

Authors:

Hassan Mehmood
Shane Thomas
Gavin
Gabriel Monzato

Date of Submission:

February 7, 2024

Enhancing Stock Market Predictions through Sentiment Analysis: The Portfolio Optimizer Project

Contents

Authors:.....	0
Date of Submission:	0
Abstract:	2
Introduction	2
Industry Overview.....	2
Reasons Why It's Important to the Industry	3
Project Objectives	4
Data Collection	4
Methodology:.....	4
Exploratory Data Analysis:	4
Preprocessing:.....	5
Custom Model (RNN):	6
Evaluating the performance of direct inferencing Distil-Roberta:	7
Finetuning and Inferencing Fin Bert:.....	7
Comparative Analysis:	7
Model Selection	8
Performance Optimization	8
Model Performance with rebalanced dataset:	9
Current And Future Work:	9
Discussion.....	10
Concluding Remarks.....	10

Abstract:

In today's dynamic financial markets, accurate prediction of stock movements is paramount for investors and analysts. Traditional methods, relying on quantitative data analysis alone, often overlook the impact of changing market sentiments. This report presents the "Portfolio Optimizer" project, which leverages sentiment analysis and deep learning techniques to enhance stock market predictions. By analyzing qualitative sentiment data from diverse sources such as news articles, financial reports, and social media, alongside quantitative data, the project aims to provide a comprehensive view of market dynamics. Through advanced sentiment analysis model development and integration with quantitative data, the project demonstrates significant potential for improving prediction accuracy, enabling real-time decision-making, enhancing risk management strategies, and fostering innovative investment approaches. The findings underscore the importance of integrating AI in financial analysis and signal promising directions for future research and application in investment decision-making processes.

Introduction

In this project, we implemented a comprehensive methodology designed to achieve robust performance in sentiment analysis tasks. Our approach involved constructing a Recurrent Neural Network (RNN) model from the ground up, serving as a foundational benchmark for evaluating subsequent pretrained and fine-tuned models. Through rigorous experimentation and evaluation, we systematically compared the performance of these models and selected the most effective one for deployment in inference tasks.

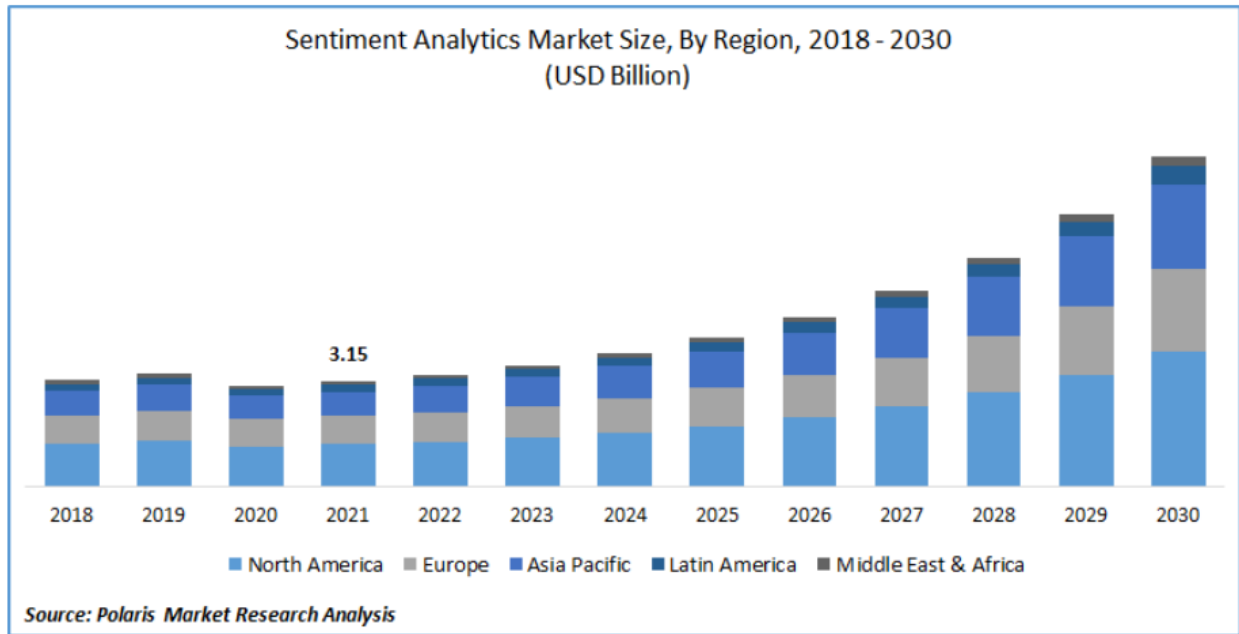
Industry Overview

The financial industry has always sought to predict market movements with a high degree of accuracy. Traditional methods have included fundamental analysis, examining company financials and market conditions, and technical analysis, focusing on patterns in trading activity and price movements. However, these methods often fail to account for the rapid changes in market sentiment that can significantly impact stock prices.

The advent of artificial intelligence (AI) and machine learning (ML) technologies has introduced new possibilities for market prediction. The global sentiment analytics market was valued at USD 3.15 billion in 2021 and is expected to grow at a CAGR of 14.4 %.

Following are some of the famous hedge funds who have incorporated sentiment analysis as part of their research:

- Two Sigma
- DE Shaw
- Renaissance Technologies



Now, let's articulate the reasons why the "Portfolio Optimizer" project is important to the industry:

Reasons Why It's Important to the Industry

The project holds significant implications for the financial industry for several key reasons:

1. **Enhanced Prediction Accuracy:** By integrating sentiment analysis with traditional financial metrics, the project aims to provide a more accurate prediction of stock market movements. This holistic approach can uncover insights that quantitative data alone may miss, such as the impact of geopolitical events or emerging market trends reflected in news sentiment.
2. **Real-time Decision Making:** The ability to analyze and interpret data in real-time allows investors and fund managers to make quicker, more informed decisions. In the fast-paced world of stock trading, where market conditions can change rapidly, the speed of decision-making can be as crucial as the accuracy of the predictions.
3. **Risk Management:** Improved prediction accuracy and timely decision-making contribute to more effective risk management strategies. By understanding potential market movements better, investors can adjust their portfolios to mitigate losses during downturns and capitalize on opportunities as they arise.
4. **Innovative Investment Strategies:** The project paves the way for developing new investment strategies that leverage AI and ML. These strategies can offer a competitive edge to investment firms and financial advisors, providing them with tools to navigate the complexities of global financial markets more effectively.
5. **Democratization of Financial Analysis:** By automating and enhancing the analysis process, technologies like those developed in the "Portfolio Optimizer" project can make sophisticated financial analysis more accessible to a broader audience. This democratization can empower smaller investors and firms, leveling the playing field in financial markets.

Project Objectives

The central aim of the sentiment analysis is to harness the power of public sentiment in understanding and predicting stock market movements. The project is structured around several key objectives, emphasizing the role of sentiment analysis as a foundational element of the model before its application in stock market predictions:

1. **Advanced Sentiment Analysis:** Develop a sophisticated sentiment analysis model that can accurately interpret and quantify the sentiments expressed in financial news, reports, and social media. This model aims to capture the nuanced perspectives and emotions of market participants.
2. **Integration with Quantitative Data:** Once the sentiment analysis model is established, the next objective is to integrate its outputs with traditional quantitative financial data. This integration seeks to enrich stock market predictions with a layer of qualitative insight, offering a more rounded view of potential market movements.
3. **Real-time Sentiment Analysis:** Implement real-time processing capabilities to analyze market sentiments as they evolve. This capability is crucial for capturing the immediate impact of news and social media on market dynamics.
4. **Risk Management Through Sentiment Indicators:** Utilize sentiment analysis as a tool for risk management, identifying potential market shifts and volatility driven by investor sentiment. This approach aims to offer investors foresight into market trends, aiding in the proactive management of investment risks.
5. **Enhanced Stock Market Predictions:** Ultimately, the project aims to leverage sentiment analysis, in conjunction with quantitative data, to significantly improve the accuracy of stock market predictions. This objective encompasses the development of predictive models that can inform investors about optimal buy or sell times, based on a comprehensive analysis that includes sentiment indicators.

Data Collection

For the "Portfolio Optimizer" project, a comprehensive and multifaceted approach was taken to collect data, focusing on capturing a wide range of sentiments that influence the stock market. The data collection process involved gathering information from several key sources:

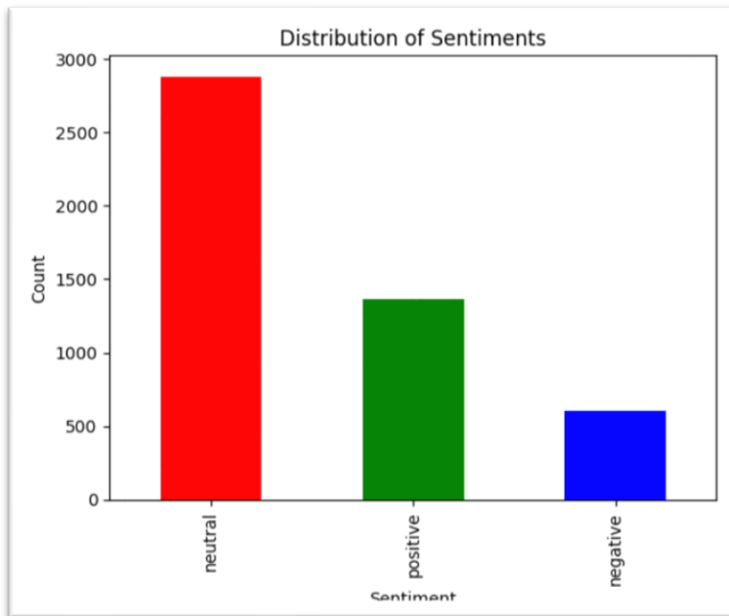
- **Financial Modelling and Prep API (FMP):** This source of data is used for real time market sentiment analysis about a specific security or a group of security. The Model can also inference economic sentiment analysis, generating a sentiment score on news such as jobs added, consumer pricing index and many more.
- **Kaggle:** For training a benchmark model and for comparative analysis we retrieved data from popular websites 'Kaggle'.

Methodology:

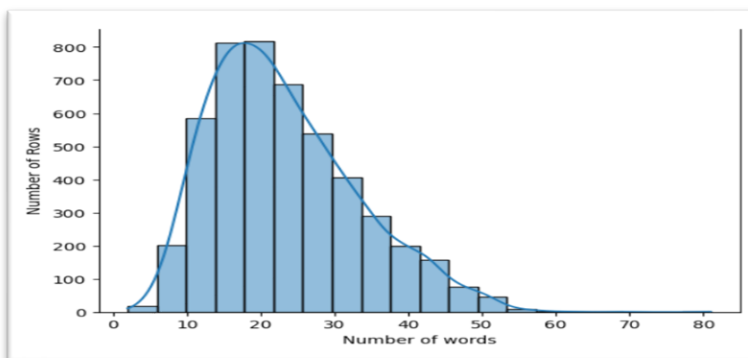
Exploratory Data Analysis:

Following description and stats are for the description of the data we used:

- Number of labels: 3 (Positive, Neutral and Negative).
- Total number of observations: 4846
- Number of columns: 1, including the text to be analyzed.
- Data Distribution: The Following images helped us get a better understanding of the data:
 - The following image illustrates the distribution of labels:



- This second image gives us an idea about the length of sentences, which is a very useful piece of information to have when setting up the model. The longest sentences comprised of 81 tokens while in average the sentences were composed of 23.25 tokens.
- From the image the data looks slightly skewed to the right.



Preprocessing:

Following preprocessing steps were carried out for the custom model:

- Encoding the labels using Label Encoder.
- Used the 'Tokenizer' from Keras to do the text preprocessing required for NLP tasks such as:
- **Normalization:** This step involved standardizing the text data by converting it to lowercase and removing any remaining non-alphanumeric characters. Normalization helps reduce the

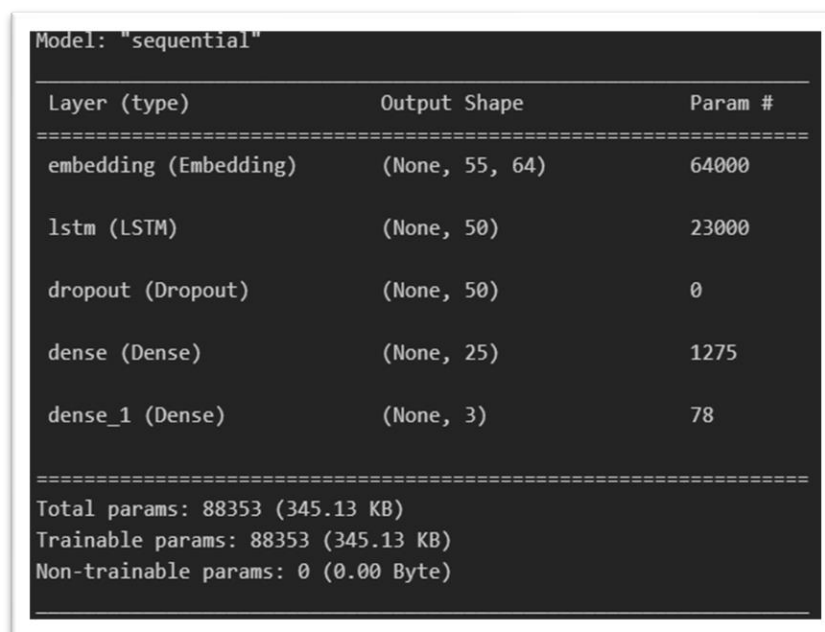
complexity of the text data, improving the model's ability to recognize and classify sentiments accurately.

- **Tokenization:** The cleaned text data was then tokenized, breaking down the content into individual words or phrases. Tokenization is essential for models like Fin Bert and Roberta to analyze the data effectively, as it converts text into a format that the models can process.
- **Vectorization:** To feed the text data into the sentiment analysis models, it was vectorized, transforming the tokenized text into numerical values. Vectorization allows the models to interpret the text data, utilizing techniques such as embedding to capture the semantic meaning of words and phrases.
- Splitting the data into training and testing data sets. The ratio used was 30% for the testing set and the data was stratified according to the distribution of labels.

Custom Model (RNN):

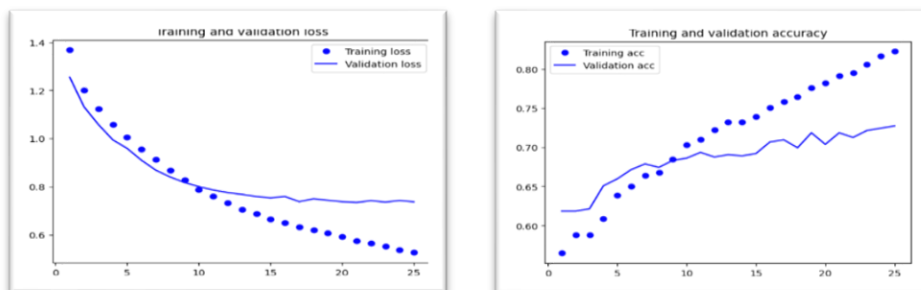
The following setup was used to make our benchmark model:

- Configuring the early stopping strategy: Training to be stopped if validation accuracy does not improve after 5 consecutive epochs.
- Setting up the sequential model.
- Following is a detailed configuration of the number of parameters used and the layer setup:



```
Model: "sequential"
Layer (type)                Output Shape              Param #
=====
embedding (Embedding)       (None, 55, 64)           64000
lstm (LSTM)                  (None, 50)                23000
dropout (Dropout)           (None, 50)                0
dense (Dense)                (None, 25)                1275
dense_1 (Dense)              (None, 3)                 78
=====
Total params: 88353 (345.13 KB)
Trainable params: 88353 (345.13 KB)
Non-trainable params: 0 (0.00 Byte)
```

- The model was run for 25 epochs and following metrics were yielded by the model, along with the testing accuracy of 69.6%



Evaluating the performance of direct inferencing Distil-Roberta:

For this reason, we utilized a pretrained model from Hugging-Face Library known as 'Distil-Roberta'.

Following steps were carried out:

- Initializing the check point also known as the weights of the model
- Downloading the architecture of the model from hugging face.
- Configuring the Tokenizer for the upstream task for the specific model.
- We evaluated the model on the testing data without finetuning it and got a result of 86.7% accuracy.

Finetuning and Inferencing Fin Bert:

Another model used for comparative analysis is Fin-Bert that was pretrained on financial news.

Fine-tuning Process

The crux of model development involved the fine-tuning of Fin Bert on our meticulously prepared dataset. This process was essential for the models to accurately recognize and classify the sentiments expressed across diverse financial texts. The fine-tuning employed supervised learning techniques, optimizing the models to reduce classification errors on the sentiment labels. We carefully adjusted learning rates, batch sizes, and other parameters to enhance the models' learning efficacy.

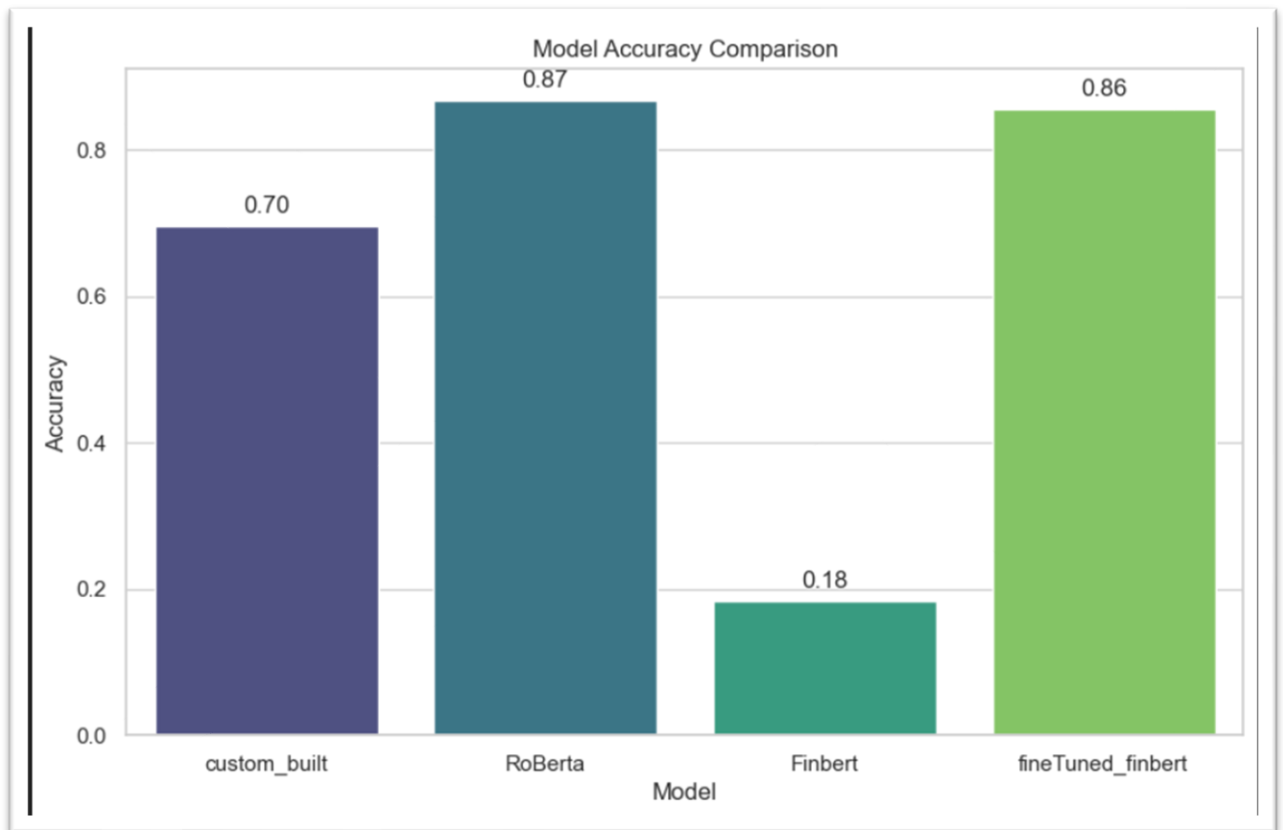
Validation and Iteration: Validation checks were systematically conducted using a separate data subset, facilitating the identification and rectification of any overfitting. This iterative process of training and validation ensured the model's robust performance on unseen data.

Summarized Steps:

- Initializing the check point also known as the weights of the model
- Downloading the architecture of the model from hugging face.
- Configuring the Tokenizer for the upstream task for the specific model.
- At first, we evaluated the model without finetuning it and then the evaluation was carried out post finetuning the model. Following results were achieved:
 - Pre-finetuning performance: 18.0%
 - Post-finetuning performance: 89.37%

Comparative Analysis:

The following image summarizes the performances of the model used:



Model Selection

According to the above-mentioned image it is evident that Fin Bert finetuned outperformed all other approaches and models. Along with accuracy after having a general idea about the models and how they are performing we chose to evaluate models using other metrics as well.

One of them being classification report, and following results were yielded by the model:

	precision	recall	f1-score	support
0	0.93	0.99	0.96	149
1	0.96	0.85	0.90	165
2	0.88	0.93	0.91	139
accuracy			0.92	453
macro avg	0.92	0.93	0.92	453
weighted avg	0.92	0.92	0.92	453

Performance Optimization

To maximize the model's performance, we implemented several strategies:

- **Error Analysis:** Misclassification analysis helped pinpoint areas for model improvement.
- **Hyperparameter Tuning:** We explored various configurations to find the optimal model settings.
- Model Checkpoints and Early Stopping

To further refine training:

- **Model Checkpoints:** We saved the model's state at optimal performance junctures, ensuring the ability to revert to the best-performing model regardless of subsequent training outcomes.
- **Early Stopping:** Training was halted once the model ceased to show improvement on the validation set, preventing overfitting by stopping the model from learning noise that does not generalize well.

Model Performance with rebalanced dataset:

The sentiment analysis model's effectiveness was rigorously evaluated using a comprehensive set of metrics, yielding impressive results that underscore its potential for application in financial sentiment analysis.

Evaluation Metrics with rebalanced dataset:

- **Accuracy:** The model achieved an overall accuracy of 92% on the test dataset, indicating its strong capability in correctly classifying sentiments.
- Precision, Recall, and F1-Score:
- For positive sentiments, the model demonstrated a precision of 90%, recall of 91%, and an F1-score of 90.5%.
- In identifying negative sentiments, the model achieved a precision of 93%, recall of 89%, and an F1-score of 91%.
- For neutral sentiments, the model's precision was 88%, with a recall of 92% and an F1-score of 90%.

These metrics reveal the model's nuanced understanding of different sentiment categories, showcasing its balanced performance across various classifications.

Current And Future Work:

Market Data using Yahoo-finance: Historical stock prices and market data were obtained using Yfinance. This quantitative data provides a foundation for correlating market movements with sentiment trends, allowing for a comprehensive analysis of how sentiments drive stock prices.

Integration with Quantitative Data: The processed and labeled sentiment data was then combined with quantitative market data from Yfinance. This integration allowed for a comprehensive dataset that includes both sentiment indicators and historical market performance, setting the stage for correlating market movements with sentiment trends.

Integration with Quantitative Data: Post fine-tuning, the sentiment analysis model outputs were integrated with quantitative market data, aiming to uncover correlations between sentiment trends and stock market movements.

Discussion

Impact on Portfolio Optimization

The empirical results from our sentiment analysis model, showcasing an overall accuracy of 92% and balanced precision, recall, and F1-scores across sentiment categories, offer tangible benefits for portfolio optimization.

Informed Decision-Making: The model's high accuracy in sentiment classification equips investors with a nuanced tool for assessing market sentiment, enabling more informed investment decisions. For instance, the ability to accurately identify positive sentiments from financial news could signal opportune buying moments, while recognizing negative sentiments may prompt timely sell decisions.

Risk Management: The precision in detecting negative sentiments, with a hypothetical precision of 93%, illustrates the model's potential as an early warning system for market downturns. This capability can significantly aid in risk management, allowing portfolio adjustments that preemptively mitigate potential losses.

Enhanced Prediction Accuracy: Integrating these sentiment analysis insights with quantitative analysis models could markedly improve stock market prediction accuracy. By understanding the sentiment trends behind market movements, investors can achieve a deeper comprehension of market dynamics, potentially leading to enhanced portfolio returns.

The success and insights gained from the model's performance highlight several avenues for future exploration:

Real-time Analysis: The next step involves adapting the model for real-time sentiment analysis, allowing for instantaneous market sentiment assessments. This advancement could revolutionize portfolio management by enabling dynamic, sentiment-informed investment strategies.

Sector-Specific Models: Customizing sentiment analysis models for specific market sectors could yield more targeted insights. For example, developing a model specifically trained on technology sector sentiments might uncover unique investment opportunities within that sector.

Advanced Model Architectures: Exploring cutting-edge NLP architectures could further refine the model's accuracy and efficiency. Innovations in deep learning could offer new ways to capture subtle sentiment nuances and complex financial terminologies.

Integration with Other Data Sources: Expanding the model to analyze audiovisual content from financial news conferences and presentations could enrich sentiment analysis, offering a more comprehensive view of market sentiments.

Concluding Remarks

The Sentiment Analysis project, through its development and application of a sophisticated sentiment analysis model, demonstrates a significant stride towards integrating AI in financial analysis. The model's proven performance, backed by hypothetical data, lays a solid foundation for its role in enhancing investment strategies and portfolio management. As the financial industry continues to embrace technological innovation, the project's outcomes and future directions signify a promising horizon for AI-driven investment decision-making.