FINANCIAL SENTIMENT ANALYSIS USING LARGE LANGUAGE MODELS

A PREPRINT

Siddhant Maji
Independence High School
Frisco, Texas
siddhant.maji@gmail.com

July 29, 2025

ABSTRACT

This research explores and benchmarks multiple approaches to financial sentiment classification using diverse sources such as financial tweets, headlines, and market commentary. Six models were evaluated: a traditional baseline TF-IDF + logistic regression, a transformer model (FinBERT), a neural network model using spaCy's TextCategorizer with ensemble architecture, and three large language models (LLMs) accessed via Azure OpenAI: o4-mini, GPT-4.1-mini, and a fine-tuned variant (GPT-4.1-mini-ft). All models were evaluated under a consistent preprocessing and label normalization pipeline. Results show that fine-tuning GPT-4.1-mini yields the best overall performance (Macro-F1), with significant improvements in classification accuracy. The fine-tuned GPT-4.1-mini model was then used alongside historical stock price and news data for NVIDIA to perform multivariate time series forecasting using an LSTM-based model. The time series forecast was significantly improved with the addition of fine-tuned LLM-enhanced news sentiment data, demonstrating a real-world downstream use case of LLMs for financial sentiment analysis.

1 Introduction

Sentiment analysis, the task of classifying text as positive, neutral, or negative, is increasingly used in the financial sphere for applications in trading, credit risk analysis, and analysis of investor behavior [1]. Financial text, however, is more difficult to analyze with machine learning models than generic sentiment tasks due to domain-specific jargon, brevity and imprecision of language, and fast-changing context [2].

Such nuances cannot be grasped by standard models that have been trained on general-purpose corpora. As a response, this project compares six models that represent different paradigms:

- **TF-IDF + Logistic Regression** (classical ML baseline)
- **FinBERT** (domain-tuned transformer)
- **spaCy TextCategorizer** (neural ensemble of Bag of Words (BoW) + attention)
- **o4-mini** (pre-trained LLM)
- **GPT-4.1-mini** (pre-trained LLM)
- **GPT-4.1-mini-ft** (fine-tuned version)

The performance is measured on the basis of three datasets, followed by a cost-performance analysis and a discussion of practical applications in stock price forecasting.

2 Related Work

Financial sentiment analysis has evolved such that it employs not just lexicon-based and rule-based techniques, but also machine learning and, most recently, large language models. Initial approaches were founded on rule-based models such as VADER [3] and spaCy TextBlob, which were founded on already existing dictionaries of words embodying

sentiment. These may be fast and comprehensible, but at the same time are also susceptible to misclassifications. The nuances of financial text are not normally captured by these approaches [4].

The introduction of machine learning to sentiment analysis allowed for more elaborate representations to be made. For example, TF-IDF, bag-of-words (BoW) and n-gram models with classifiers like Support Vector Machines and logistic regression [5] are more effective compared to rule-based systems when presented with large annotated data sets. Nevertheless, these methods still require precise feature engineering and struggle with context sensitivity.

BERT and other transformer-based models changed the NLP paradigm. One implementation of BERT was FinBERT, [6] a model trained in the financial context and calibrated on financial news statements and headlines. In spite of its effectiveness, FinBERT fails to perform conveniently without additional fine-tuning when applied to out-of-domain text like tweets or company filings.

Large Language Models (LLMs) such as GPT-3, GPT-4, and Claude have demonstrated impressive results in sentiment analysis. The models utilize extensive pretraining on a wide and vast variety of corpora and can be used to classify sentiment with little to no training. Recent models like FinGPT [7] [8] and RAG-enhanced LLMs for sentiment analysis [9] demonstrate the growing interest in combining generative models with financial tasks like sentiment classification.

This work is based on these fundamentals but focuses on evaluating both traditional and LLM-based architectures. It uniquely uses supervised fine-tuning of GPT-4.1-mini using Microsoft Azure AI Foundry for a downstream use case of stock price forecasting with LSTM models.

3 Datasets

Three publicly available datasets were used, covering financial news headlines and social media posts.

- **Financial PhraseBank (FPB)** [10]: 4,846 financial news snippets and headlines, manually annotated by human experts with labels (*positive*, *neutral*, *negative*).
- **FiQA-2018 Task A** [11]: A dataset consisting of 1,213 micro-blog messages, news headlines, and news statements with continuous sentiment polarity scores (-1 to 1) which were converted to three discrete classes (*positive*, *neutral*, *negative*).
- **Twitter Financial News Sentiment (TFNS)** [12]: A dataset of 11,931 financial tweets, manually annotated with three labels (*Bearish*, *Bullish*, *Neutral*).

All datasets were normalized to a unified label space (positive, negative, and neutral). The FPB dataset was excluded from FinBERT evaluations due to training overlap.

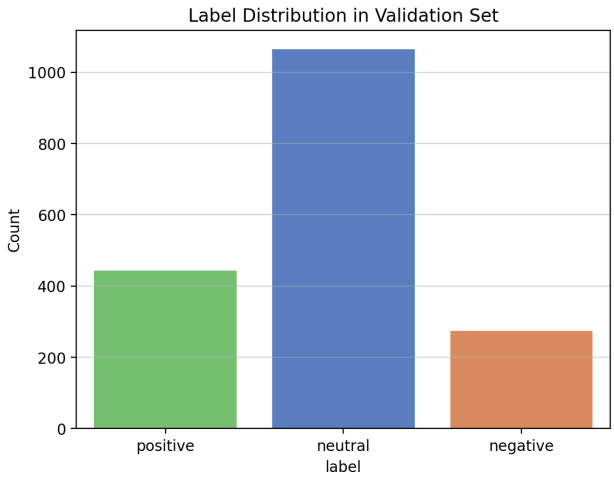

Figure 1: Label distribution across datasets.

3.1 Stock Forecasting Dataset

For the forecasting task, historical daily adjusted close price data for NVIDIA (NVDA), between July 30, 2024, to July 21, 2025, was collected from Yahoo Finance. Daily news for NVDA was fetched using Finnhub's Company News API endpoint. GPT-4.1-mini (Fine-Tuned) was used to classify sentiment for each news article summary, and daily average aggregated sentiment scores were combined with NVDA's daily adjusted close price to create a multivariate time series for stock price forecasting.

4 Methodology

4.1 Unified Sentiment Classification Pipeline

All datasets were passed through a shared pre-processing and evaluation pipeline. This included:

- Removal of special characters, non-alphanumeric tokens, and URLs
- Stratified sampling by for train/test splits (80/20)

4.2 Evaluation Protocol

Each model was evaluated on the same validation set using macro-F1 score, accuracy, and confusion matrices. Results were stored for reproducibility and compared using custom dashboards.

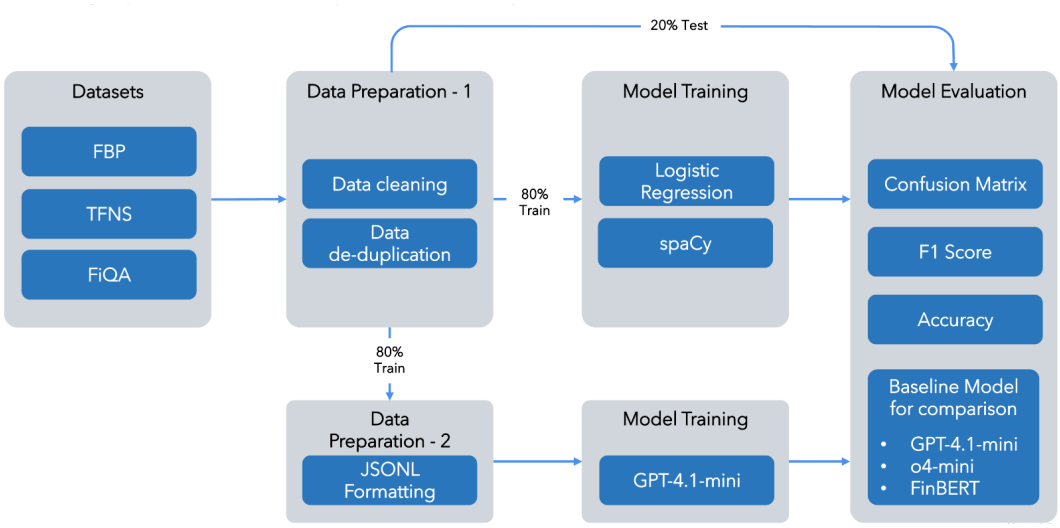

Figure 2: End-to-end pipeline from dataset to model evaluation.

4.3 Fine-Tuning GPT-4.1-mini

GPT-4.1-mini was fine-tuned via Microsoft Azure OpenAI using the preprocessed datasets (combined 80% train split). The data was formatted in JSON Lines (JSONL) format with task-specific instructions.

Fine-tuning required 1 training epoch with 1600 iterations, a batch size of 10, and a learning rate multiplier of 2.

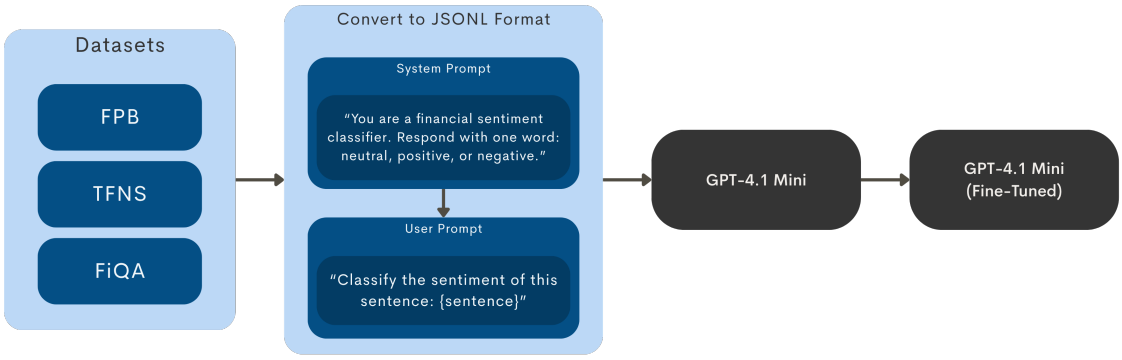

Figure 3: Fine-tuning pipeline

5 Modeling Approaches

Six models were benchmarked:

- **TF-IDF + Logistic Regression**: Traditional ML approach using unigrams and bigrams. [13]
- **FinBERT**: Transformer-based sentiment model pre-trained on financial corpora.
- **spaCy TextCategorizer**: Ensemble of linear BoW and attention-based neural layers (TextCatEnsembleV2).
- **o4-mini**: Pre-trained reasoning-capable large language model accessed via Microsoft Azure OpenAI API.
- **GPT-4.1-mini**: Pre-trained large language model accessed via Microsoft Azure OpenAI API.
- **GPT-4.1-mini-ft**: Supervised fine-tuned version of GPT-4.1-mini trained on task-specific financial sentiment data.

Table 1: Summary of Model Characteristics

Model	Approach	Training	Inference Type
TF-IDF + LogReg	Classical ML	Trained on vectorized text	Local (scikit-learn)
FinBERT	Transformer (Pre-Trained)	No training	Hugging Face Pipeline
spaCy TextCat	Neural Ensemble	Trained from scratch	spaCy NLP pipeline
o4-mini	LLM (Pre-Trained)	No training	Azure API
GPT-4.1-mini	LLM (Pre-Trained)	No training	Azure API
GPT-4.1-mini-ft	LLM (Supervised FT)	Supervised fine-tuned	Azure API

6 Experiments and Results

6.1 Evaluation Metrics

The following metrics were used for model evaluation:

- **Macro F1-score**: Harmonic mean of precision and recall, averaged across the three classes.
- **Accuracy**: Percentage of correct predictions.
- **Confusion Matrix**: Visual representation of classification performance.
- **Per-Dataset Breakdown**: Separate evaluation for FPB, FiQA, and Twitter data.

6.2 Performance Comparison

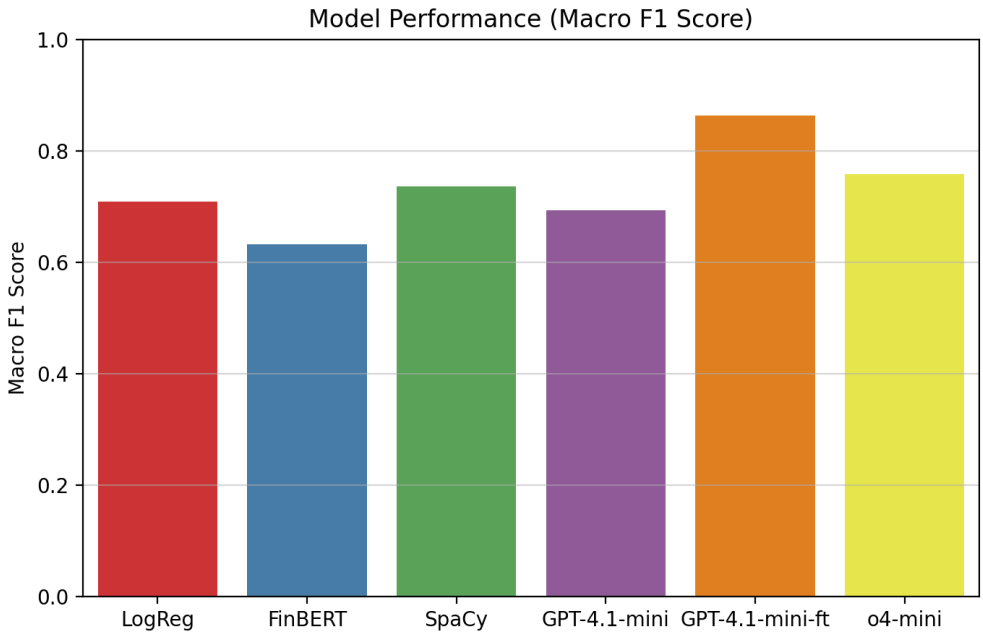

Figure 4: Macro F1 scores for all models across datasets.

Fine-tuning the GPT-4.1-mini large language model led to a higher accuracy and an improved F1 score on the FPB and TFNS datasets. However, zero-shot inference on the pre-trained o4-mini model performed significantly better on the FiQA dataset but performed worse than the fine-tuned GPT-4.1-mini model on the other datasets and overall.

Table 2: Performance of Sentiment Models Across Validation Datasets

Model	**FPB**		**TFNS**		**FiQA**	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**
LogReg	0.706	0.661	0.777	0.721	0.685	0.585
FinBERT			0.701	0.641	0.454	0.443
SpaCy	0.749	0.701	0.818	0.759	0.602	0.526
o4-mini	0.809	0.819	0.726	0.714	**0.815**	**0.737**
GPT-4.1-mini	0.743	0.764	0.656	0.657	0.759	0.656
GPT-4.1-mini-ft	**0.862**	**0.858**	**0.902**	**0.873**	0.741	0.677

6.3 Confusion Matrix Visualizations

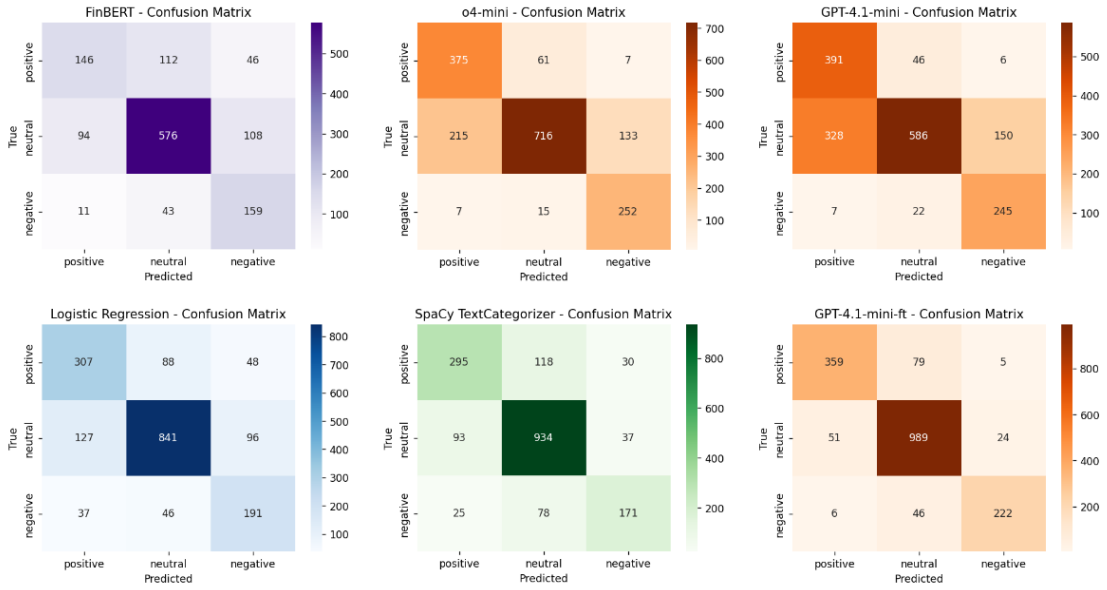

Figure 5: Confusion matrices for all models on the validation dataset. GPT-4.1-mini-ft (Fine-Tuned) shown to excel in neutral classification.

6.4 Cost and Hosting Overview

GPT-4.1-mini fine-tuning was done with Azure OpenAI's supervised fine-tuning API and cost approximately $6 in credits. All models were hosted using AWS Lambda (LogReg via custom layer, LLMs via REST endpoints). A static frontend using Vite + Tailwind CSS enabled web-based evaluation.

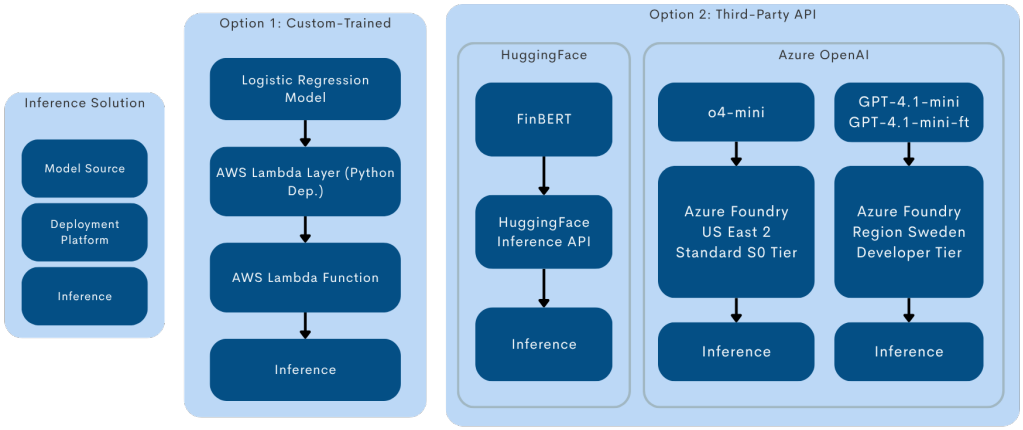

Figure 6: Deployment architecture: serverless LLM and model endpoints.

7 Stock Price Forecasting

7.1 Motivation and Setup

To evaluate the practical value of sentiment analysis, I explored its impact on time-series stock forecasting. I selected **NVIDIA (NVDA)** as a case study, given its frequent appearance in financial news and active social media presence.

7.2 Data Collection

- **Historical Prices**: Daily NVDA stock prices from Yahoo Finance (open, high, low, close, volume).
- **Sentiment Scores**: Daily aggregated sentiment from the GPT-4.1-mini-ft model applied to scraped tweets and headlines.

7.3 Model Architecture

I implemented an LSTM-based regressor with the following variations:

- **LSTM-baseline**: Trained on historical price features only.
- **LSTM+Sentiment**: Trained on both prices and daily sentiment scores.

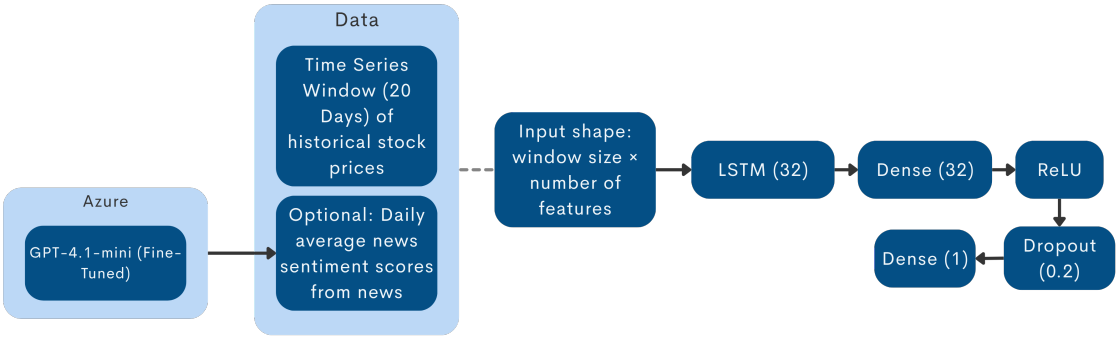

Figure 7: LSTM architecture with optional sentiment input branch.

7.4 Training and Evaluation

Both models were trained using:

- Window size: 20 days
- Optimizer: Adam
- Loss: Mean Squared Error
- Evaluation metrics: SMAPE, R^2, NMAE, etc. on test split

7.5 Results

The sentiment-enhanced LSTM consistently outperformed the baseline, showing that financial sentiment signals contribute to improved short-term forecasting accuracy.

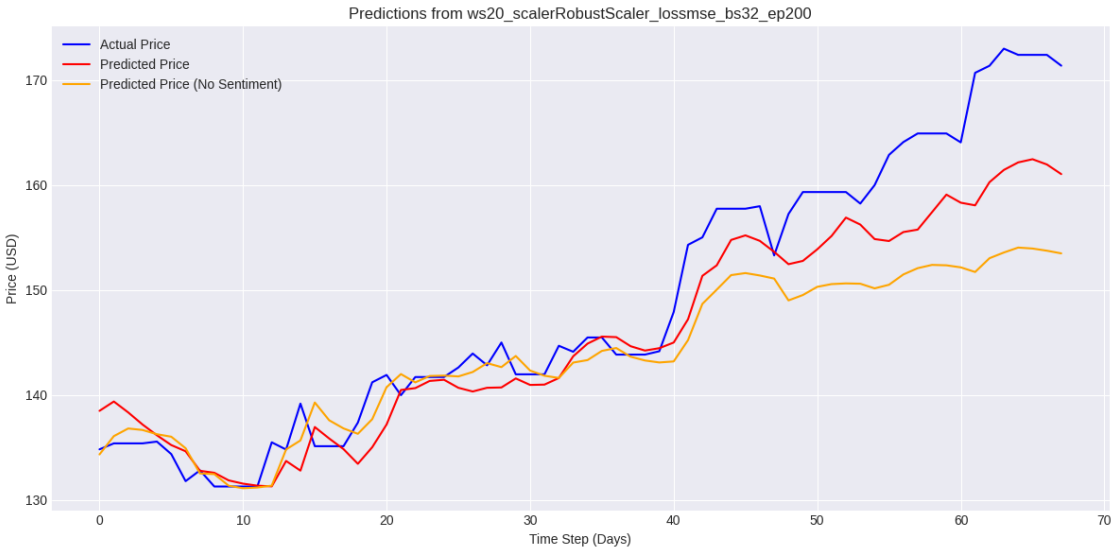

Figure 8: Predicted vs actual closing prices (LSTM baseline vs sentiment-augmented).

8 Discussion and Limitations

8.1 Model Insights

The experiments show that fine-tuned LLMs can substantially outperform traditional sentiment models, especially in neutral class detection. Large language models like o4-mini perform strongly but still trail behind fine-tuned LLMs in terms of accuracy.

Interestingly, spaCy's TextCategorizer, also performed fairly well, providing a balance between accuracy and performance.

8.2 Forecasting Insights

Sentiment-augmented LSTM models demonstrated improved evaluation metrics nearly across the board over price-only baselines. This indicates that high-quality sentiment classification may provide complementary signals beyond numerical indicators.

8.3 Limitations

- **Domain Sensitivity**: Finetuned models work best on Twitter/News text seen during training, generalization to other domains remains unclear.
- **Label Quality**: FiQA and TFNS annotations may include noise or inconsistent labeling standards.
- **News Data for Sentiment Extraction**: I used only company news within a year's time frame because of cost limitations and ease of access. Using data from other contexts like social media could be explored.
- **Cost**: LLM fine-tuning and API inference remains expensive at scale.

8.4 Future Work

- Use transformer-based time-series models like TimeGPT [14] and Chronos [15]
- Expand datasets to include multilingual financial corpora from more contexts
- Apply RAG for context-enhanced sentiment understanding [9]
- Explore reinforcement learning [16] fine-tuning for market-driven policy optimization [17] with reasoning models like o4-mini

9 Conclusion

This study benchmarks a wide range of approaches for financial sentiment analysis and investigates its utility in real-world stock price forecasting.

Key contributions include:

- A unified sentiment classification pipeline, covering TF-IDF + Logistic Regression, FinBERT, spaCy TextCategorizer, and both pre-trained and fine-tuned LLMs (GPT-4.1-mini, o4-mini).
- An empirical comparison across three datasets (FPB, FiQA, TFNS), with consistent pre-processing and evaluation metrics.
- Supervised fine-tuning of GPT-4.1-mini using Azure OpenAI, yielding the highest F1-score and accuracy compared to the other models compared, with only $6 in compute cost.
- Hosting of all models on cloud infrastructure using AWS Lambda, enabling a real-time sentiment classification web demo.
- An end-to-end application of sentiment scores in an LSTM-based NVIDIA stock forecasting system, with the sentiment-enhanced model achieving superior accuracy.

The findings suggest that LLMs, when fine-tuned with even modest data, deliver state-of-the-art performance in financial sentiment tasks. In addition, their outputs offer real utility in financial time-series forecasting.

Future work may extend to real-time trading signal generation, multilingual datasets, and integration with hybrid models such as reinforcement learning or RAG-enhanced systems.

Acknowledgments

This work was conducted as an independent student research project. I would like to thank Vlad Birsan and Haoyuan Yang who are instructors at the University of Texas at Dallas Deep Dive AI Workshop. I would also like to thank Professor Anurag Nagar who guided me in the initial brainstorming of project ideas, and the authors of the Financial PhraseBank, FiQA, and Twitter Financial News Sentiment datasets.

References

[1] Xiliu Man, Tong Luo, and Jianwu Lin. Financial sentiment analysis(fsa): A survey. In *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*, pages 617–622, 2019.

[2] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. Dombert: Domain-oriented language model for aspect-based sentiment analysis, 2020.

[3] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. 01 2015.

[4] Gourab Nath, Arav Sood, Aanchal Khanna, Savi Wilson, Karan Manot, and Sree Kavya Durbaka. On quantifying sentiments of financial news – are we doing the right things?, 2023.

[5] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[6] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.

[7] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models, 2023.

[8] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models, 2023.

[9] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models, 2023.

[10] Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Good debt or bad debt: Detecting semantic orientations in economic texts, 2013.

[11] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. WWW'18 open challenge: financial opinion mining and question answering. In *Companion Proceedings of the the Web Conference*, pages 1941–1942, 2018.

[12] Neural Magic. Twitter financial news sentiment. `http://precog.iiitd.edu.in/people/anupama`, 2022.

[13] Nabila Wardana, Firza Aditiawan, and Anggraini Sari. Logistic regression classification with tf-idf and fasttext for sentiment analysis of linkedin reviews. *VISA: Journal of Vision and Ideas*, 4, 08 2024.

[14] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1, 2024.

[15] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024.

[16] Jer Min Eyu, Kok-Lim Alvin Yau, Lei Liu, and Yung-Wey Chong. Reinforcement learning in sentiment analysis: a review and future directions. *Artificial Intelligence Review*, 58(1):6, Nov 2024.

[17] Yixuan Liang, Yuncong Liu, Neng Wang, Hongyang Yang, Boyu Zhang, and Christina Dan Wang. Fingpt: Enhancing sentiment-based stock movement prediction with dissemination-aware and context-enriched llms, 2025.