions over a period of days. The X-Axis depicts the number of days, whilst the Y-Axis depicts the change of stock values. The model's success is demonstrated by the tight alignment of anticipated and real stock prices, which has an accuracy of roughly 80%. This high degree of precision is essential because it illustrates the model's capacity to translate sentiment research into meaningful information for stock price prediction.

3. Implications for Strategy Development: Both statistics' results have far-reaching ramifications for stock trading technique development. The model's excellent accuracy in predicting stock price fluctuations, together with its ability to sustain low loss rates during training and validation, provides a good platform for building sentiment-based trading strategies. Traders can use these data to develop methods that are more in tune with market sentiments, perhaps resulting in more informed and lucrative trading decisions.

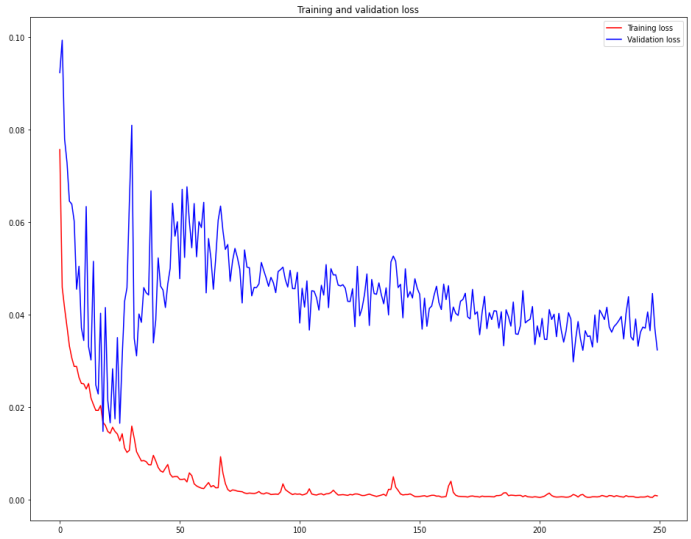

Fig. 12. Effective Losses in Strategy Development, where X-Axis shows the epochs and Y-Axis shows the loss

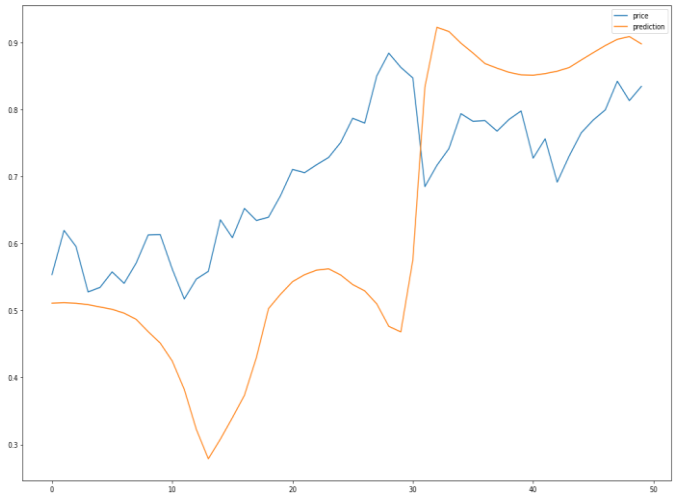

Fig. 13. Actual Stock Price vs Predicted Stock Price, where X-Axis shows the days count and Y-Axis shows the movement of stock price.

V. Conclusions and Future Work

In conclusion, this research has illuminated the significant potential of sentiment analysis, particularly utilizing Reddit trading data, in understanding and predicting stock market trends. Our study underscores the feasibility and value of leveraging online discourse to inform trading strategies, contributing to safer and more informed investment decisions. By integrating sophisticated sentiment analysis with advanced machine learning techniques, we've demonstrated a robust method for interpreting and capitalizing on market sentiment. For future research, there are several promising directions that other scholars and practitioners could pursue:

1. Algorithm Enhancement for Broader Data Sources: Future work could involve enhancing sentiment analysis algorithms to effectively process and interpret data from a broader range of online sources beyond Reddit. Expanding to platforms like Twitter, financial blogs, and news sites could provide a more comprehensive view of market sentiment.

2. Cross-Market Analysis: Researchers could explore sentiment analysis across different financial markets, including commodities, foreign exchange, and cryptocurrencies. This cross-market analysis could reveal unique sentiment dynamics and correlations, offering insights into global financial sentiment trends.

3. Real-Time Analysis and API Integration: Developing an API for real-time sentiment analysis could be a significant advancement. This would allow traders and institutions to integrate sentiment analysis into their existing trading platforms, providing real-time insights for faster and more responsive decision-making.

4. Long-Term Predictive Models: Future studies could focus on long-term predictive models, analysing how sentiment data correlates with long-term stock market trends. This would be beneficial for long-term investors and portfolio managers looking to incorporate sentiment analysis into their investment strategies.

5. Comparative Studies of Sentiment Analysis Techniques: Comparing the effectiveness of various sentiment analysis techniques, like machine learning, deep learning, and natural language processing, in the context of stock market prediction could provide valuable insights into the most effective methods for financial sentiment analysis.

6. Impact of Global Events on Market Sentiment: Investigating the impact of significant global events, like political changes or economic policies, on market sentiment could offer valuable insights into how external factors influence market dynamics.

By exploring these avenues, future research can build upon the foundation laid by this study, contributing further to the field of sentiment analysis in finance and its application in strategy development for stock trading.

References

[1] C. R. Ko and H. T. Chang, "LSTM-based sentiment analysis for stock price forecast," Peer J Computer Science, vol. 7, e408, 2021. doi: 10.7717/peerj-cs.408.

[2] M. Makrehchi, S. Shah, and W. Liao, "Stock prediction using event-based sentiment analysis," in Proc. 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Vol. 1, 2013, pp. 337-342. doi: 10.1109/WI-IAT.2013.48.

[3] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," in Proc. 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 2016, pp. 1345-1350. doi: 10.1109/SCOPES.2016.7955659.

[4] F. Valencia, A. Gómez-Espinosa, and B. Valdés-Aguirre, "Price movement prediction of cryptocurrencies using sentiment analysis and machine learning," Entropy, vol. 21, no. 6, p. 589, 2019. doi: 10.3390/e21060589.

[5] S. Kazemian, S. Zhao, and G. Penn, "Evaluating sentiment analysis in the context of securities trading," in Proc. 54th Annual Meeting of the

Association for Computational Linguistics, Volume 1: Long Papers, 2016, pp. 2094-2103. doi: 10.18653/v1/P16-1197.

[6] W. Zhang and S. Skiena, "Trading strategies to exploit blog and news sentiment," in Proc. Fourth International AAAI Conference on Weblogs and Social Media, 2010. doi: 10.1609/icwsm.v4i1.14075.

[7] K. Ahmad, "Multi-lingual sentiment analysis of financial news streams," PoS, vol. 001, 2006. doi: 10.22323/1.026.0001.

[8] Y. Guo, "Stock Price Prediction Based on LSTM Neural Network: the Effectiveness of News Sentiment Analysis," in Proc. 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME), 2020, pp. 1018-1024. doi: 10.1109/ICEMME51517.2020.00206.

[9] T. Sidogi, R. Mbuvha, and T. Marwala, "Stock Price Prediction Using Sentiment Analysis," in Proc. 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2021, pp. 46-51. doi: 10.1109/SMC52423.2021.9659283.

[10] E. Naresh, B. J. Ananda, K. S. Keerthi, and M. R. Tejonidhi, "Predicting the Stock Price Using Natural Language Processing and Random Forest Regressor," in Proc. 2022 IEEE International Conference on Data Science and Information System (ICDSIS), 2022, pp. 1-5. doi: 10.1109/ICDSIS55133.2022.9915940.

[11] Rekha G., Bhanu Sravanthi D., Ramasubbareddy S., Govinda K., "Prediction of stock market using neural network strategies, Journal of Computational and Theoretical Nanoscience," Vol:16, Issue: 43621, Pg.No:2333-2336. doi: 10.1166/jctn.2019.7895.

[12] Q. Wang, "Predicting Chinese Stock Market with Internet Key Word Hotness by Statistical Time Series Regression Analysis," in Proc. 2021 International Conference on Computer, Blockchain and Financial Development (CBFD), 2021, pp. 286-291. doi: 10.1109/CBFD52659.2021.00064.

[13] S. S. Abdullah, M. S. Rahaman, and M. S. Rahman, "Analysis of stock market using text mining and natural language processing," in Proc. 2013 International Conference on Informatics, Electronics and Vision (ICIEV), 2013, pp. 1-6. doi: 10.1109/ICIEV.2013.6572673.

[14] Lakshya, Prateek, and D. Sethia, "Stock Price Prediction Using News Sentiment Analysis," in Proc. 2022 2nd International Conference on Intelligent Technologies (CONIT), 2022, pp. 1-6. doi: 10.1109/CONIT55038.2022.9847747.

[15] A. Pandey, S. Shukla, and K. Mohbey, "Comparative Analysis of a Deep Learning Approach with Various Classification Techniques for Credit Score Computation," Recent Advances in Computer Science and Communications, 2020. doi: 10.2174/2666255813999200721004720.