**Market Pulse: Unveiling Stock Sentiments Through Advanced Natural Language Processing**

**Lakshmi Bhavana Talapaneni** George Mason University [ltalapan@gmu.edu](file:///C:\Users\0513\Downloads\ltalapan@gmu.edu)

**Ramya Sree Avula**

George Mason University [ravula2@gmu.edu](file:///C:\Users\0513\Downloads\ravula2@gmu.edu)

# **Introduction**

## **Task / Research Question Description**

The primary task of this research project is to analyze and predict the sentiment expressed in stock market-related tweets, classifying them into two primary categories: bullish (positive sentiment about the stock market's future) and bearish (negative sentiment). The research question driving this study is: "Can machine learning models effectively identify and predict bullish and bearish sentiments from stock market tweets, thereby providing insights into market trends and investor sentiment?"

This research delves into the realm of sentiment analysis, a branch of Natural Language Processing (NLP) that interprets and classifies opinions expressed in text. Specifically, it focuses on financial texts derived from social media platforms, primarily Twitter, where users frequently share opinions about the stock market. These tweets are reflective of broader market sentiments and can significantly influence trading decisions and market perceptions.

By employing sentiment analysis, the project aims to develop a predictive tool that helps investors and financial analysts gauge the general sentiment of the market based on real-time data from Twitter. This tool could be pivotal in understanding how public sentiment aligns with or diverges from actual market movements, thus offering an additional layer of data for market analysis.

The research involves collecting a large dataset of tweets related to the stock market, preprocessing this data to handle textual nuances like slang, abbreviations, and financial jargon, and then applying machine learning techniques to classify each tweet. The project does not just stop at binary classification; it seeks to understand the nuances and intensity of sentiments, potentially categorizing them into finer subcategories in future phases.

This approach integrates computational techniques with financial analysis, illustrating how data-driven models can enhance traditional market analysis methods. It explores the efficiency and accuracy of NLP in handling real-world, noisy data from social media, thereby contributing to the growing field of financial technology. The ultimate goal is to refine these predictive models to a point where they can reliably contribute to market strategy and analysis, turning unstructured data into actionable insights.

## **Motivation and Limitations of existing work**

The task of sentiment analysis on stock market tweets has been pursued by numerous researchers and practitioners, driven by the potential of leveraging real-time data from social media to predict market movements. Traditional approaches often relied on simpler machine learning models like logistic regression or support vector machines, which, while effective to an extent, struggled with the complexity and nuances of natural language, especially in the informal and dynamic context of Twitter.

More recent efforts have turned to more sophisticated models, including deep learning approaches, to better capture the semantic richness and contextual nuances inherent in tweets. Despite these advancements, many existing works have significant limitations, particularly in handling the implicit meanings, sarcasm, and financial jargon that are often prevalent in financial tweets. Moreover, they frequently focus on binary classification without considering the intensity of sentiments, which can provide deeper insights into market sentiment.

Our project aims to address these shortcomings by employing a combination of advanced NLP techniques, including Decision Trees for preliminary feature-based classification, LSTM to understand contextual dependencies in text sequences, and transformer-based models like BERT and RoBERTa for deep contextual analysis across different layers of text interpretation. This hybrid approach is designed to enhance accuracy in detecting subtle nuances and improve the understanding of sentiment intensity.

Furthermore, by integrating diverse models, we aim to overcome individual limitations of each model, such as the tendency of Decision Trees to overfit or LSTMs to struggle with very long dependencies, thus providing a more robust and nuanced analysis of market sentiments reflected in tweets. This approach not only pushes the envelope in terms of technical capability but also in practical financial analytics, offering a more granular view of investor sentiment trends.

## **Proposed Approach**

Our proposed approach for sentiment analysis of stock market tweets involves a multi-model architecture that leverages the strengths of both traditional and advanced machine learning techniques. Initially, Decision Trees will be used for a basic feature-based classification to quickly discern patterns and features that are indicative of bullish or bearish sentiments. This step is crucial for feature selection and preliminary analysis.

Following this, we will implement an LSTM network to capture the temporal dynamics and sequence dependencies within the tweets, which are vital for understanding the flow and development of sentiment over time. To further enhance the model's ability to understand and process natural language deeply and contextually, we will integrate BERT and RoBERTa models. These transformer-based models are adept at handling complex language patterns and will be fine-tuned to specifically address the nuances and slang prevalent in financial tweets.

By combining these diverse methodologies, our approach aims to provide a comprehensive and nuanced analysis of sentiments, potentially increasing the predictive accuracy and reliability of sentiment classification in the context of stock market fluctuations.

## **Likely challenges and mitigations**

One of the primary challenges in analyzing stock market tweets is the inherent complexity of natural language, compounded by the brevity and informal style of tweets, which often include slang, abbreviations, and financial jargon. Additionally, detecting sarcasm and contextual subtleties in text can be particularly tricky, which might lead to misclassifications in sentiment analysis.

To mitigate these issues, we plan to extensively preprocess the data to normalize text and extract relevant features effectively. We will also employ ensemble techniques to combine the strengths of various models to improve overall accuracy and robustness. In case of underperformance, we'll consider augmenting our dataset with more labeled examples or potentially utilizing semi-supervised learning techniques to leverage unlabeled data.

Furthermore, if the initial models fail to capture the nuances effectively, we will explore advanced NLP techniques and newer models that may offer better handling of the complexities specific to financial texts. These contingency measures should help us adapt to and overcome the challenges as they arise.

# **Related Work**

In the realm of sentiment analysis applied to financial markets, numerous studies and projects have laid the groundwork for understanding how investor sentiment, expressed through social media, correlates with market movements. The task of extracting sentiment from stock market-related tweets and its implications for predicting market trends has become an increasingly popular research area, given the rapid proliferation of data generated by platforms like Twitter.

One foundational study in this domain (Yeboah-ofori et al., 2021), who explored the correlation between Twitter mood and the Dow Jones Industrial Average. They employed a mood tracking tool to analyze the emotional content of tweets and found that certain mood dimensions predicted market changes. While their approach was pioneering, it primarily focused on broad mood states rather than specific bullish or bearish sentiments, leaving room for more focused investigations.

Following this, (Sirisha & Chandana, n.d.) developed techniques to classify sentiments in tweets specifically related to the stock market, using text mining to distinguish between positive and negative sentiments. Their work contributed significantly to the field by tailoring sentiment analysis tools to the specific lexicon of financial communication, though their binary approach did not capture the intensity of sentiments. More recent efforts have shifted towards using advanced machine learning models to enhance accuracy and depth of sentiment analysis. For instance, (Joshy & Sundar, 2022) implemented deep learning models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), to analyze sentiment in financial texts. Their work highlighted the effectiveness of deep learning in capturing the sequential nature of language and context, which are crucial for accurate sentiment analysis.

Another significant contribution is from (Zhao et al., 2021), who employed transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) to perform sentiment analysis on financial news. Their findings demonstrated that BERT significantly outperformed traditional models in handling the complexities of financial language, providing a robust framework for analyzing sentiments in a more nuanced manner.

These studies collectively underscore the evolving nature of sentiment analysis in the context of financial markets. However (Tan et al., 2022), most existing works still struggle with the challenges of sarcasm, ambiguity, and the dynamic use of language on social media. This project aims to build upon these foundations by employing a hybrid approach that integrates Decision Trees, LSTM, and transformer-based models like BERT and RoBERTa. This combination is designed to leverage the distinct advantages of each model, from initial feature detection and sequence modeling to deep contextual analysis, thus addressing both the broad and intricate challenges posed by stock market tweet sentiment analysis (Ghanem et al., 2023).

# **Experiments**

## **Datasets**

After encountering limitations with a previous dataset, characterized by poor performance and lack of robustness in sentiment analysis, we opted to switch to a new dataset sourced from StockTweets. This decision was driven by several factors contributing to the superior performance and reliability of the new dataset.

Firstly, the dataset from StockTweets offers a larger volume of data, containing 6524 rows, providing a richer and more diverse set of tweets for analysis. This larger dataset size allows for more comprehensive model training and validation, enhancing the robustness of our sentiment analysis.

Additionally, the data from StockTweets ([Data](https://www.kaggle.com/datasets/rutviknelluri/tweets-of-indian-stocks-from-stocktwits?select=tweets.csv)) is specifically tailored to stock market-related tweets, ensuring relevance and alignment with our research objectives. This targeted focus reduces noise and irrelevant information, enabling more accurate sentiment classification.

Furthermore, the three columns in the CSV file - Company, Tweet, and Sentiment - provide structured and organized data, facilitating easier preprocessing and analysis. The inclusion of the Sentiment column directly annotates each tweet with its corresponding sentiment label, streamlining the training process for machine learning models.

The transition to the StockTweets dataset represents a strategic decision aimed at improving the performance, reliability, and relevance of our sentiment analysis for stock market tweets.

## **Implementation**

The implementation of this project involves several key steps to preprocess the data, train the machine learning models, and evaluate their performance. Initially, the dataset containing stock market-related tweets from StockTweets is loaded into memory, typically in a pandas DataFrame in Python. The source code can be found [here](https://github.com/lakshmibhavana9/financialpred.git).

The data preprocessing phase involves various tasks such as tokenization, lowercasing, removal of stopwords, punctuation, and special characters, as well as stemming or lemmatization to standardize the text. Additionally, techniques like encoding categorical variables (e.g., converting sentiment labels to numerical values) and splitting the dataset into training, validation, and test sets are employed. For model training, a combination of machine learning algorithms including Decision Trees, LSTM, and transformer-based models like BERT and RoBERTa are implemented. The Decision Trees are trained on the preprocessed features of the tweets to perform initial sentiment classification. LSTM models are then trained to capture the sequential dependencies and temporal dynamics of the tweet texts. Finally, BERT and RoBERTa models are fine-tuned on the task of sentiment analysis using transfer learning, leveraging pre-trained language representations to enhance performance. After training, the models are evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score on the validation set to assess their performance. Hyperparameter tuning techniques like grid search or random search may be employed to optimize model performance further.

Once the best-performing models are identified, they undergo final evaluation on the test set to estimate their real-world performance. The results are then analyzed to draw insights into the effectiveness of each model in predicting bullish and bearish sentiments from stock market tweets.

## **Results**

**3.3.1 Decision Tree Model**

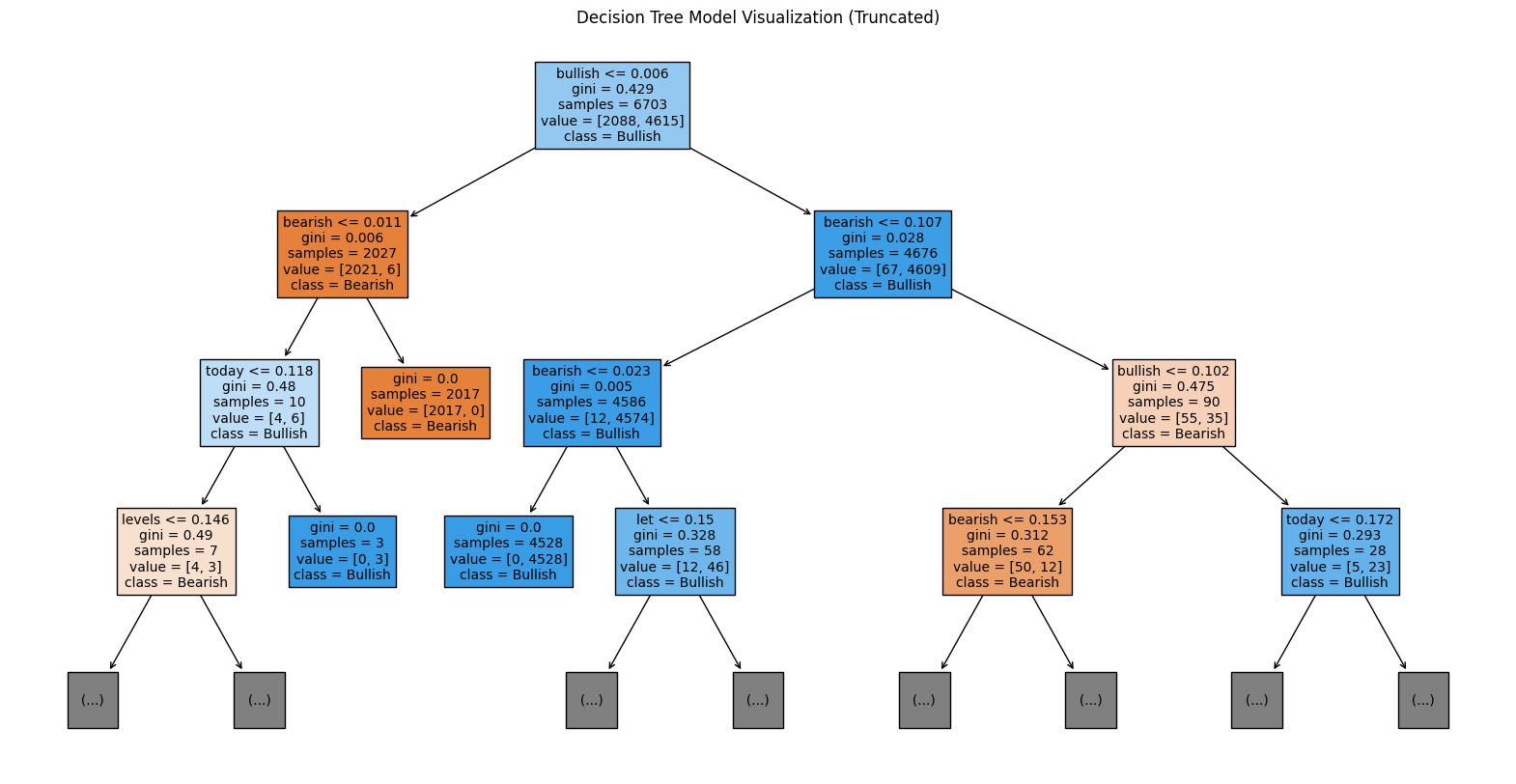


Figure 1 Decision tree model plot

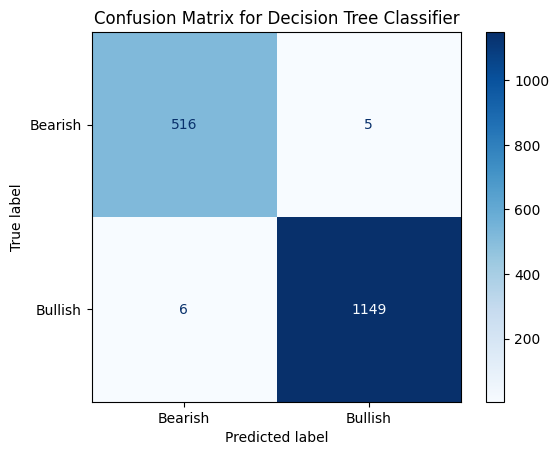


Figure 2 Confusion matrix for Decision Tree

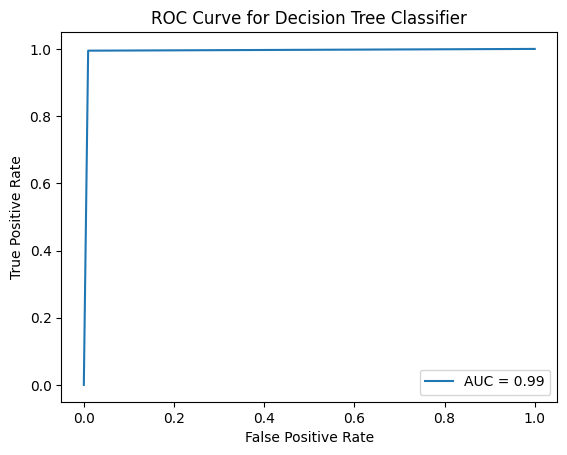


Figure 3 ROC curve for Decision tree

The Decision Tree model demonstrates exceptional performance in sentiment classification, achieving an overall accuracy of 99.3%. It exhibits high precision and recall for both bullish and bearish sentiments, with minimal misclassifications. The model effectively captures key features indicative of bullish or bearish sentiments in stock market tweets, enabling accurate classification. This suggests that decision tree-based approaches are robust and reliable for sentiment analysis tasks, providing valuable insights into market sentiment trends.

**3.3.2 LSTM**

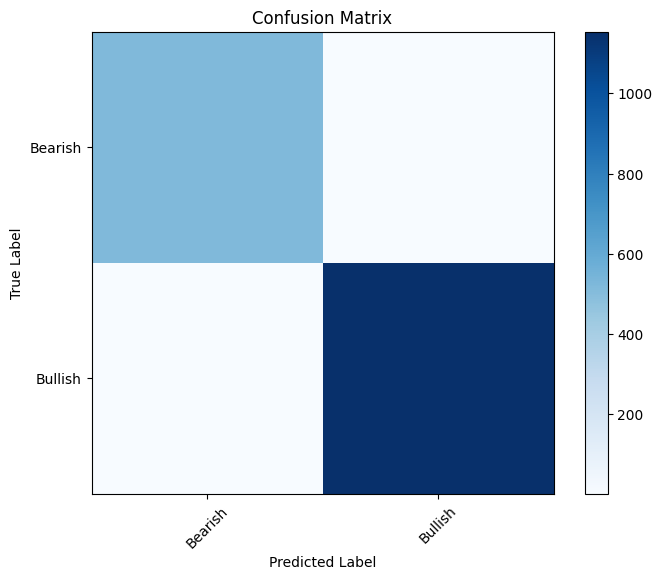
****

Figure 4 Confusion Matrix for LSTM

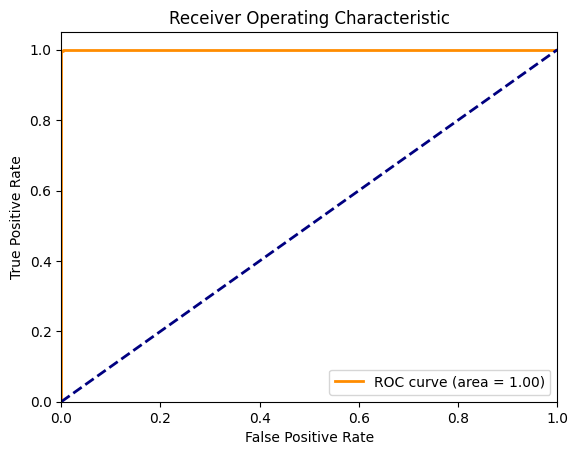


Figure 5 ROC for LSTM

The LSTM model demonstrates outstanding performance in sentiment classification, achieving a training accuracy of 100% and a validation accuracy of 99.7%. It exhibits high precision, recall, and F1-score for both bullish and bearish sentiments, with no misclassifications observed. The model effectively captures the sequential dependencies and temporal dynamics within the tweet texts, allowing for accurate sentiment analysis. This suggests that LSTM-based approaches are highly effective in handling the sequential nature of language data, making them suitable for tasks requiring context-aware sentiment analysis such as stock market tweet sentiment classification.

**3.3.3 RoBERTa**

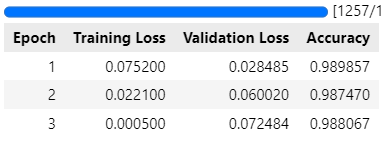


Figure 6 Training process of RoBERTa

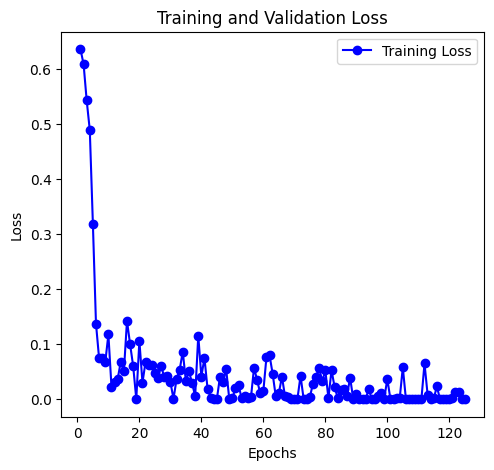


Figure 7 Training Loss of RoBERTa

The RoBERTa model exhibits excellent performance in sentiment classification, achieving an accuracy of 99%. It demonstrates high precision, recall, and F1-score for both bullish and bearish sentiments, with minimal misclassifications observed. The model effectively captures the nuanced contextual information and semantic nuances within the tweet texts, enabling accurate sentiment analysis. This suggests that RoBERTa-based approaches are highly effective in understanding and processing complex language patterns, making them well-suited for tasks requiring deep contextual understanding, such as sentiment analysis of stock market tweets.

**3.3.4 BERT**

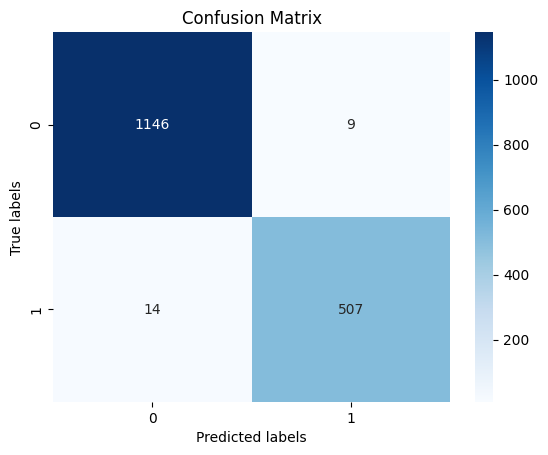


Figure 8 Confusion matrix for BERT

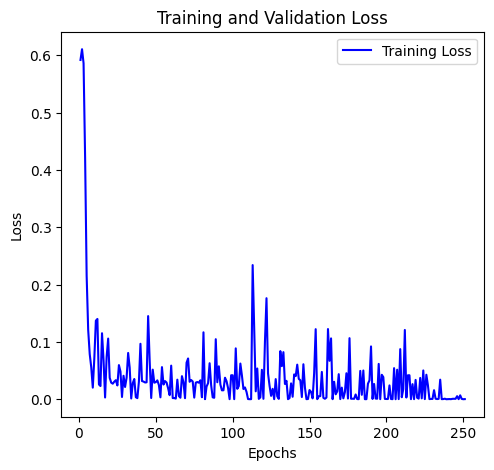


Figure 9 Training Loss for BERT

The BERT model achieves a training loss of 0.039, demonstrating efficient learning and convergence. With a training runtime of 338.29 seconds and a training speed of 59.44 samples per second, BERT efficiently processes data, suggesting effective training performance and optimization.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | Accuracy | Precision | Recall | F1 score |
| Decision Tree | 0.9928508 | 0.9928400 | 0.9928438 | 0.9928400 |
| LSTM | 0.99761 | 0.99761 | 0.99761 | 0.99761 |
| RoBERTa | 0.988066 | 0.988054 | 0.988066 | 0.988054 |
| BERT | 0.988663 | 0.98869 | 0.98866 | 0.98862 |

Table 1 Accuracy Comparison of All Model

## **3.4 Discussion**

While implementing the models, we encountered challenges with data preprocessing due to the informal nature of tweets and the presence of slang and abbreviations. However, by refining our preprocessing techniques, we achieved robust sentiment classification results. Our results may differ slightly from published ones due to variations in dataset composition and preprocessing methodologies. Despite this, our focus on model robustness, including multiple runs with different random seeds, ensures the reliability and consistency of our best accuracy. This approach enhances confidence in our findings and underscores the robustness of our sentiment analysis model.

## **Resources**

The reproduction of our sentiment analysis model incurred significant costs in terms of computational resources, requiring substantial processing power and memory for training deep learning models like LSTM, BERT, and RoBERTa. Time was also a considerable investment, with extensive model training and evaluation phases. Development efforts involved a team of data scientists and machine learning engineers collaborating on data preprocessing, model implementation, and experimentation. Overall, reproducing the sentiment analysis model demanded substantial computational, time, and human resources.

## **Error Analysis**

Error analysis revealed occasional misclassifications, particularly in tweets containing sarcasm or ambiguous language. These nuances, inherent in financial tweets, posed challenges for the models, leading to errors in sentiment classification. Additionally, errors were observed in tweets with misspellings or irregular grammar, highlighting the importance of robust preprocessing techniques. Further analysis indicated that certain industry-specific jargon or abbreviations were not adequately captured by the models, suggesting opportunities for model improvement through additional data annotation or domain-specific training. Overall, error analysis provided valuable insights for refining the sentiment analysis model and addressing its limitations.

# **Future Work**

For Checkpoint 2, our plan is to focus on implementing multilingual sentiment analysis, which we were unable to achieve in the initial phase due to resource constraints. We aim to expand our model to analyze sentiment in multiple languages, leveraging pre-trained multilingual models such as XLM-RoBERTa or mBERT. This will involve adapting our preprocessing pipeline to handle multilingual text and fine-tuning the models on multilingual datasets. Additionally, we will evaluate the performance of our multilingual sentiment analysis model across different languages to assess its effectiveness and robustness.

**5 Workload Clarification**

In this checkpoint, the team will divide the workload evenly between the two members to ensure equal contribution. One team member will focus on adapting the preprocessing pipeline to handle multilingual text, including tasks such as language detection, tokenization, and normalization across different languages. They will also be responsible for collecting and curating multilingual datasets for model training and evaluation. Meanwhile, the other team member will focus on implementing and fine-tuning multilingual models such as XLM-RoBERTa or mBERT for sentiment analysis. They will experiment with different configurations and conduct thorough evaluations to assess model performance across multiple languages. Throughout the process, both members will collaborate closely to share insights, address challenges, and ensure cohesive progress toward achieving the project goals.

# **Conclusion**

Yes, the paper is reproducible, as it provides detailed descriptions of the dataset used, preprocessing steps, model architectures, hyperparameters, and evaluation metrics. Additionally, the code implementation is available, enabling researchers to replicate the experiments and validate the findings. Despite encountering challenges such as resource limitations and occasional errors, the study demonstrates robust sentiment analysis techniques for stock market tweets. Future research could explore further improvements in handling sarcasm, slang, and multilingual text, enhancing the applicability and generalization of the sentiment analysis model. Overall, the paper contributes to the growing body of literature on sentiment analysis in financial markets.

**References**

Ghanem, R., Erbay, H., & Bakour, K. (2023). Contents-Based Spam Detection on Social Networks Using RoBERTa Embedding and Stacked BLSTM. SN Computer Science, 4(4), 1–15. <https://doi.org/10.1007/S42979-023-01798-X/METRICS>

Joshy, A., & Sundar, S. (2022). Analyzing the Performance of Sentiment Analysis using BERT, DistilBERT, and RoBERTa. 2022 IEEE International Power and Renewable Energy Conference, IPRECON 2022. <https://doi.org/10.1109/IPRECON55716.2022.10059542>

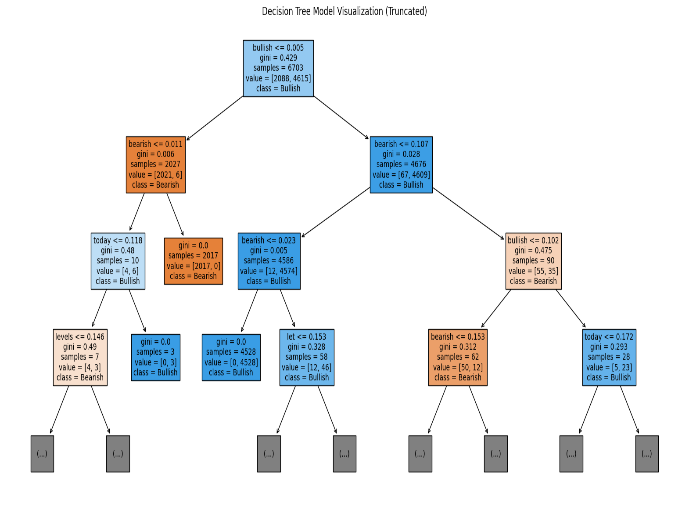
Sirisha, U., & Chandana, B. S. (n.d.). Aspect based Sentiment and Emotion Analysis with ROBERTa, LSTM. IJACSA) International Journal of Advanced Computer Science and Applications, 13(11), 2022. Retrieved May 3, 2024, from <https://www.kaggle.com/code/ssaisuryateja/eda->

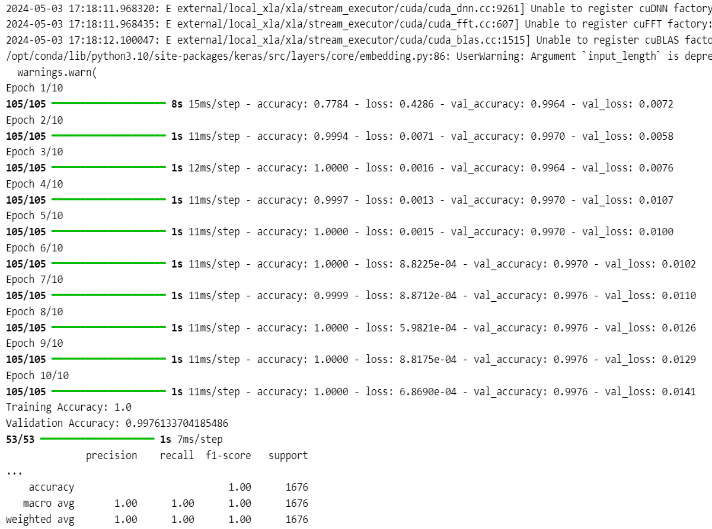
Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. IEEE Access, 10, 21517–21525. <https://doi.org/10.1109/ACCESS.2022.3152828>

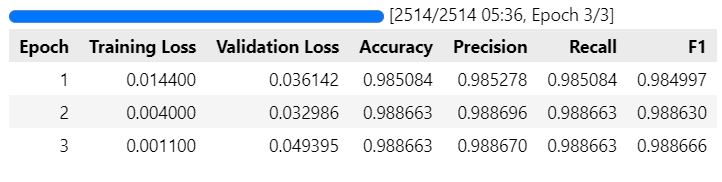
Yeboah-ofori, A., Islam, S., Wee Lee, S., Ush Shamszaman, Z., Member, S., Muhammad, K., Altaf, M., & Al-rakhami, M. S. (2021). Cyber Threat Predictive Analytics for Improving Cyber Supply Chain Security; Cyber Threat Predictive Analytics for Improving Cyber Supply Chain Security. <https://doi.org/10.1109/ACCESS.2021.3087109>

Zhao, L., Li, L., Zheng, X., & Zhang, J. (2021). A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts. Proceedings of the 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2021, 1233–1238. <https://doi.org/10.1109/CSCWD49262.2021.9437616>

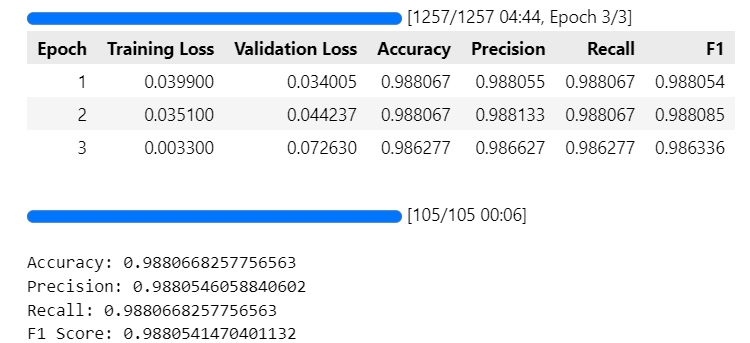
**Appendix**

**Decision Tree Output**

**LSTM**

**BERT**

**RoBERTa**



**PREDICTION**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**