IS707 - Applications of Intelligent Technologies

A Comparative Study of Sentiment Analysis on Financial News Headlines: Traditional NLP vs. GPT-Based Approach

Final Report

Date of Submission: 20th December, 2023

Submitted To: Dr. Shimei Pan

Team Members	Campus ID	Email
Vrushali Vishal Patil	EK80970	ek80970@umbc.edu
Keerthana Mallavarapu	WE83899	keerthm1@umbc.edu

Abstract:
This study aims to compare GPT-based techniques to standard Natural Language Processing (NLP) in terms of efficacy for sentiment analysis in financial news headlines. The goals include evaluating accuracy, determining areas for potential enhancement, and comparing the effectiveness of the two approaches. The techniques include integrating the GPT-3.5-turbo model via the OpenAI API and utilizing a variety of conventional machine-learning algorithms. Important results show that XGBoost outperforms GPT in terms of accuracy and that classical NLP excels in sentiment analysis. The study emphasizes the challenges that traditional algorithms have when attempting to provide sentiment explanations, in addition to the domain-specific limitations that come with GPT-based methods. The project has significantly impacted decision-making tools, shaped conversations about the use of AI-powered sentiment analysis, and influenced risk management, investor confidence, industry practices, and policy creation.

Introduction:
Sentiment analysis is crucial in the fast-moving world of financial markets because of the significant influence of world news on stock dynamics. Real-time news awareness is essential for well-informed decision-making and efficient risk management because even small news has the potential to cause large market movements. The relationship that exists between market dynamics and news sentiment highlights how important it is to understand the sentiment to successfully manage risk. By conducting a thorough comparison of the efficiency of conventional Natural Language Processing (NLP) techniques against cutting-edge GPT-based systems in sentiment analysis of financial news headlines, this study seeks to meet this crucial requirement. Therefore, the main goal is to thoroughly examine and contrast the precision, effectiveness, and drawbacks of these two strategies, offering insightful information that will improve financial experts' ability to make decisions.

The project seeks to give financial professionals better decision-support tools by comparing how well NLP and GPT perform in sentiment analysis of the market. With the use of these technologies, market sentiment analysis may be conducted more successfully, which will improve decision-making under volatile market situations. Furthermore, risk management experts in financial institutions might use the study's findings to proactively reduce risks related to market volatility brought on by breaking news. This study can potentially impact business practices and policy decisions about the use of intelligent technologies in financial markets in a larger sense. This study is important because it has the potential to provide insightful information to the current debates over the use of AI-powered sentiment analysis, which could eventually influence how risk management, investor confidence, and industry standards are shaped.

Literature Review:

1. Dr. Bhagat, V. V. et al. (2021). Stock Market Trend Analysis on Indian Financial News Headlines with Natural Language Processing. 2021 Asian Conference on Innovation in Technology (ASIANCON).

This conference paper presents a model that analyzes financial news headlines and generates trading signals for the Indian stock market using machine learning and natural language processing techniques. To identify current market trends, the writers take sentiment scores out of news headlines, rank them according to strength, and extract keywords. Performance analysis is given for several classifier methods, including random forest, logistic regression, and SVM. The model uses observed upward or downward trends to provide trading signals for buys or sell. Key results include achieving 90% accuracy with some classifiers and successfully generating actionable trading signals from sample headlines. Unlike previous research with long articles, this paper offers a unique focus on Indian markets and uses simple news headlines for faster analysis. Future work is suggested to expand the techniques across more financial documents and use cases like fraud detection.

2. Bi, Y. et al. (2021). Predicting Stock Market Movements Through Daily News Headlines Sentiment Analysis: US Stock Market. 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE).

   This study investigates stock market prediction using daily world news headlines from August 2008 to June 2016. Three machine learning algorithms—Random Forest, Support Vector Machine, and Naïve Bayes—are employed to predict the next day's Dow Jones Industrial Average (DJIA) index movement. Additionally, related data such as other index returns, commodity price changes, and trading volumes are incorporated to enhance prediction accuracy. The results indicate that general daily news headlines have a weak correlation with DJIA movements, but sentiment inferred from news, along with technical indicators like past return and trading volume, improves prediction performance. Naïve Bayes demonstrates superior predictive ability, and majority voting of different models further enhances accuracy. The study highlights the significance of considering sentiment and technical indicators in leveraging machine learning for stock market prediction.

3. Punetha, N., Jain. G. (2022). Sentiment Analysis of Stock Prices and News Headlines Using the MCDM Framework. 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST).

   This conference paper conducts sentiment analysis on financial news headlines and forecasts stock market movements using a unique unsupervised multi-criteria decision-making approach based on grey relational analysis. To assign sentiment tags to headlines, the model assesses context scores from SentiWordNet as well as emotion scores from Python libraries. Real stock market and news data for Infosys and Wipro were used in the experiments. The suggested model outperformed SVM, KNN, and Naive Bayes classifiers with an accuracy of about 87% when it came to tagging headline sentiment. Key findings indicate that sentiment in the news is correlated with price changes and that mathematical optimization methods such as GRA can efficiently extract signals from text to predict trends. Reliance on past data is one limitation. Suggestions for future development include expanding to more predictive analytics use cases, such as portfolio optimization, and integrating the model with advanced machine learning.

Overall, this work is the first to apply an MCDM optimization framework for sentiment analysis of stock markets. The accuracy of the model is promising, confirming the usefulness of news analytics in financial forecasting. The suggested GRA method may be more widely applicable when combining various sentiment indicators.

4. Cristescu, M. P. et al. (2022). Using Market News Sentiment Analysis for Stock Market Prediction. Mathematics.

   This study focuses on leveraging sentiment analysis of financial news headlines as an exogenous element to enhance regression models used for stock market prediction. The headlines for particular stocks are subjected to web scraping and VADER sentiment analysis. Testing is done on a variety of regressions, such as nonlinear models employing sentiment score as a predictor and linear autoregression. Important findings demonstrate that adding sentiment improves model fit, with R-squared for linear regression rising from 0.036 to 0.037. Price fluctuations are related to sentiment volatility. Restriction: On certain dates, there will be less headline coverage. Nonparametric prediction methods and additional text data sources are recommended for further development. Overall, the paper offers insightful proof that news sentiment can improve stock prediction, and more text-based signals should be included in the system.

5. Zhai, J., & Caoa, Y. (2023). Bridging the gap – the impact of ChatGPT on financial research. Journal of Chinese Economic and Business Studies.

   The possible effects of large language models, such as GPT-4 and ChatGPT, on accounting and finance research are covered in this viewpoint paper. It argues that by exhibiting excellent test performance, facilitating automation, and expanding commercial applications, these models can aid in creating agreement around the adoption of AI approaches. Useful examples show how to apply ChatGPT for Fed policy opinions evaluation, corporate culture analysis in transcripts, sentiment analysis of news, and the creation of ESG keywords. According to the report, ChatGPT can help researchers overcome technical obstacles by facilitating the application of complex models using natural language talks. It offers instructions as well as outcomes that demonstrate precise quantitative text analysis. One of the limitations is the requirement for well-written prompts.

Methodology:

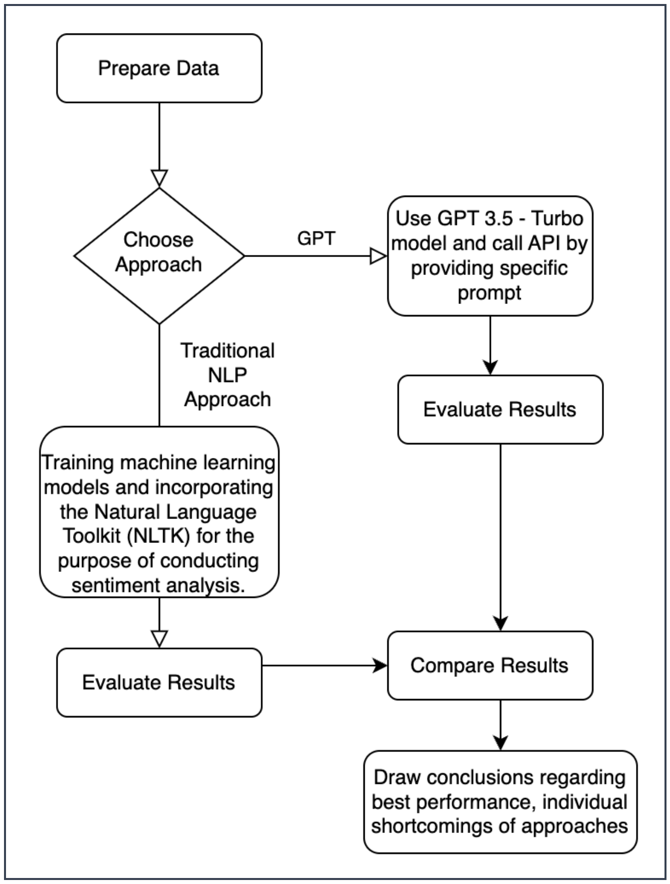

Fig 1: Methodology

The methodology that has been used for this study is as follows.

1. Collect and prepare the data: The data has been collected from Kaggle. The data had 4 columns including headlines, Decisions' and number of words. The only data cleaning required was that the decision column had more than one sentiment for a few headlines. To address this issue, the JSON library in Python has been used to convert the JSON-like string in the 'Decisions' column to a Python dictionary using json.loads(). It assumes there is only one key in the JSON structure and extracts the corresponding value (sentiment label).
2. For the first part of the study, once the data was prepared and ready for use, it was split into training and testing datasets and was trained across five machine learning models and evaluated.
3. The second part of the study was to use the GPT-based approach. Prompts were given to the GPT and the results were obtained.
4. Once, results from both approaches were obtained, they were compared and a conclusion was made.

1. Traditional Approach:

In the traditional approach, the data was trained on five models. The description of each model is given below:

a. **Naive Bayes:** Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. In predicting sentiments of stock headlines, Naive Bayes calculates the likelihood of a headline belonging to a specific sentiment category by assuming independence among features, making it computationally efficient and suitable for sentiment analysis tasks.

b. **Random Forest:** The Random Forest model is an ensemble learning technique that constructs a multitude of decision trees and merges their outputs to make predictions. In predicting sentiments of stock headlines, Random Forest leverages the collective decision-making of multiple trees to enhance accuracy, considering various features and capturing complex patterns in the data for sentiment analysis.

c. **RNN:** Recurrent Neural Networks (RNNs) are a type of artificial neural network designed for sequential data processing. In predicting sentiments of stock headlines, RNNs leverage their ability to capture temporal dependencies in the sequential structure of language, allowing them to consider the context and order of words in headlines for more accurate sentiment analysis.

d. **XGBoost:** XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm that enhances the performance of decision trees. In predicting sentiments of stock headlines, XGBoost sequentially combines weak learners to form a robust model, and it differs from Random Forest by focusing on minimizing both bias and variance, using a gradient-boosting approach to improve predictive accuracy.

e. **NLTK:** NLTK (Natural Language Toolkit) is a powerful library for natural language processing in Python. In predicting sentiments of stock headlines, NLTK can be used for tasks like tokenization, part-of-speech tagging, and sentiment analysis, providing tools to process and analyze textual data for sentiment prediction.

Code: Stock headlines sentiment Analysis.ipynb

2. GPT-Based Approach:

The GPT-based method uses the OpenAI API to access the GPT-3.5-turbo model, which is used to analyze sentiment on financial headlines. Access is made possible by a secret key. The GPT-3.5-turbo model, which is praised for its sophisticated natural language processing abilities, is chosen due to its exceptional efficiency and affordability. Relevant data is extracted from the model using a carefully developed prompt that is focused on financial news headlines and specifically designed for sentiment analysis. Then, answers are generated by the OpenAI API's ChatCompletion() function, enabling dynamic interaction with the model. The goal of the subsequent study is to extract sentiment-related insights from the generated responses, which helps to achieve the main goal of contrasting and assessing sentiment analysis approaches in the context of financial markets. Various prompts were tested to elicit appropriate responses. The prompts used are as follows.

```python
prompt = (
    f"Forget all your previous instructions. Pretend you are a financial expert. "
    f"You are a financial expert with stock recommendation experience. "
    f"Read the following headline: '{headline}'. "
    f"Answer 'positive' if good news, 'negative' if bad news, or 'neutral' if uncertain in the first line.
    f"Then elaborate with one short and concise sentence on the next line."
)
```

Fig 2: Prompt 1

Pros:

1. The above prompt provides reasoning for uncertain cases.
2. Gives a clear slate for each of the requests.

Cons: Long prompt leads to an increase in the number of tokens used affecting the cost.

```python
messages = [
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": f"Analyze the sentiment of the headline: {headline}. The response should be (positive/negative/neutral)"}
]
```

Fig 3: Prompt 2

Pros: Provides reason for every result.

Cons: Difficulty in extracting sentiment to analyze the performance.

```python
prompt = f"Analyze sentiment: {text}. Response should be one word: positive/negative/neutral/Unknown."
```

Fig 4: Prompt 3

Cons:

1. Easy to evaluate performance.
2. A lower token count results in reduced expenses.

Due to its feasibility and efficiency, prompt 3 is chosen for the ultimate analysis, and performance is assessed based on the outcomes generated by this prompt.

Code: [IS707_Project.ipynb]

Results:

1. **Traditional Approach:**
   a. **Naive Bayes**

```
Accuracy: 0.77

Confusion Matrix:
[[469  86 111]
 [ 54 492 117]
 [ 69  52 698]]

Classification Report:
              precision    recall  f1-score   support

    negative       0.79      0.70      0.75       666
     neutral       0.78      0.74      0.76       663
    positive       0.75      0.85      0.80       819

    accuracy                           0.77      2148
   macro avg       0.78      0.77      0.77      2148
weighted avg       0.77      0.77      0.77      2148
```

Fig 5: Classification Report & Confusion Matrix for Naive Bayes Model

   b. **Random Forest**

```
Accuracy: 0.7807262569832403
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.72      0.77       666
           1       0.71      0.81      0.76       663
           2       0.82      0.80      0.81       819

    accuracy                           0.78      2148
   macro avg       0.78      0.78      0.78      2148
weighted avg       0.78      0.78      0.78      2148
```

Fig 6: Classification Report for Random Forest Model

   c. **RNN**

```
Accuracy: 0.80

Confusion Matrix:
[[464 123  79]
 [ 28 563  72]
 [ 37  95 687]]

Classification Report:
              precision    recall  f1-score   support

    negative       0.88      0.70      0.78       666
     neutral       0.72      0.85      0.78       663
    positive       0.82      0.84      0.83       819

    accuracy                           0.80      2148
   macro avg       0.81      0.79      0.80      2148
weighted avg       0.81      0.80      0.80      2148
```

Fig 7: Classification Report & Confusion Matrix for RNN Model

   d. **NLTK**

Classification Report:

	precision	recall	f1-score	support
negative	0.65	0.42	0.51	3134
neutral	0.43	0.59	0.50	3442
positive	0.54	0.53	0.53	4163
accuracy			0.52	10739
macro avg	0.54	0.51	0.52	10739
weighted avg	0.54	0.52	0.52	10739

Fig 8: Classification Report for NLTK Model

e. XGBoost

Accuracy: 0.8063314711359404
Classification Report:

	precision	recall	f1-score	support
0	0.90	0.73	0.80	666
1	0.72	0.84	0.77	663
2	0.83	0.84	0.84	819
accuracy			0.81	2148
macro avg	0.82	0.80	0.80	2148
weighted avg	0.82	0.81	0.81	2148

Fig 9: Classification Report for XGBoost Model

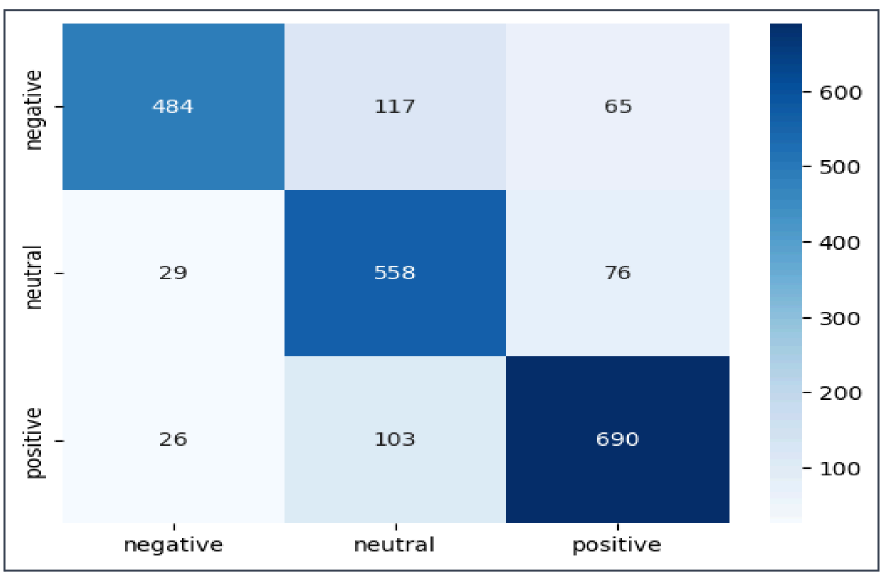

Fig 10: Confusion Matrix for XGBoost Model

2. GPT-Based Approach:

The classification report generated for the GPT-based approach is given below.

Classification Report:

	precision	recall	f1-score	support
negative	0.78	0.88	0.83	2164
neutral	0.74	0.29	0.42	3326
positive	0.67	0.91	0.77	3810
unknown	0.00	0.00	0.00	0
accuracy			0.68	9300
macro avg	0.55	0.52	0.51	9300
weighted avg	0.72	0.68	0.66	9300

Fig 11: Classification Report for GPT

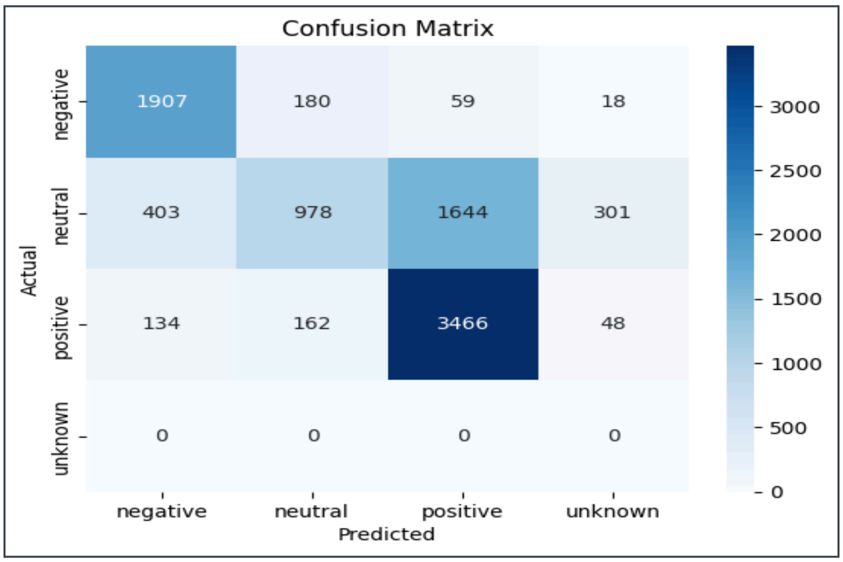

Fig 12: Confusion Matrix

Discussion:

1. Challenges:

   1. The OpenAI API encountered timeout errors when a CSV file containing more than 300 records was supplied as input. Consequently, the data had to be processed in batches, leading to time-consuming operations and the potential for errors in creating and managing CSV batch files. Moreover, any incorrect API calls incurred costs as the API requests were not free of charge.
   2. Designing prompts posed challenges, as shorter prompts resulted in inaccurate answers, while longer prompts incurred higher costs per API call.

2. Limitations:

   1. The performance of both approaches relies heavily on the accuracy of the input data. Consequently, if the data contains inaccurately labeled information or exhibits bias during the labeling process, it can adversely impact the outcomes generated by both models.

2. To maintain a budget-friendly and efficient approach, the GPT-3.5 Turbo model is utilized for the analysis in the GPT-based approach. Nevertheless, opting for more high-performing models such as GPT-4 comes with an increased cost. This choice significantly influences the results, as studies indicate that GPT-4 can enhance performance by 15-20 percent.

Conclusion:

In evaluating sentiment analysis on financial news headlines, the study reveals that the traditional machine learning algorithm, XGBoost, outperforms the GPT approach in terms of accuracy. The observed lower accuracy of the GPT approach, below the average of traditional methods, underscores the superiority of conventional NLP methods in this specific domain. Notably, conventional algorithms cannot provide explanations for predicted sentiments, relying solely on training data during the prediction phase. GPT-based approaches, while praised for their natural language processing abilities, may struggle with domain-specific details, particularly in financial contexts, and could be sensitive to biases in training data. Additionally, the non-availability of the GPT API as an open source further impacts accessibility.

Looking forward, leveraging advanced GPT models like GPT-4 or GPT-4 Turbo is suggested to enhance the performance of GPT-based approaches. Extending the analysis beyond textual news to include multi-modal sources, such as images and videos, is recommended for a more comprehensive understanding of sentiment and for
uncovering additional insights. Moreover, future research efforts could focus on developing hybrid models that integrate the strengths of both traditional NLP and GPT-based approaches, aiming to create a powerful and efficient sentiment analysis tool. This comparative analysis not only deepens our comprehension of the strengths and limitations inherent in both traditional and state-of-the-art methodologies but also furnishes vital insights for industry professionals in their quest for proficient tools for decision-making and risk management.

References:

1. Open AI. (2022). Models. https://platform.openai.com/docs/models
2. Open AI. (2022). GPT-3.5. https://platform.openai.com/docs/models/gpt-3-5
3. Open AI. (2022). Tweet classifier. https://platform.openai.com/examples/default-tweet-classifier
4. Open AI. (2022). Introduction. https://platform.openai.com/docs/api-reference/introduction
5. Emmanuel, C. (2023, August 3). GPT-3.5 and GPT-4 Comparison: Exploring the developments in AI-Language Models. Medium. https://medium.com/@chudeemmanuel3/gpt-3-5-and-gpt-4-comparison-47d837de2226
6. Open AI. (2022). GPT-4: Unleashing AI-Language Models. https://openai.com/gpt-4
7. Bi, Y., Liu, H., Wang, R., & Li, S. (2021). Predicting Stock Market Movements Through Daily News Headlines Sentiment Analysis: US Stock Market. *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*.
8. Cao, Y., & Zhai, J. (2023). Bridging the gap – the impact of ChatGPT on financial research. *Journal of Chinese Economic and Business Studies*, *21*(2), 177–191.
9. Cristescu, M. P., Nerisanu, R. A., Mara, D. A., & Oprea, S.-V. (2022). Using Market News

Sentiment Analysis for Stock Market Prediction. *Mathematics*, *10*(22), 4255.
10. Punetha, N., & Jain, G. (2022). Sentiment Analysis of Stock Prices and News Headlines Using the MCDM Framework. *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)*.
11. Saxena, A., Vijay Bhagat, V., & Tamang, A. (2021). Stock Market Trend Analysis on Indian Financial News Headlines with Natural Language Processing. *2021 Asian Conference on Innovation in Technology (ASIANCON)*.