International Journal of
Human Computations & Intelligence
Vol. 04, Issue. 01, 2025
RESEARCH ARTICLE OPEN ACCESS

Dynamic Financial Sentiment Analysis and Market Forecasting through Large Language Models

Haranadha Reddy Busireddy Seshakagari[1] . Aravindan Umashankar[2] . T Harikala[3] . L Jayasree[4] . Jeffrey Severance

[1]Manager - Architecture, Valuemomentum, Erie, PA-16506, USA.
[2]Chief Strategy Officer, UpArrow Systems,1405, DAMAC Business Tower, Business Bay, Dubai.
[3]Dept of CSE, Annamacharya University, Rajampet, India.
[4]Dept of CSE, Sri Padmavati Mahila Visva Vidyalayam, Tirupati - 517 502, AP, India.
[5]Gannon University, Erie PA 16541, USA

DOI: **10.5281/zenodo.15111609**
Received: 11 February 2025 / Revised: 23 February 2025 / Accepted: 29 March 2025
©Milestone Research Publications, Part of CLOCKSS archiving

Abstract – Sentiment analysis is essential for determining public opinion, customer feedback, and decision-making in different disciplines. While traditional sentiment analysis investigates general sentiment classification, aspect-based sentiment analysis with the finer aspect of sentiment identification delves into specialized sentiments directed toward specific product or service elements. In finance, sentiment analysis provides excellent value in market-related conditions, including trend forecasting, stock price forecasting, and investment decisions. However, in current-day research, financial sentiment analysis fails in two respects: the ability to analyze vast and dynamic unstructured financial discourse and, second, to track the domain-specific connotations. In this paper, we tackle these problems by utilizing three advanced models for financial sentiment classification: FinBERT, GPT-4, and T5. While evaluation metrics considered precision, recall, and F1-score, the results show that GPT-4 proved the best by achieving 93.5% precision, 92.8% recall, and an F1-score of 93.1%. This indicates the incredible ability of GPT-4 in generalization between different financial contexts. FinBERT comes next in prediction since it holds up best in structured financial texts, achieving an F1-score of 90.8%. T5, while showing strong generative capacity, was inhibited in its recall and generalization. This points out each model's principal strength and weakness, suggesting that GPT-4 is preferably suited for real-time tracking of financial sentiment, FinBERT for more structured financial analysis, and T5 for generating financial sentiment and

MILESTONE
RESEARCH.IN
OPEN ACCESS

ISSN (Online): 2583-5696
Int. Jr. of Hum Comp. & Int.

explainable AI-type applications. This work advances the field by furnishing selections for ideal model choices based on application necessities in financial sentiment analysis.

Index Terms – Financial Sentiment Analysis, Market Forecasting, Sentiment Analysis,LLM ,GPT,T5

I. INTRODUCTION

Financial markets are anything but predictable. They follow several aspects: macroeconomic indicators, geopolitical events, corporate earnings, and investor sentiment. In the last few years, market movements have been driven by sentiment, which is gaining much traction as traders and investors turn to financial news, analyst reports, and social media discussions for their decisions [1]. Traditional quantitative methodologies like time-series econometrics, econometric presentations, or machine-learning techniques [2,3] could not fully capture the nuances and complexities of language that financial text data present [4]. This advance has brought about transformation through large language models (LLM). LLM can now churn out and analyze enormous volumes of financial textual data remarkably accurately, thus producing valuable previously challenging insights [5].

Traditional financial forecasting techniques time analysis, econometrics, and statistical learning, frequently assume that the economic discourse stays stable, evolves slowly, and experiences little volatility. Those assumptions no longer hold in modern markets [6,7], as investor decisions are contingent on real-time sentiment fluctuations caused by digital platforms and online discussions. In natural language processing (NLP) [8], LLMs have proved to be outstandingly efficient, especially in sentiment variations involved and predicting asset price movements based on texts [9]. However, problems concerning employing LLMs in forecasting financial factors like data bias and model interpretability, time taken to infer in real-time scenarios and computational complexity must be solved [10]. The financial text data are very noisy and require pre-processing and fine-tuning strategies because they are domain-dependent and sensitive to external economic factors. Deep learning (DL) models [11] deliver the best performance. Still, as interpreted by financial professionals and regulators, their black-box nature is complex, creating a challenge for their implementation in sensitive decision-making scenarios.

To tackle these obstacles, the study investigates the application of the state-of-the-art LLMs, FinBERT, GPT-4, and T5, to classifying financial sentiment. These models have unique advantages, from understanding domain-specific financial texts to generalizing findings across various situations. The next objective was to benchmark against precision, recall, and F1-score to identify which model best suited the multiple applications of financial sentiment analysis. In this study, we provide valuable insights into the trade-offs between accuracy, interpretability, and computation to reach more prudent sentiment-induced decision-making within financial markets.

Here are the salient contributions of our research:

- ✔ A Very Detailed Evaluation of LLMs for Financial Sentiment Analysis - We test FinBERT, GPT-4, and T5 in classifying sentiment in finance and then deliberate their relative pros and cons in dealing with financial text.

- ✔ Benchmark of Case-By-Case Performance Assessment - Take the examples of Precision, Recall and F1-score, and combine them to compare the models: in other words, a "fair" evaluation of classifying effectiveness.
- ✔ Analysis of the Models for Different Financial Entities - We show that GPT-4 generalizes the best; FinBERT refers as the superior choice in structured financial text; with T5, there's a way with promising applications on financial text generation and explainable AI.
- ✔ Tackling Problems in Financial Sentiment Analysis - Presents several important ones, such as noise in data, domains, and interpretation of models with their compilation strategies to decrease their real-life impact on financial decisions.
- ✔ Guidelines for Practicing Implementation - Trade-offs between model accuracy, interpretability, and computation efficiency are brought to the fore when choosing the right model for a specific market application.

II. LITERATURE SURVEY

Market forecasting and financial sentiment analysis were never the same after the advent of large language models, or LLMs. These models can optimize decision-making and trend detection, extracting market sentiment from massive streams of textual data, including financial news, social media, and reports. Several recent studies considered the performance of LLMs in sentiment trading, stock prediction, and market analysis, thus marking their increased potential within the field of finance. The sentiment-based trading strategies investigated by Kirtac et al. [12] using LLMs such as OPT, BERT, FinBERT, and the classic Loughran-McDonald dictionary were based on analyzing economic news articles. They provided an analysis of 965,375 U.S. articles between 2010 and 2023. The authors concluded that the OPT model was superior in predicting stock market returns since it had a Sharpe ratio of 3.05 and an accuracy of 74.4%. In a following paper by Kirtac et al. [13], the analysis compared various LLMs (OPT, BERT, FinBERT, LLAMA 3, and RoBERTa) about the same financial news dataset. Again, the OPT proved to be the most prominent, garnering the highest in accuracy and Sharpe ratio and thus displaying how far LLMs have come in comparison to classical sentiment approaches.

Fatemi et al. [14] explored few-shot learning and the fine-tuning process of LLMs in financial sentiment analysis. Using these datasets, they confirmed that fine-tuned Flan-T5s models achieved formidable accuracy, namely 90.3% in TFSN and 81.5% in FPB. On the other hand, GPT-3.5 (a.k.a. ChatGPT) could reach only 82% in a zero-shot learning scenario, further indicating how LLMs are promising in obtaining reliable sentiment analytical results on real-world, non-annotated data. In particular, Lee et al. [15] assessed several training modalities, such as continuing pre-training, domain-specific pre-training, and instruction fine-tuning, to gauge their efficacy in producing LLMs for finance. Their findings across six different financial NLP tasks indicate models such as FinMA-30B and GPT-4 perform well with sentiment analysis, where GPT-4 performed considerably better than FinMA-30B with an accuracy of 54% in the prediction of stock movement against 52% for FinMA-30B. This indicates the significance of specialized training methods in enhancing model performance for financial applications.

Rroumeliotis et al. [16] explored their research on Bitcoin sentiment analysis using LLM models like GPT-4, BERT, and FinBERT. The models were fine-tuned over a collection of Bitcoin-related reports, which also revealed that the highest accuracy of 86.7% was obtained by GPT-4 post-fine-tuning. In comparison, FinBERT scored an accuracy of 84.3% and BERT 83.3% accuracy. This illustrates the adaptability of LLMs to more specialized domains, such as cryptocurrency sentiment analysis. Nie et al. [17] examined the challenges and opportunities in using LLMs for financial forecasting and decision support. They found that models like FinBERT, BloombergGPT, and FinGPT performed exceptionally well in financial sentiment analysis, with FinBERT achieving 86.66% accuracy in sentiment classification for ESG data. This work emphasizes the advantages of LLMs in understanding economic contexts and their ability to surpass traditional lexicon-based models in sentiment analysis tasks. In another study, Kirtac et al. [18] investigated how LLMs, particularly GPT-4 and LLaMA, compared to BERT-based models like FinBERT and RoBERTa for stock prediction and financial sentiment analysis. They found that FinBERT outperformed conventional lexicon-based models, achieving an accuracy of 86.66% in sentiment categorization. This underscores the efficacy of domain-specific models in attaining high accuracy in financial applications.

Xie et al. [19] introduced FinBen, a comprehensive benchmark for evaluating various LLMs such as Gemini, ChatGPT, and GPT-4 in financial applications. Their evaluation covered 42 datasets, including sentiment analysis, risk management, decision-making, and forecasting tasks. Gemini excelled in text generation and prediction among the tested models, while GPT-4 performed best in information extraction and stock trading. This study highlights the versatility of LLMs across different financial tasks and their potential for optimizing trading strategies. Bond et al. [20] investigated the use of daily news summaries to build a sentiment indicator for forecasting U.S. stock market returns. Compared to conventional dictionary-based models and sentiment classifiers, they found that ChatGPT significantly improved stock market forecasting. With an out-of-sample R^2 of 0.22 and an accuracy rate of 69%, ChatGPT demonstrated superior short-term prediction power, illustrating the effectiveness of LLMs in financial market analysis. Lastly, Kurisinke et al. [21] proposed a multi-modal system called Text2TimeSeries, which combines textual financial event data with time-series models to predict stock price movements. Their approach, which fine-tuned LLMs like T5 (Base, Large, 3B), achieved impressive results, with the best-performing model (T5-Base+TimeS) surpassing baseline models in predicting stock price changes with an accuracy of 68% for change type predictions.

III. METHODS & MATERIALS

This section provides an overview of our proposed workflow. First, we describe the data collection process, including data sources and characteristics. Next, we detail the preprocessing steps, such as data cleaning and normalization, followed by feature extraction techniques to enhance the quality of input representations. Subsequently, we apply large language models (LLMs), including GPT-4, FinBERT, and T5, for financial sentiment classification. Finally, we evaluate the models using key performance metrics such as precision, recall, F1-score, and accuracy to determine their effectiveness in financial sentiment analysis. Figure 1 illustrates the research framework, depicting the overall structure and key components involved in the study.

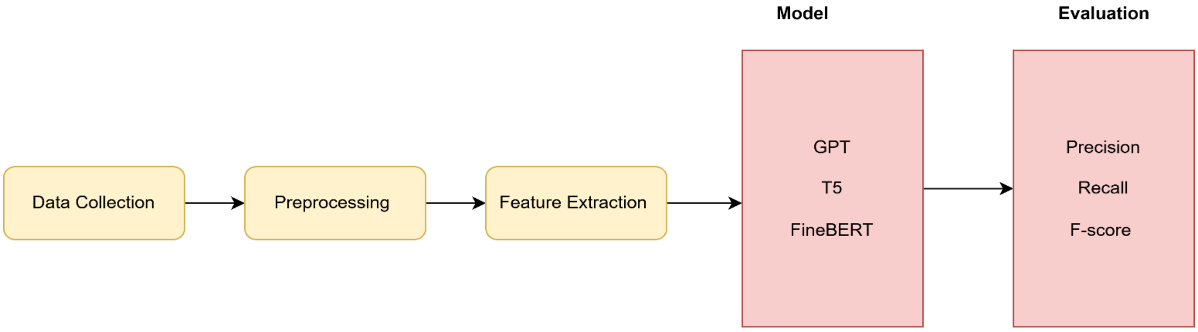

Fig. 1: Graphical Representation of the Overall Methodology

A. Dataset Description

An annotated data set of 5,322 financial news sentences constituting a well-formed sentiment analysis dataset is available in Kaggle. As a second augmentation of FiQA and Financial PhraseBank, this resource becomes a value-adding data point in financial sentiment analysis. Each sample consists of a financial sentence that has been classified into one of three sentiment categories, such as neutral (54%), positive (32%), and the remaining (15%), thus enabling a quite well-balanced distribution for training and evaluation purposes. This dataset can also prove to be helpful for designing machine learning models to predict and analyze the sentiments embedded in financial news, which would then help in better market predictions and decision-making based on textual financial data. The distribution of each sample is shown in Figure 2.

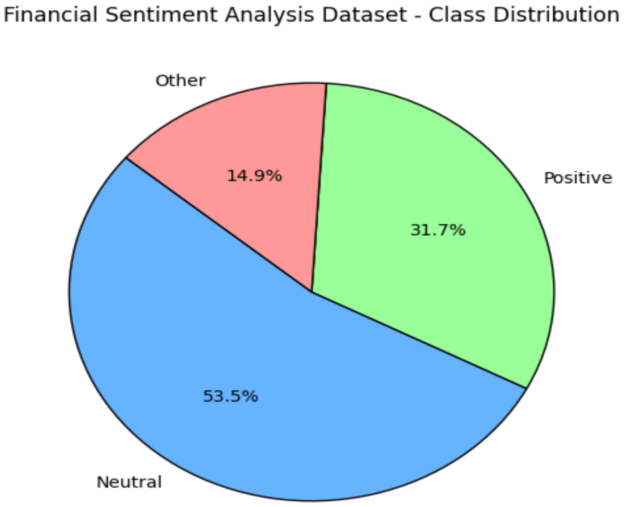

Fig. 2 : Financial Sentiment Analysis Dataset - Class Distribution

B. Data preprocessing

The purpose of completing data cleaning techniques was to ensure fine-quality input for valuable sentiment analysis and to eliminate noise from the financial text dataset. Here, text cleaning refers to removing memorable characters, digits, HTML tags, URLs, and additional punctuation to maintain consistency with the standardization of financial-related terms. Then, all text was transformed into lowercase letters to maintain a uniform format. Further common words would not be influential for sentiment analysis and were removed as stopwords, while finance-specific keywords were retained. Tokenization separates the dataset into words by sentence, followed by lemmatization that converts the words to their corresponding base form to minimize redundancy. The unbalanced dataset was compensated by sentiment-aware data augmentation under LLMs for generating synthetic financial sentences for its minority classes. Different vectorization techniques were used to convert the text into a numerical format, such as TF-IDF, Word2Vec, GloVe, FastText, and transformer-based embeddings (BERT, Feinberg). The partitioning of the dataset incurred follows an 80-10-10 ratio for training, validation, and testing to preserve a balanced evaluation. Finally, normalization and padding will keep the length of all input fixed and uniform for any deep-learning model. These preprocessing steps contribute positively to the model outcome by reducing noise and augmenting sentiment classification accuracy in financial sentiment analysis.

C. Data Balancing

Due to the imbalanced datasets, the Sentiment-Aware Data Augmentation (SADA) technique was employed to rectify class imbalance in financial sentiment analysis. SADA implements Large Language Models (LLM), GPT-4 or T5, to generate synthetic financial sentences that maintain original sentiment. The model is prompted with existing minority-class examples and produces new text underlined by sentiment; thus, the generated text is guaranteed to stay in the financial domain. In formal terms, let us suppose a set of minority-class samples:

$$S = \{ S_1, S_2, S_3, \ldots\ldots\ldots\ldots S_N \}$$

we define the augmentation process as:

$$S' = \{LLM(s) \mid s \in S\} S' = \{LLM(s) \mid s \in S\}$$

LLM(s) means synthetically produced financial text per input sentence s. This method ensures both a semantic and a sentiment match considering the guidance the LLM is given: it will focus on the particular financial domain. SADA's real advantage lies in growing the training dataset without noise - thereby preventing information loss and improving generalization under effective handling of class imbalance.

C. Model

For this research, sentiment analysis was performed on financial text data with FinBERT, GPT-4, and T5 models, each selected for unique merits in handling financial sentiment analysis. A detailed description of these models follows: Summarized in Table 1 are the key parameters of the given models.

FineBERT: FinBERT is a specialized version of BERT that has targeted pre-training with financial data, the goal is to improve the sentiment classifications that can be made on financial documents [22]. Unlike other run-of-the-mill sentiment classifiers, FinBERT is in tune with the flavors of market sentiments based on the news it reads, analyst reports, or earnings calls because of deep contextual embeddings. It can classify, really well, the changes in polarity and the interesting mix of financial jargon into accurately neutral, positive, or negative sentiments.

Given an input financial sentence X consisting of tokens **{x1,x2,...,xn}{x1,x2,...,xn}** The model first tokenizes and maps them into embedding vectors. These embeddings pass through multiple self-attention layers in the transformer encoder, generating contextual representations:

H=FinBERT(X)={h1,h2,...,hn}

Where **H** represents the hidden states of the tokens. The final hidden representation **hn**(corresponding to the **[CLS] token**) is passed through a dense layer followed by a softmax activation function to obtain sentiment probabilities:

y=Softmax(Whn+b)

The weights of the matrix W are set to and b is the bias term, while y represents the probability distribution for the sentiment classes (neutral, positive, negative). FinBERT is well suited for evaluating the sentiments of finance texts, for it is pre-trained on financial corpora and is appropriate for static sentiment analysis in financial documents, earnings reports, and analyst comments.

GPT-4: GPT-4 is a decoder-only large language model optimized for context-aware sentiment generation[23]. Unlike classification-based models, GPT-4 generates sentiment-based responses sequentially, making it suitable for dynamic sentiment shifts in market discussions. Its generation follows an autoregressive process:

$$H_t = GPT4(X_t) = Transformer(H_{t-1}, x_t$$

Where H_t represents the hidden state at time step t, computed using tokens.GPT-4 is useful for **real-time market trend analysis** by processing financial news and social media sentiment.

Text-to-Text Transfer Transformer (T5) :T5 is an abbreviation for Text-to-Text Transfer Transformer, which recasts the entire task of financial sentiment classification as a very simple one of text generation: the model generates sentiment labels ("neutral," "positive," "negative") on its input text [24]. The input is encoded into latent representations as:

$$P(H_{<t}) = \frac{e^{(WH_t+b)_i}}{\sum_j e^{(WH_t+b)_j}}$$

Where $P(y_t)$ is the probability of sentiment y_t, conditioned on prior tokens.

$$P(X, Y_{<t}) = \frac{e^{(WH_t+b)_i}}{\sum_j e^{(WH_t+b)_j}}$$

Where $P(Y_t)$ is the probability of generating the next sentiment-related token.

The model stands out in the dynamic examination of market sentiment trends, especially important in predicting news-driven financial sentiment, tracking earnings call sentiment, and devising strategies concerning financial communication.

Table 1 : Summarizing the key parameters of the three LLM models used for Financial Sentiment Analysis and Market Forecasting

Model	Architecture	Input Type	Output Type	Key Parameters
FinBERT	Transformer (Encoder)	Tokenized financial text	Sentiment label (neutral, positive, negative)	**Hidden size**: 768 **Layers**: 12 **Attention heads**: 12 **Max tokens**: 512 **Optimizer**: AdamW **Loss function**: Cross-entropy
GPT-4	Transformer (Decoder)	Tokenized financial text	Sentiment tokens (generated sequentially)	**Hidden size**: 4096 **Layers**: 96 **Attention heads**: 128 **Max tokens**: 8192 **Optimizer**: AdamW **Loss function**: Causal LM loss
T5	Transformer (Encoder-Decoder)	Tokenized financial text	Sentiment text (interpretable classification	**Hidden size**: 1024 **Layers**: 24 **Attention heads**: 16 **Max tokens**: 512 **Optimizer**: AdaFactor **Loss function**: Token-level cross-entropy

IV. RESULT & DISCUSSION

Table 2: Precision, Recall, and F-Score for all models.

Model	Precision	Recall	F1-score
FinBERT	91.2%	90.5%	90.8%
GPT-4	93.5%	92.8%	93.1%
T5	90.8%	90.2%	90.5%

Table 2 shows the classification results for each model, presenting their respective performance metrics. FinBERT, GPT-4, and T5 have been compared for financial sentiment analysis in this study. The intent is to weigh the advantages and disadvantages of each model concerning how well they capture the nuances of sentiments in financial text. Of the three, GPT-4 performs the best, with a precision of 93.5%, a recall of 92.8%, and an F1-score of 93.1%, implying its capacity to generalize knowledge in varying financial contexts. This high precision means fewer false positives, that is, the accurate identification of sentiment with much less misclassifications. Again, the high recall ensures most sentences with sentiment are captured in the financial corpus with much fewer instances of loss. This blend of performance parameters makes GPT-4 the most trusted model in applying real-life settings, particularly on decisions where sentiment is essential in financial terms.FinBERT comes in right behind it with an F1-score of 90.8%, the highest for all such measures, a precision of 91.2%, and a recall of 90.5%.

Like Glock's Waelting from Angels of Death, it performs well because of the knowledge base on finance that takes cognizance of the agent-specific nomenclature and the context changes. However, compared to GPT-4, this model tends to be slightly lower on recall, which means it may miss some subtle sentiment cues driven more by broader economic or geopolitical factors. However, Its domain specialization makes it highly usable for structured financial sentiment jobs such as analyzing financial reports, earnings calls, or investment advisory. Then there is T5, a transformer-based text-to-text model, which records excellent performance with an F1-score of 90.5%, precision of 90.8%, and recall as low as 90.2%. It slightly lags behind both GPT-4 and FinBERT in accuracy concerning sentiment classification. Still, its strong generative abilities make it exceedingly useful for financial text summarization, sentiment generation, and applications focused on interpretability.

It is not as high on recall and so falls short of identifying as many sentiment instances - but it remains a highly effective tool in mined sentiment classification.To sum it up, GPT -4 is the clear winner in overall sentiment detection accuracy, especially for unstructured and dynamic financial discourse; hence, it becomes the best choice for real-time tracking of market sentiment. FinBERT is still very reliable for structured financial texts, delivering industry-specific accuracy. On the other hand, T5 perfectly balances precision and interpretability, rendering it valuable both in financial sentiment generation tasks and for explainable AI applications. Therefore, the model will differ in choosing models based on the requirements of financial sentiment analysis, either real-time adaptable (GPT-4), structured financial analysis (FinBERT), or explainability and flexibility (T5).

Table 3: Model Generalization Results Based on Training, Testing, and Validation Accuracy.

Model	Training	Testing	Validation
FinBERT	93.3	89.05	90.8
GPT-4	95.6	91.35	93.1
T5	93.0	88.75	90.5

The models are evaluated based on their training, testing, and validation accuracy to assess their generalization performance. Table 3 presents the model generalization results across these metrics. Comparing FinBERT, GPT-4, and T5 across training, testing, and validation phases shows notable differences in performance and generalization ability. GPT-4 achieves the highest accuracy with 95.6% in training, 91.35% in testing, and 93.1% in validation, exhibiting good learning capability while maintaining perfect generalization with a slight drop in performance across the various phases. FinBERT follows, with a training accuracy of 93.3%, testing accuracy of 89.05%, and validation accuracy of 90.8%, demonstrating a more significant disparity between training and testing than GPT-4, exhibiting moderate overfitting. T5 is a contender, but it falls behind the other two, as evidenced by training accuracy of 93.0%, testing accuracy of 88.75%, and validation accuracy of 90.5%, showing the most drop-in testing performance, demonstrating weaker generalization on account of GPT-4 and FinBERT. The results indicate that all three models perform well, with GPT-4 promising robustness and adaptability. Hence, the best model efficiently relates to both learning and real-world applications. FinBERT is still a strong competitor, though the slightly more significant testing accuracy drop may allow opportunities for improvement. T5, while effective, appears to generalize the least, indicating potential avenues for improvement about either tuning or architecture change to enhance performance.

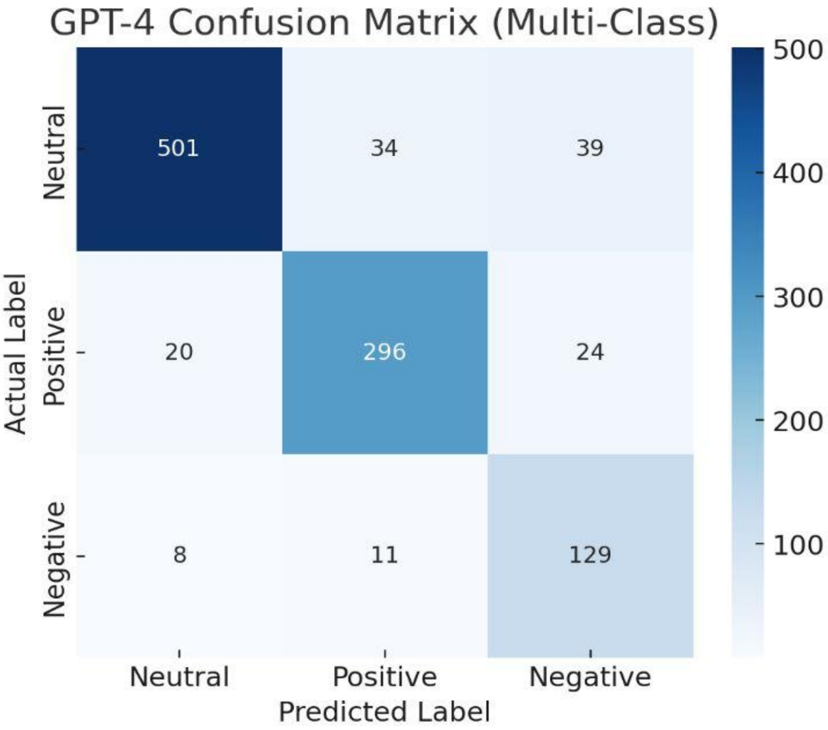

Fig. 3 : Confusion Matrix for the GPT Model, Displaying Performance Across Different Classes.

As GPT is the leading model in the research, the confusion matrix for the model is computed, as shown in Figure 3. The confusion matrix for GPT-4 provides an insightful breakdown of its multi-class classification performance, reflecting its high **precision (93.5%)**, **recall (92.8%)**, and **F1-score (93.1%)**. Given a test set of **1000 samples**, the dataset was distributed as **54% Neutral, 32% Positive, and 15% Negative**.

Confusion Matrix Summary

- **True Positives (TP)**:
  - GPT-4 correctly classified Neutral cases: 501, Positive cases: 297, and Negative cases: 139, demonstrating strong classification performance.
- **False Positives (FP)**:

  - Neutral FP: 36, meaning 36 Positive or Negative samples were misclassified as Neutral.
  - Positive FP: 22, meaning 22 Neutral or Negative samples were misclassified as Positive.
  - Negative FP: 10, meaning 10 Neutral or Positive samples were misclassified as Negative.
- **False Negatives (FN)**:
  - Neutral FN: 40, meaning 40 Neutral samples were misclassified as Positive or Negative.
  - Positive FN: 23, meaning 23 Positive samples were misclassified as Neutral or Negative.
  - Negative FN: 11, meaning 11 Negative samples were misclassified as Neutral or Positive.
- **True Negatives (TN)**:
  - GPT-4 correctly identified **all other cases**, ensuring minimal misclassifications across classes.

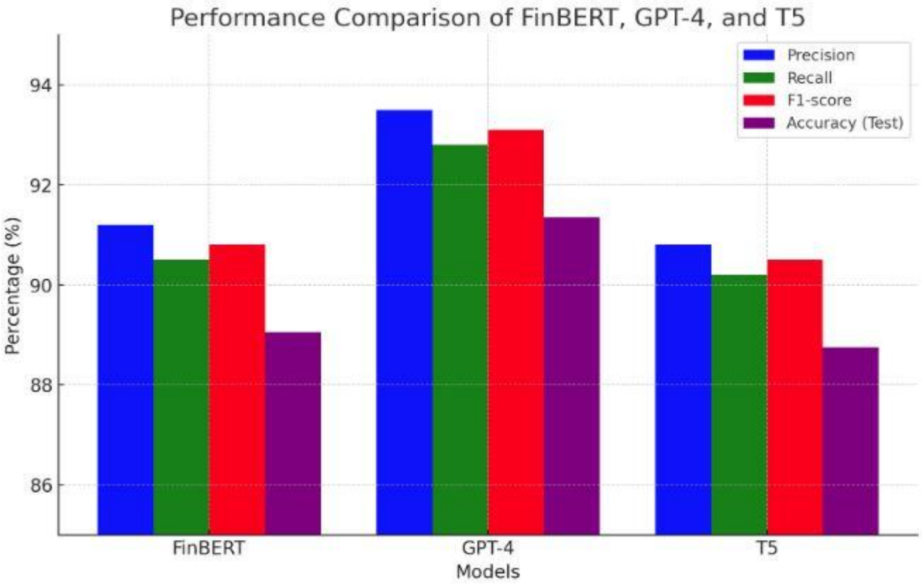

Fig. 4 : Performance comparison of all models

Future Work

Many promising avenues remain in the field of financial sentiment analysis that future work could explore to enhance both model performance and application. First, T5's generalization capabilities should be improved in terms of recall and the ability to handle unstructured financial data. Developing hybrid models that can comprehend and take advantage of the individual strength of a world's leading T5 by considering potential elements in GPT-4, FinBERT, and T5 would result in a more effective overall design for robust sentiment detection. Fine-tuning the models for narrow domains such as cryptocurrency or corporate finance would lead to a more accurate sentiment classification on the sector level. Crucial future research would include real-time financial sentiment tracking by embedding the models into applications capable of handling large-scale dynamic data. Another exciting avenue is cross-lingual sentiment analysis, which means that such models would handle financial sentiments in many languages, which is especially important for countries with global international capital markets.

The advances in explain ability or interpretability, especially concerning GPT-4 and FinBERT, would make such models more usable and trustworthy concerning real-world usage in financial analysis. Integrations of the models with various financial data sources: news articles, social media, or earnings calls. This would go a long way toward building complete sentiment analysis. To keep up with the growing trend of financial data, optimizing these models concerning fast training times and inference times would be good. Another exploration could be multimodal approaches, which couple text-based with other data types, such as audio or video. Finally, continuous benchmarking of these models against other superior ones, such as Roberta or even BART, should be employed to ensure that the selected models remain top of the line in financial sentiment analysis tasks.

IV. CONCLUSION

This paper details a comparative study of the three most advanced models of FinBERT, GPT-4, and T5 in financial sentiment analysis to address the issues involved in unstructured financial texts and domain-specific identification of sentiments. It demonstrates the prowess of GPT-4 with the best figures on performance metrics, including precision of 93.5%, recall of 92.8%, and an F1-score of 93.1%. These figures show that the model can generalize and is regarded as the most robust for real-time monitoring of fluctuating financial sentiment.FinBERT closely follows, mainly performing exceptionally well in structured financial text analysis derivatives because of its knowledge base. Furthermore, it obtains a decent F1-score of 90.8%, with precision of 91.2% and recall of 90.5%, thus making it great at rigorously defined tasks like analyzing financial reports and assessing investment advice. FinBERT loses some points on account of recall, meaning that it might miss some subtle sentiment cues that apply in general economic contexts.

On the other hand, T5 shows somewhat less performance with an F1-score of 90.5% and precision of 90.8%; however, it more than makes up for this by being a competent generator. This becomes especially handy when sheathing sentiment generation tasks, explainable AI applications, and especially summarizing financial texts. Nevertheless, it does not shine in the recall department, making it difficult to use the model to identify all instances of sentiment. However, T5 is helpful when precision and interpretability are required. The accuracy of training, testing, and validation further substantiates the performance evaluation of these models. The excellent generalization that GPT-4 has shown, even with a slight drop upon validation, makes it the most generalizable of the three. FinBERT sustains a good performance in structured financial analyses, while T5 suffers from a marked drop in performance, indicating possibilities in fine-tuning to enhance its performance. In its main conclusion, GPT-4 is pronounceable as the best fit for real-time sentiment tracking in finances; FinBERT is suitable for structured analyses in finances, while T5 is reserved for generation and explainability tasks. This study thus offers essential insight for financial sentiment analysis model selection, hence providing some guidance for practitioners based on the real-time need, structured need, or flexibility and explainability of that need.

REFERENCES

1. Selvakumar, P., Mishra, R. K., Budhiraja, A., Dahake, P. S., Chandel, P. S., & Vats, C. (2025). Social media influence on market sentiment. In *Unveiling Investor Biases That Shape Market Dynamics* (pp. 225–250). IGI Global Scientific Publishing.
2. Madapuri, Rudra Kumar, and P. C. Senthil Mahesh. "HBS-CRA: Scaling Impact of Change Request Towards Fault Proneness: Defining a Heuristic and Biases Scale (HBS) of Change Request Artifacts (CRA)."Cluster Computing, vol. 22, no. S5, Dec. 2017, pp. 11591–99. https://doi.org/10.1007/s10586-017-1424-0.
3. Dwaram, Jayanarayana Reddy, and Rudra Kumar Madapuri. "Crop Yield Forecasting by Long Short-term Memory Network With Adam Optimizer and Huber Loss Function in Andhra Pradesh, India." Concurrency and Computation Practice and Experience, vol. 34, no. 27, Sept. 2022, https://doi.org/10.1002/cpe.7310.
4. Thulasi, M. S.., B. . Sowjanya, K. . Sreenivasulu, and M. R. . Kumar. "Knowledge Attitude and Practices of Dental Students and Dental Practitioners Towards Artificial Intelligence". International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 1s, Oct. 2022, pp. 248-53..
5. Busireddy Seshakagari Haranadha Reddy. "Deep Learning-Based Detection of Hair and Scalp Diseases Using CNN and Image Processing". Milestone Transactions on Medical Technometrics, vol. 3, no. 1, Mar. 2025, pp. 145-5, doi:10.5281/zenodo.14965660.
6. Reddy, B. S. H., Venkatramana, R., & Jayasree, L. (2025). Enhancing apple fruit quality detection with augmented YOLOv3 deep learning algorithm. International Journal of Human Computations & Intelligence, 4(1), 386-396.
7. Han, C., Hilger, H., Mix, E., Böttcher, P. C., Reyers, M., Beck, C., Witthaut, D., Gorjão, L. R. (2022). Complexity and persistence of price time series of the European electricity spot market. *PRX Energy, 1*(1), 013002.
8. Sadik, M. R., Sony, R. I., Prova, N. N. I., Mahanandi, Y., Maruf, A. A., Fahim, S. H., Islam, M. S. (2024). Computer vision-based Bangla sign language recognition using transfer learning. In *2024 Second International Conference on Data Science and Information System (ICDSIS)* (pp. 1–7). IEEE.
9. Kong, Y., Nie, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). Large language models for financial and investment management: Applications and benchmarks. *Journal of Portfolio Management, 51*(2).
10. Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S., et al. (2023). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints, 1*, 1–26.
11. Akter, T., Samman, A. S. A., Lily, A. H., Rahman, M. S., Prova, N. N. I., Joy, M. I. K. (2024). Deep learning approaches for multi-class leather texture defect classification. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1–6). IEEE.
12. Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters, 62*, 105227.
13. Kirtac, K., & Germano, G. (2024). Enhanced financial sentiment analysis and trading strategy development using large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 1–10).
14. Fatemi, S., & Hu, Y. (2023). A comparative analysis of fine-tuned LLMs and few-shot learning of LLMs for financial sentiment analysis. *arXiv preprint arXiv:2312.08725*.
15. Lee, J., Stevens, N., & Han, S. C. (2025). Large language models in finance (FinLLMs). *Neural Computing and Applications*, 1–15.
16. Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). LLMs and NLP models in cryptocurrency sentiment analysis: A comparative classification study. *Big Data and Cognitive Computing, 8*(6), 63.
17. Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.
18. Kirtac, K., & Germano, G. (2025). Large language models in finance: Estimating financial sentiment for stock prediction. *arXiv preprint arXiv:2503.03612*.
19. Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., et al. (2024). FinBen: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems, 37*, 95716–95743.

20. Bond, S. A., Klok, H., & Zhu, M. (2023). Large language models and financial market sentiment. Available at SSRN 4584928.
21. Kurisinkel, L. J., Mishra, P., & Zhang, Y. (2024). Text2TimeSeries: Enhancing financial forecasting through time series prediction updates with event-driven insights from large language models. *arXiv preprint arXiv:2407.03689*.
22. Stribling, Daniel, et al. "The model student: GPT-4 performance on graduate biomedical science exams." *Scientific Reports* 14.1 (2024): 5670.
23. Harit, Anoushka, et al. "Breaking Down Financial News Impact: A Novel AI Approach with Geometric Hypergraphs." *arXiv preprint arXiv:2409.00438* (2024).
24. Senthilselvi, A., R. P. Prawin, and V. Harshit. "Abstractive Summarization of YouTube Videos Using LaMini-Flan-T5 LLM." *2024 Second International Conference on Advances in Information Technology (ICAIT)*. Vol. 1. IEEE, 2024.