Enhancing Financial Sentiment Analysis using Large Language Models

Sentiment Classification • LLM Fine-Tuning • Stock Forecasting

Presented by

Siddhant Maji

Independence High School

Date

July 31, 2025

Website

https://finsa-utd.web.app/

GitHub

https://github.com/sidmaji/financial-sentiment-analysis/

Agenda

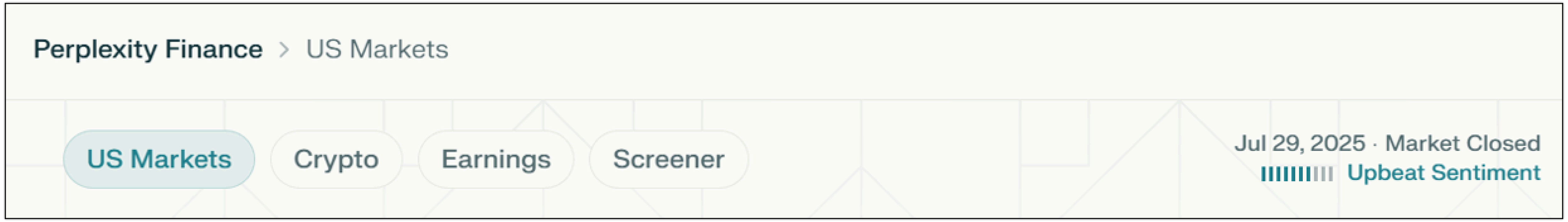
- Introduction
- Solution Overview
- Tenets of Solution
- Deployment & Demo
- Application to NVDA Stock Prediction
- Conclusions
- References

https://finsa-utd.web.app/

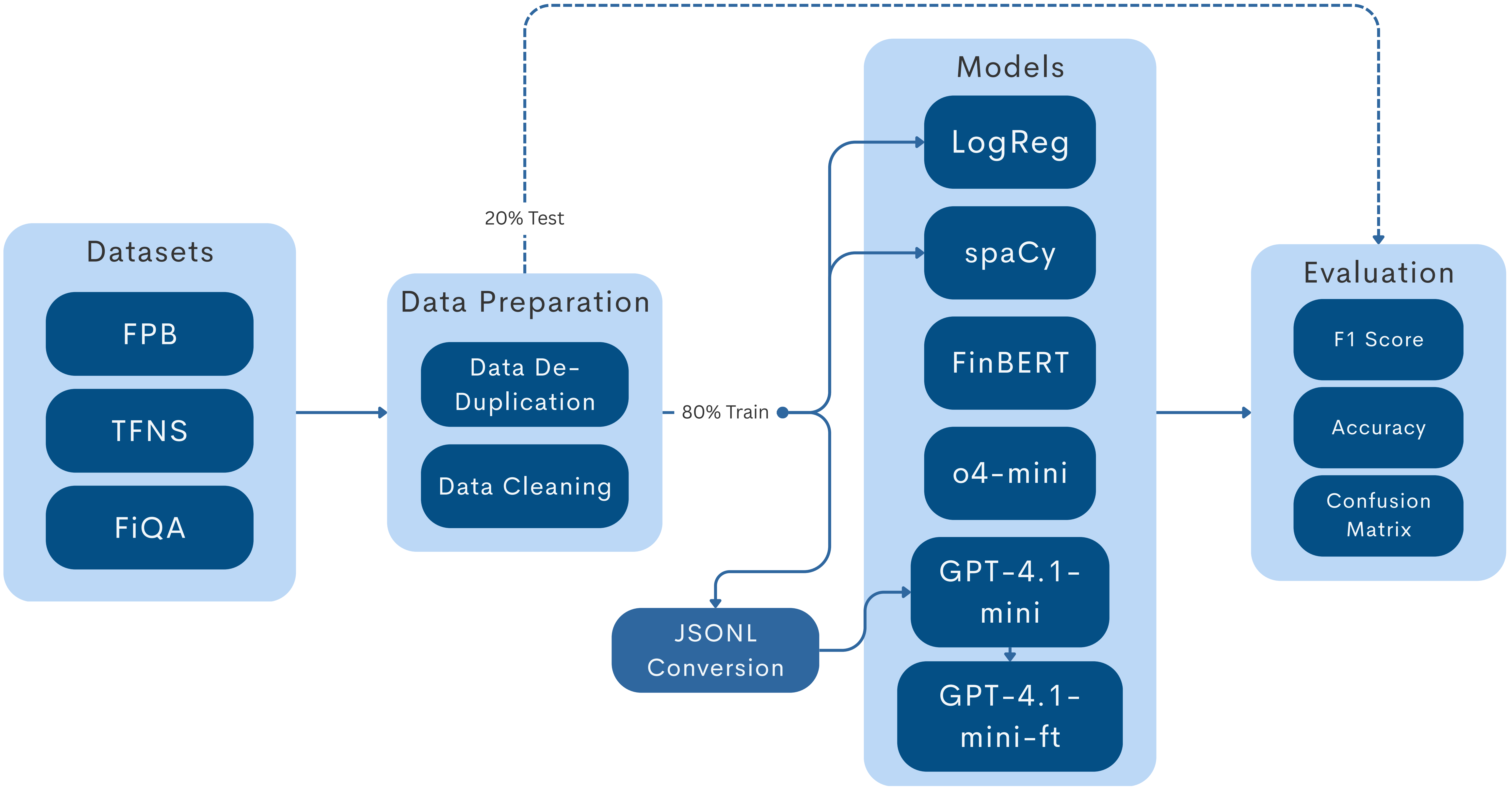
Importance of Sentiment Analysis in the Financial Domain

- Finance professionals dedicate a significant amount of time to analyze reports, news, social media todetermine the market sentiment and make crucial investment and trading decisions. [1]
- Objective of this research:
    - Perform Sentiment Analysis of news and financial texts using Large Language Models (LLMs).
    - Improve accuracy of these models by fine-tuning them using supervised learning.
- Compare and evaluate the trained LLM models against traditional ML and untrained Finance models.
- Apply News Sentiment Analysis using the trained models to predict stock prices.

Screenshot from Perplexity Finance



Solution Overview



Datasets

Datasets

FPB

TFNS

FiQA

Dataset	Source	License	Description	Size
Financial PhraseBank (FPB)	HuggingFace	CC BY-NC-SA 3.0	Manually annotated news sentences	4,838
Twitter Financial News Sentiment (TFNS)	Kaggle	MIT	Real-world tweets	11,931
Financial Question Answering (FiQA)	HuggingFace	MIT	News sentences + blog messages.	1,111

Data Preprocessing - 1

```python
def clean_text(text):
    if pd.isnull(text):
        return ""
    # Remove URLs
    text = re.sub(r"http\S+|www\.\S+", "", text)
    # Remove non-alphanumerics
    text = re.sub(r"[^a-zA-Z0-9\s]", "", text)
    # Normalize spacing
    text = re.sub(r"\s+", " ", text).strip()
    return text
```

"Macy's has shrunk to the size of Beyond Meat. It's going to get worse, technician says $M $BYND (via @TradingNation) https://t.co/h86IfyLmqy"

“Macys has shrunk to the size of Beyond Meat Its going to get worse technician says M BYND via TradingNation”

Data Preprocessing - 2

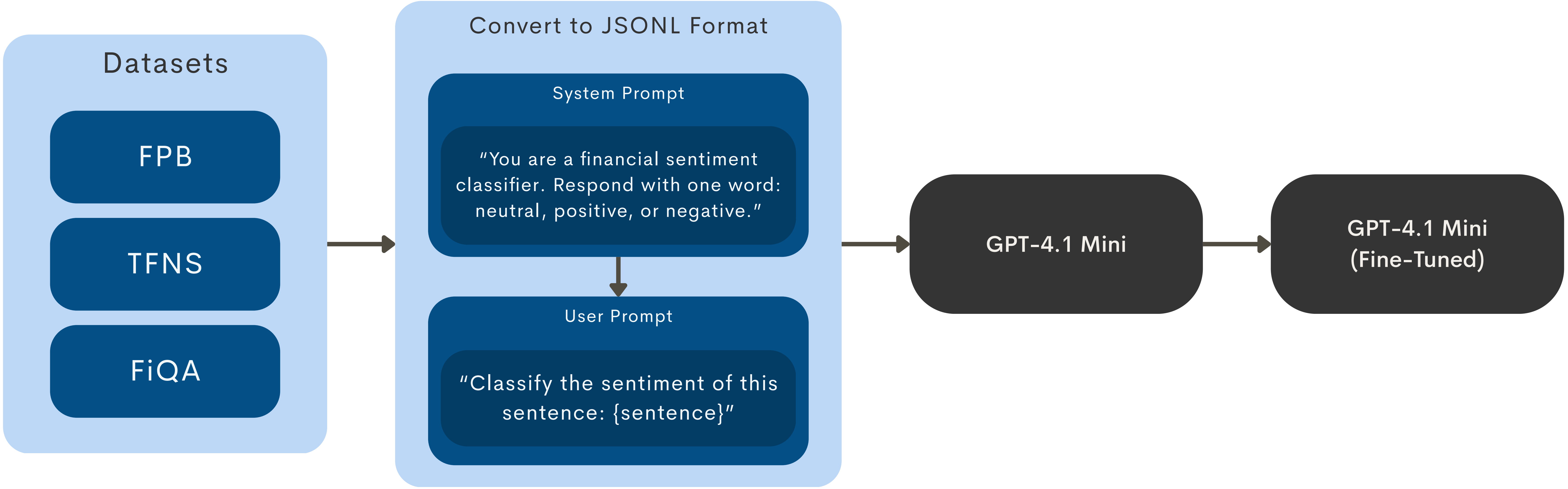
For GPT-4.1-mini – Conversion to JSONL Format

```
system_prompt = "You are a financial sentiment classifier. Respond with one word: neutral, positive, or negative."

with open("../data/processed/train.jsonl", "w", encoding="utf-8-sig") as out_file:
    for _, row in train_df.iterrows():
        sentence = row["sentence"]
        label = row["label"].lower().strip()

        item = {
            "messages": [
                {"role": "system", "content": system_prompt},
                {
                    "role": "user",
                    "content": f"Classify the sentiment of this sentence: {sentence}",
                },
                {"role": "assistant", "content": label},
            ]
        }
        json.dump(item, out_file, ensure_ascii=False)
        out_file.write("\n")
```

Sentiment Classification Models

Model	Type	Description
TF-IDF + Logistic Regression	Classical ML	Converts text to numerical TF-IDF vectors; predicts sentiment using a logistic regression classifier.
FinBERT	Transformer (BERT)	Pretrained BERT model fine-tuned on Financial PhraseBank for sentiment analysis.
spaCy TextCategorizer	Lightweight Neural Net	Ensemble of BoW linear model and neural network (Tok2Vec + attention); trained on financial data.
GPT-4.1 Mini	Large Language Model	Smaller general-purpose GPT model used in zero-shot setting for sentiment classification.
o4-Mini	LLM (Reasoning-focused)	Optimized GPT model for reasoning and precision; used in zero-shot sentiment classification.
GPT-4.1 Mini (Fine-Tuned)	LLM (Supervised FT)	GPT-4.1 Mini fine-tuned on labeled financial sentiment data for improved domain-specific accuracy.

Fine-Tuning Process



Fine-Tuning Process

Create a fine-tuned model

Method

Supervised

Base model

gpt-4.1-mini-2025-04-14

Training type

Standard

Standard training offers data residency guarantees. No data leaves the current resource region.

Automatically deploy when fine-tuning is complete (Preview)

No

Training data *

train.jsonl Edit

Validation data

val.jsonl Edit

Refresh | Continual fine-tuning | Delete | Pause | Cancel

Model attributes

ID
ftjob-9e2dec8fc72c499da71bc5668f52a1be

Status
Queued

Base model
gpt-4.1-mini-2025-04-14

Created on
Jul 20, 2025 12:26 AM

Training file
train.jsonl

Validation file
val.jsonl

Azure OpenAI resource
SentimentLLM-UTD-Sweden

Weights & Biases integration enabled?
No

Method of Customization
Supervised

Training type
Standard

Download results | Download training file | Download validation file

Details | Metrics | Logs | Checkpoints

Use this model | Refresh | Continual fine-tuning | Delete | Pause | Cancel

Model attributes

ID
ftjob-9e2dec8fc72c499da71bc5668f52a1be

Status
Completed

Base model
gpt-4.1-mini-2025-04-14

Created on
Jul 20, 2025 12:26 AM

Training file
train.jsonl

Validation file
val.jsonl

Azure OpenAI resource
SentimentLLM-UTD-Sweden

Weights & Biases integration enabled?
No

Method of Customization
Supervised

Training type
Standard

Training completed on
Jul 20, 2025 1:47 AM

Duration
1h 21m 28s

Training tokens billed
937,000

Download results | Download training file | Download validation file

Task parameters

Batch size
10

Learning rate multiplier
2

Number of epochs
1

Seed
42

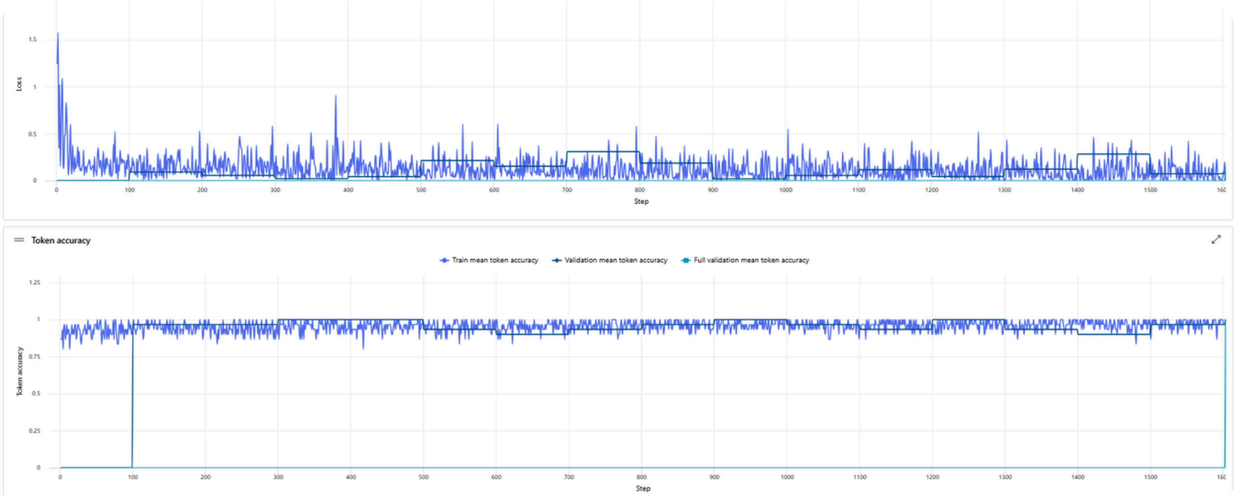
← gpt-4.1-mini-2025-04-14.ft-9e2dec8fc72c499da71bc5668f52a1be-001

Details | Metrics | Logs | Checkpoints

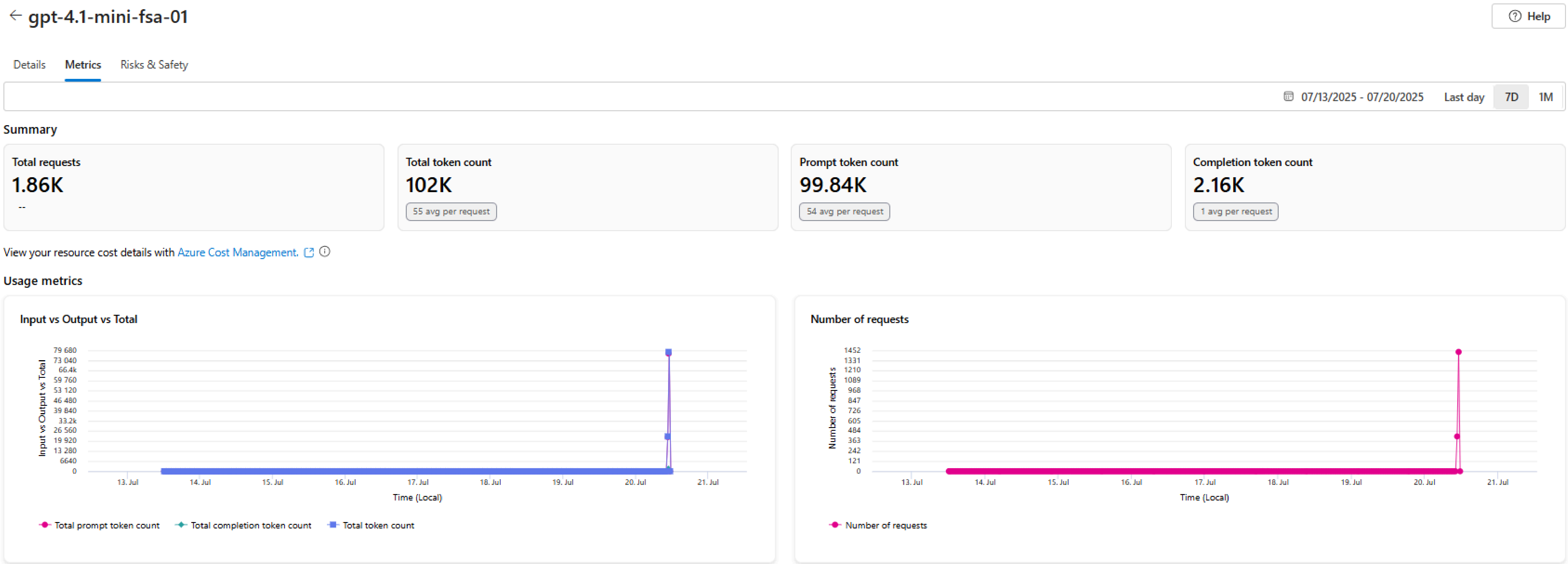
Refresh | Continual fine-tuning | Reset view

Search

Checkpoint name	Full validation loss	Created on	Step number	Checkpoint ID
gpt-4.1-mini-2025-04-14.ft-9e2dec8fc72c499da71bc5668f52a1be-001	0.09915062463308082	Jul 20, 2025 1:35 AM	1603	ftchkpt-f94a286beb674c3c9a5eb7d5228b690c



Azure Model Inference



Model Evaluation Results

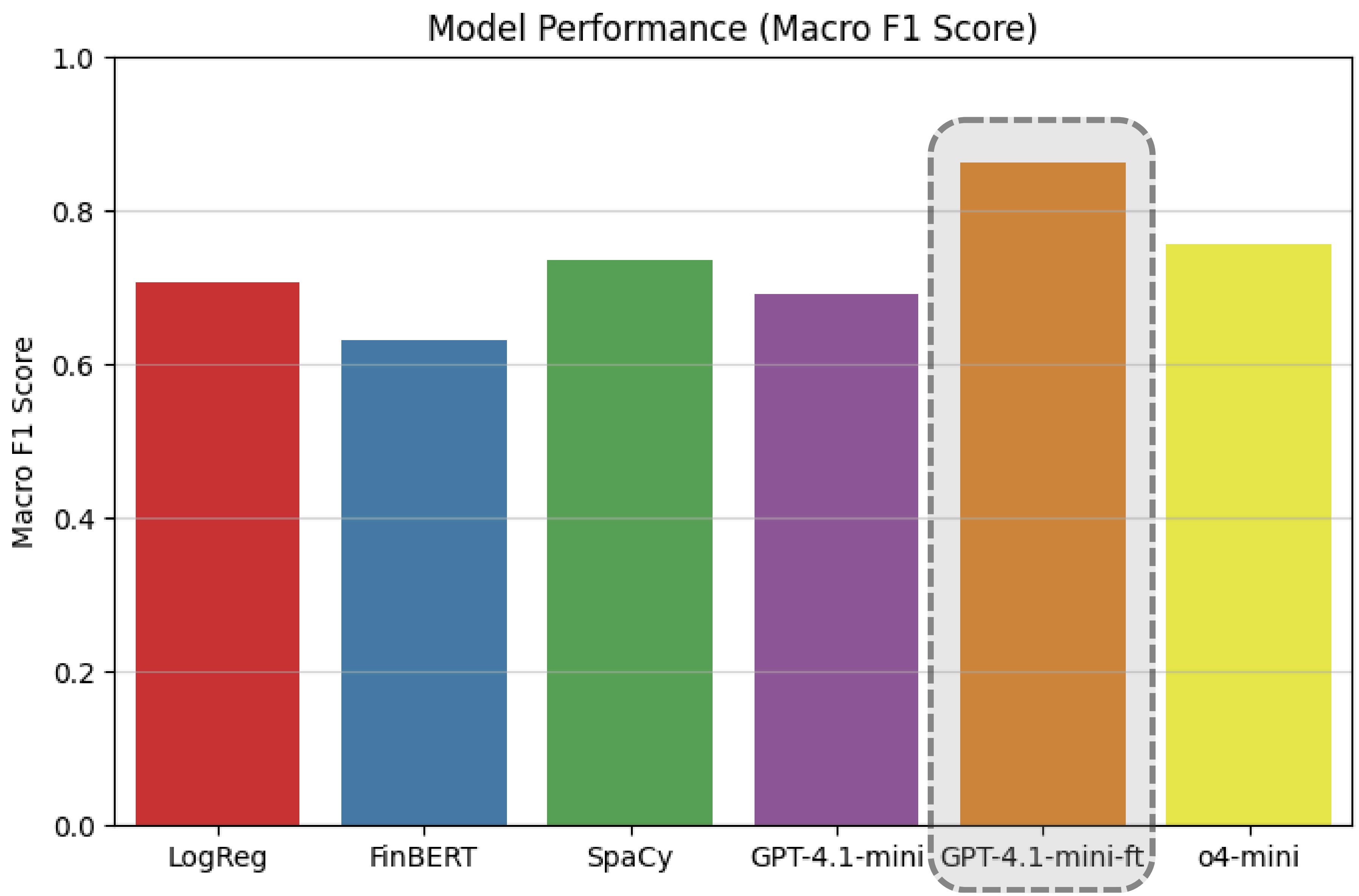
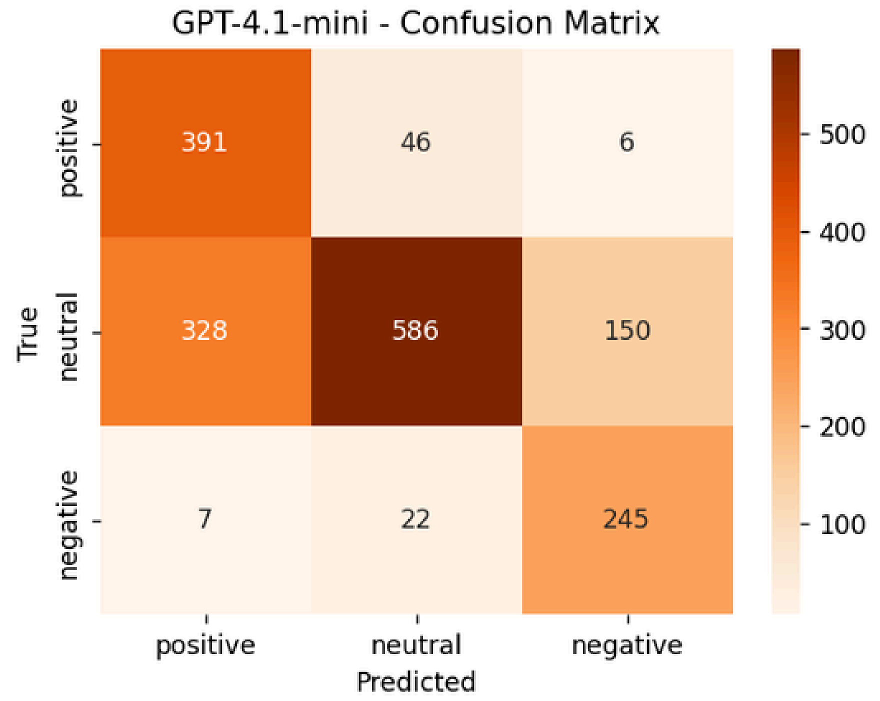
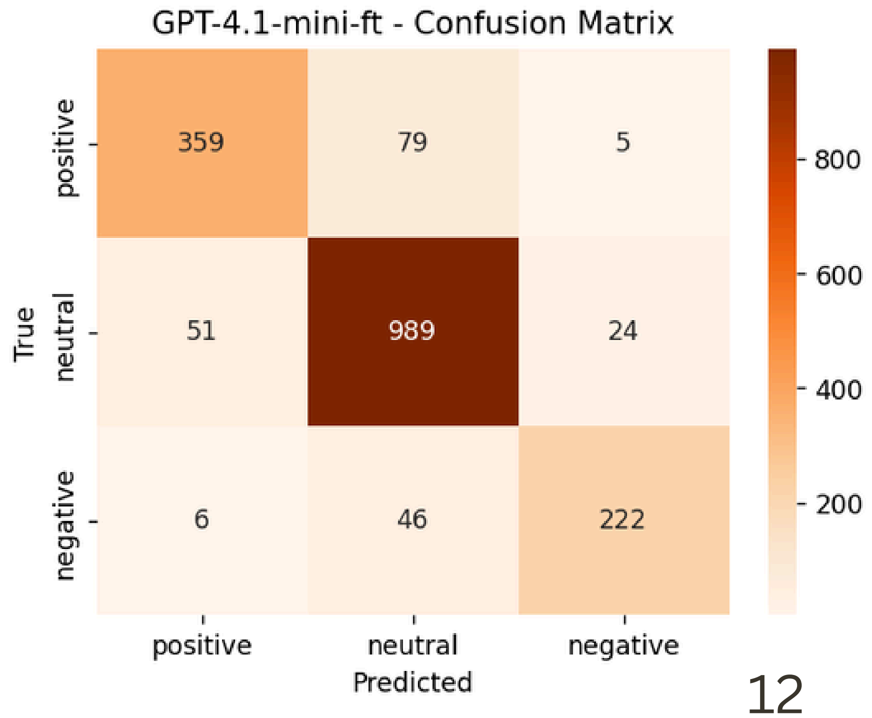


Table 2: Performance of Sentiment Models Across Validation Datasets

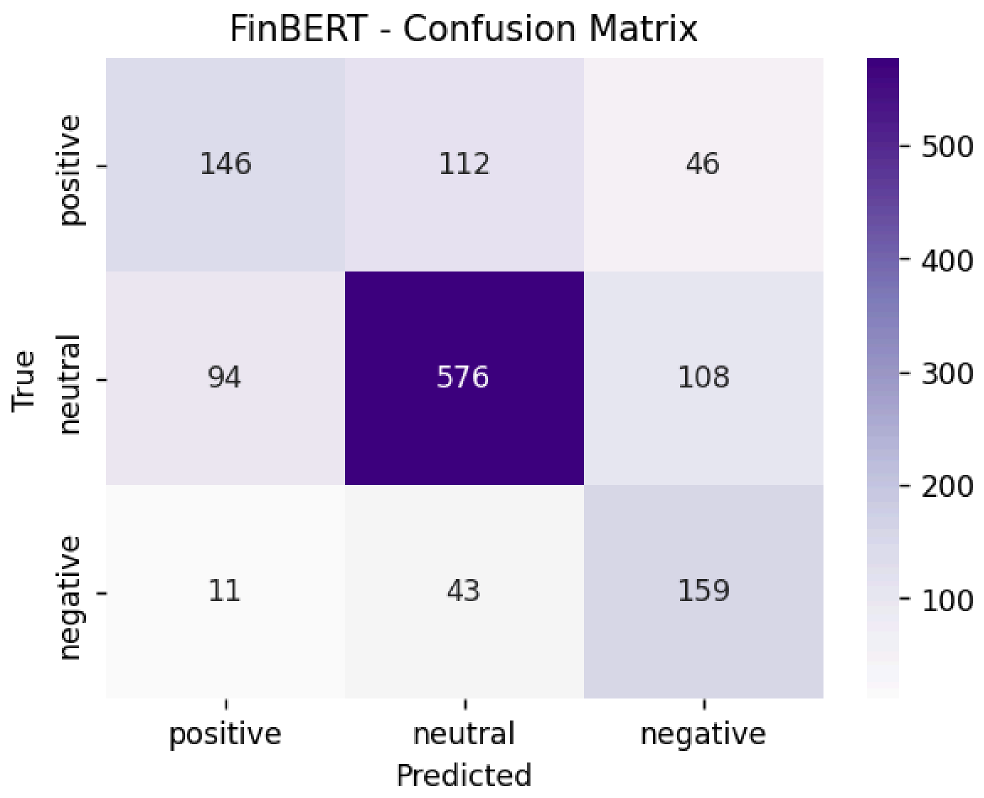
Model	FPB		Twitter		FiQA	
	Acc	F1	Acc	F1	Acc	F1
LogReg	0.706	0.661	0.777	0.721	0.685	0.585
FinBERT			0.701	0.641	0.454	0.443
SpaCy	0.749	0.701	0.818	0.759	0.602	0.526
o4-mini	0.809	0.819	0.726	0.714	**0.815**	0.737
GPT-4.1-mini	0.743	0.764	0.656	0.657	0.759	0.656
GPT-4.1-mini-ft	**0.862**	**0.858**	**0.902**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

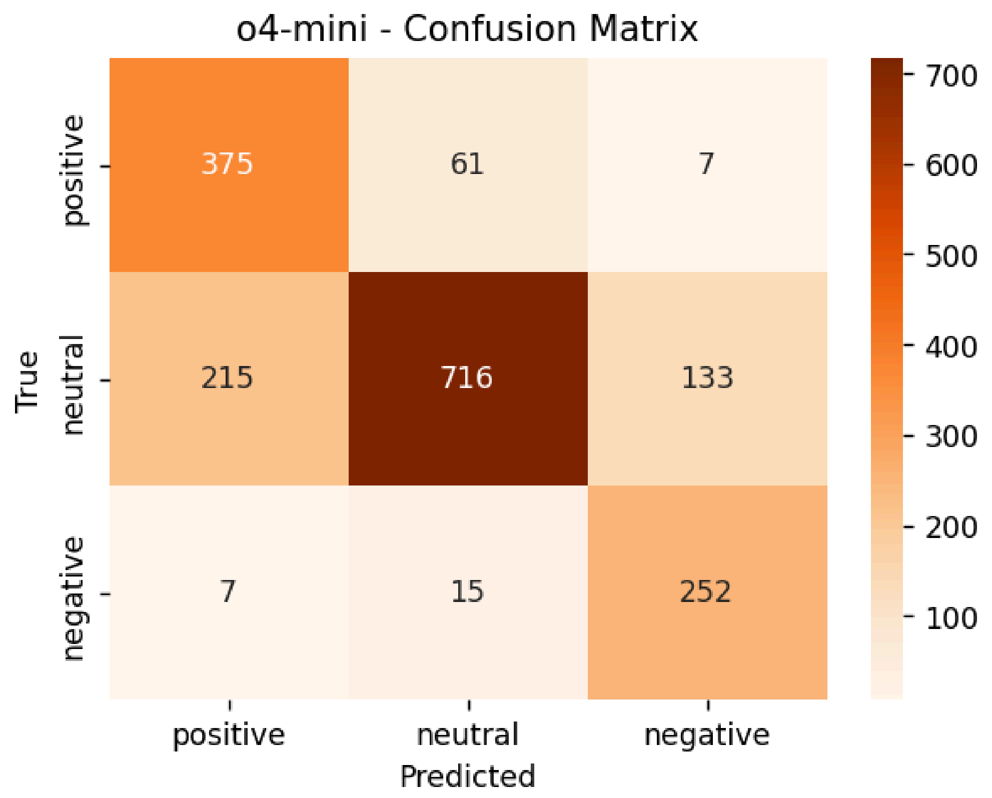


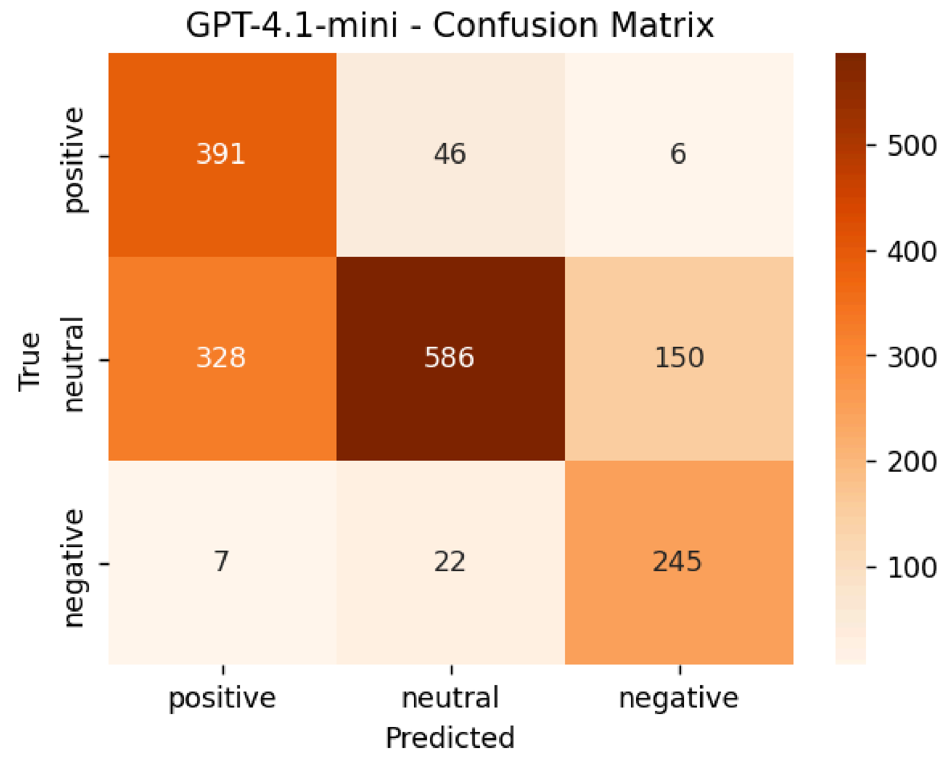


Model Evaluation – Confusion Matrices

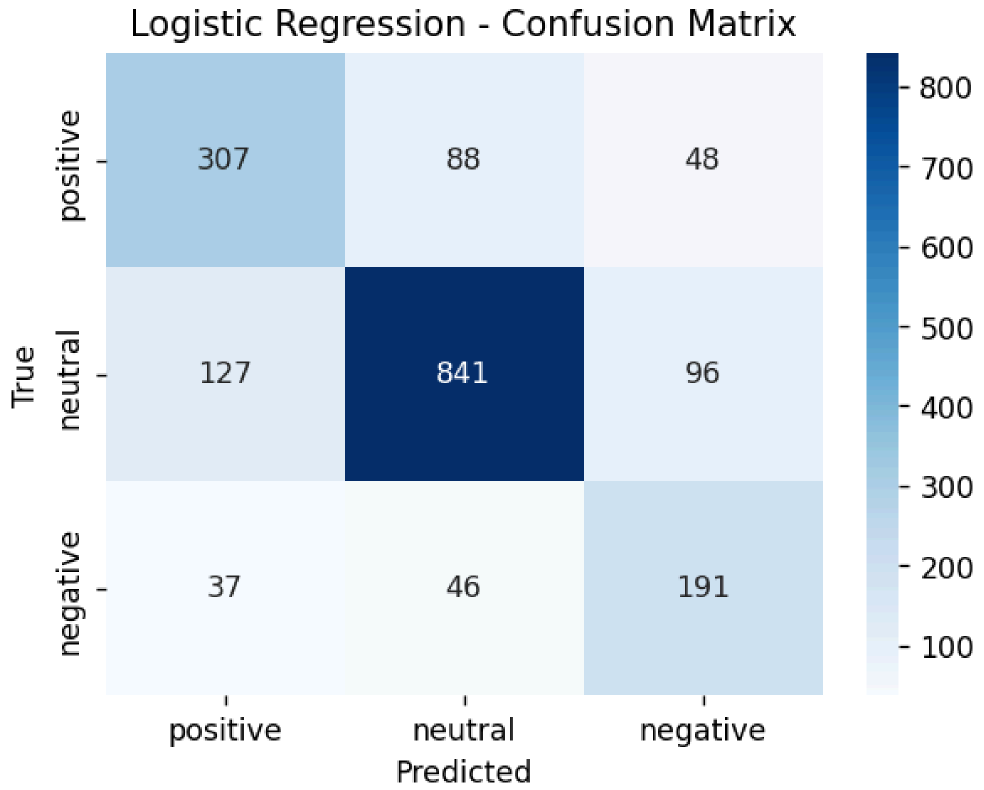
Pre-Trained

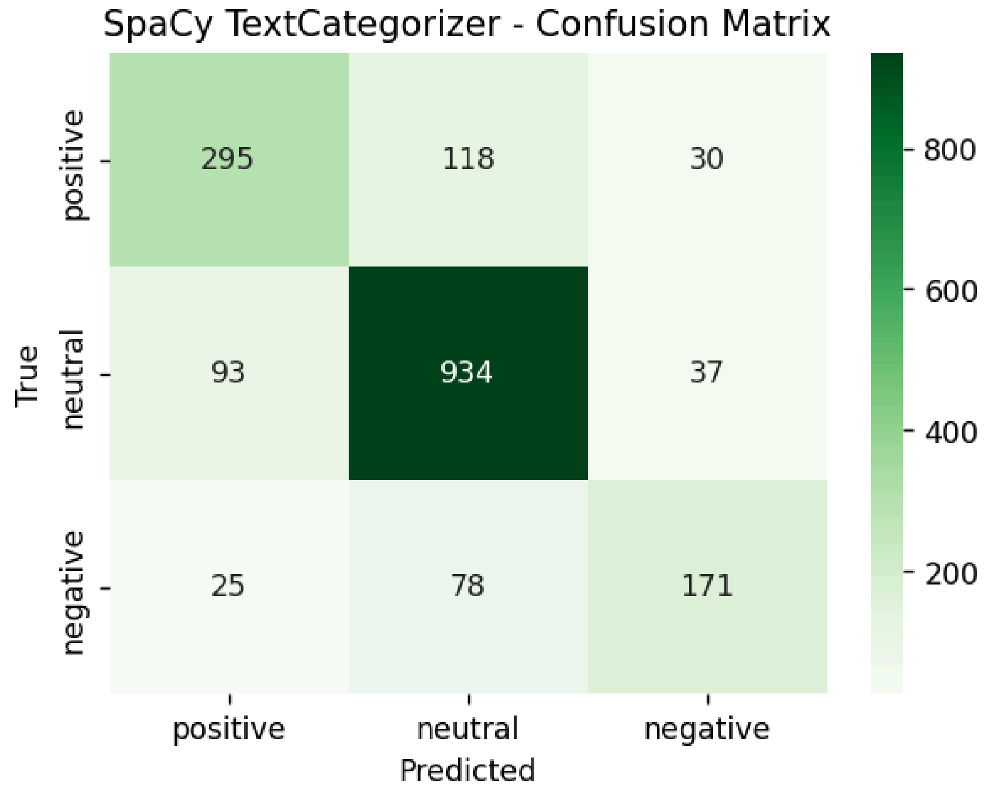


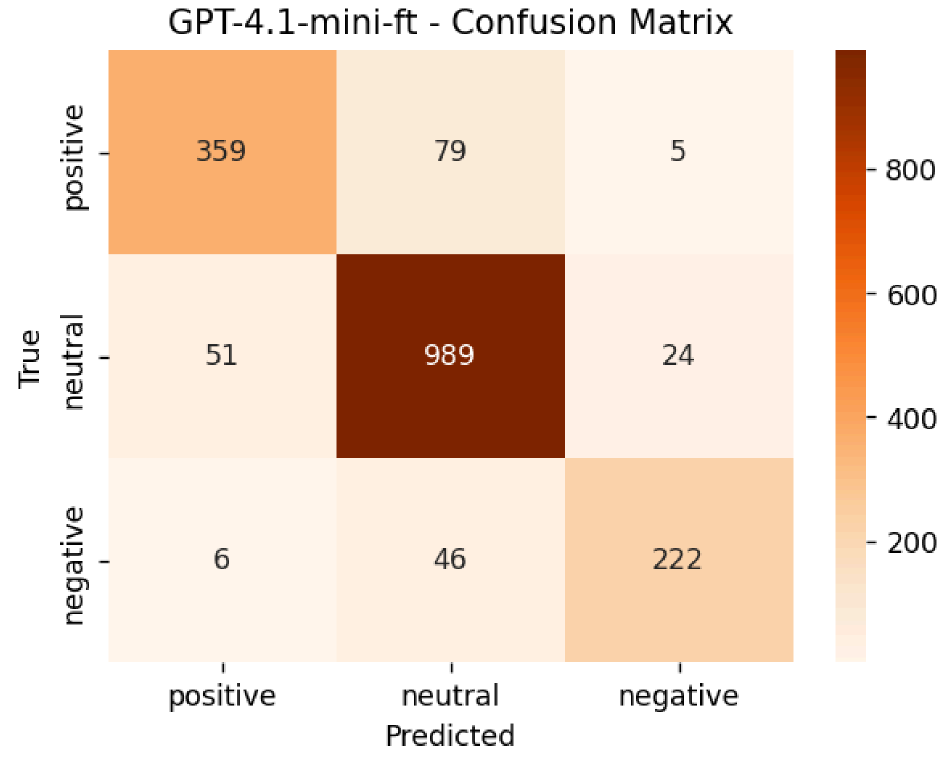




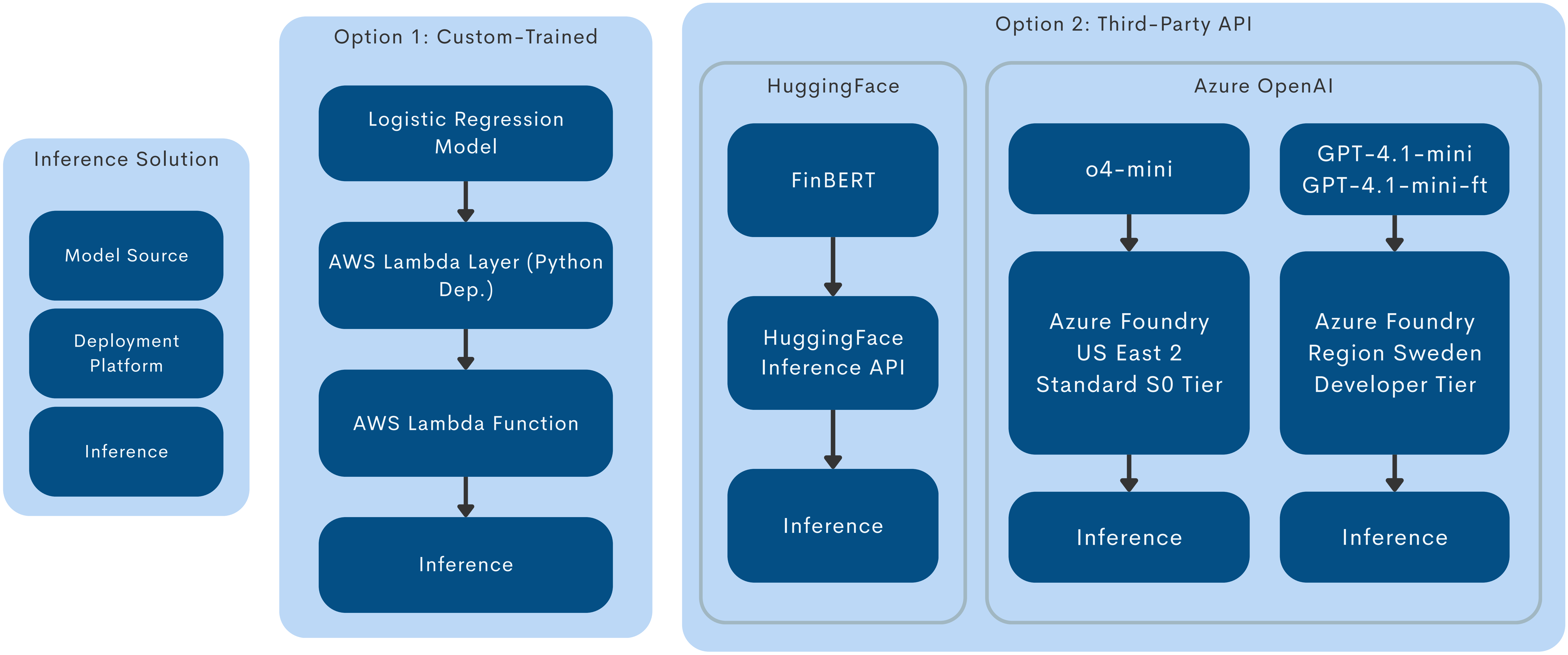
Trained



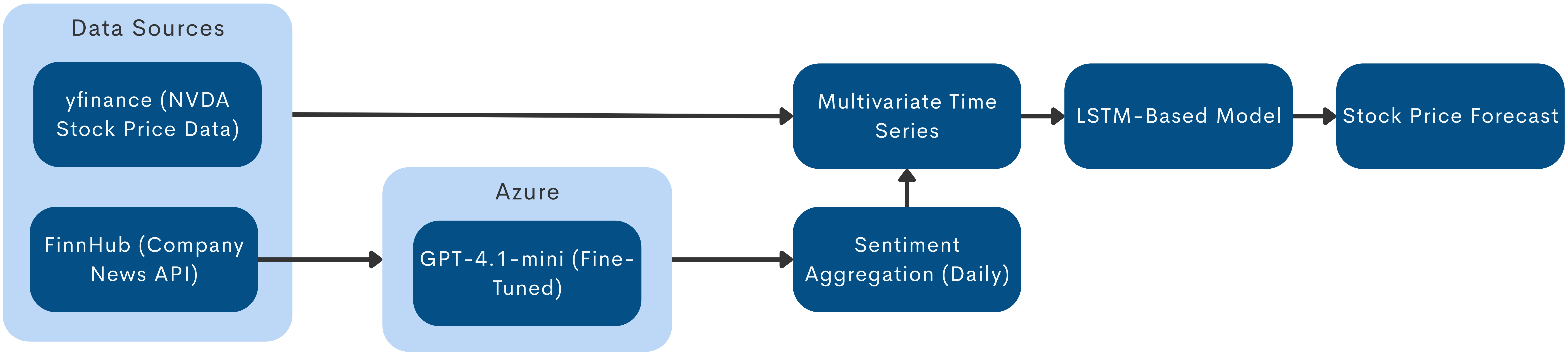




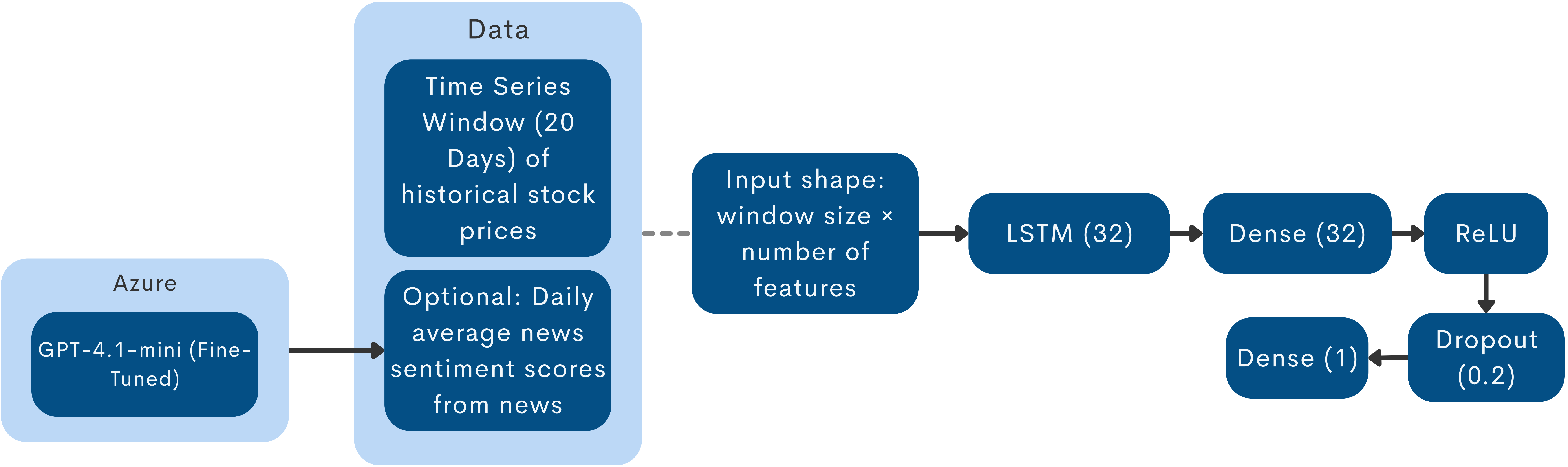
Model Deployment and Inference



Multivariate Stock Price Forecasting



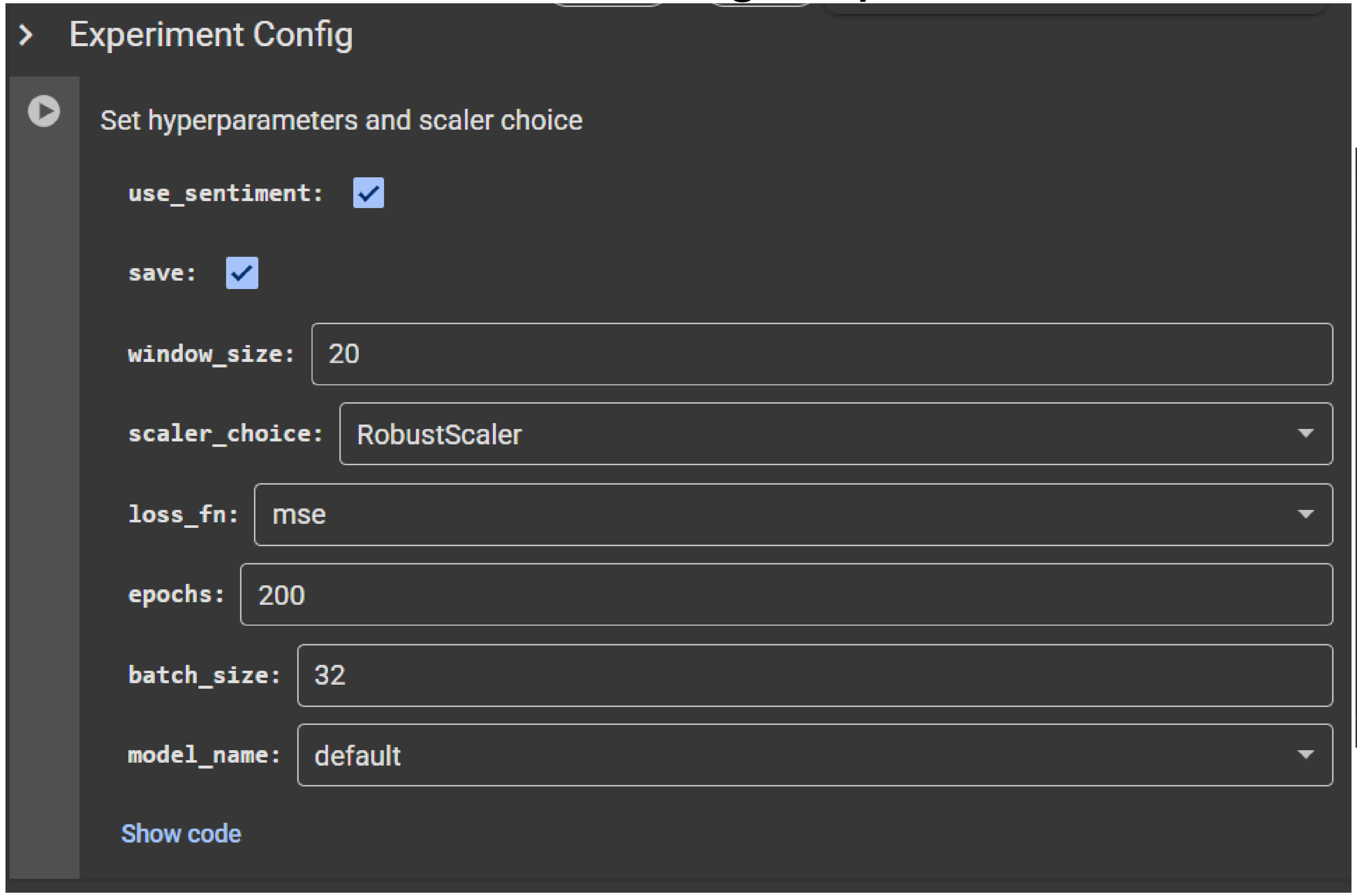
Multivariate Stock Price Forecasting

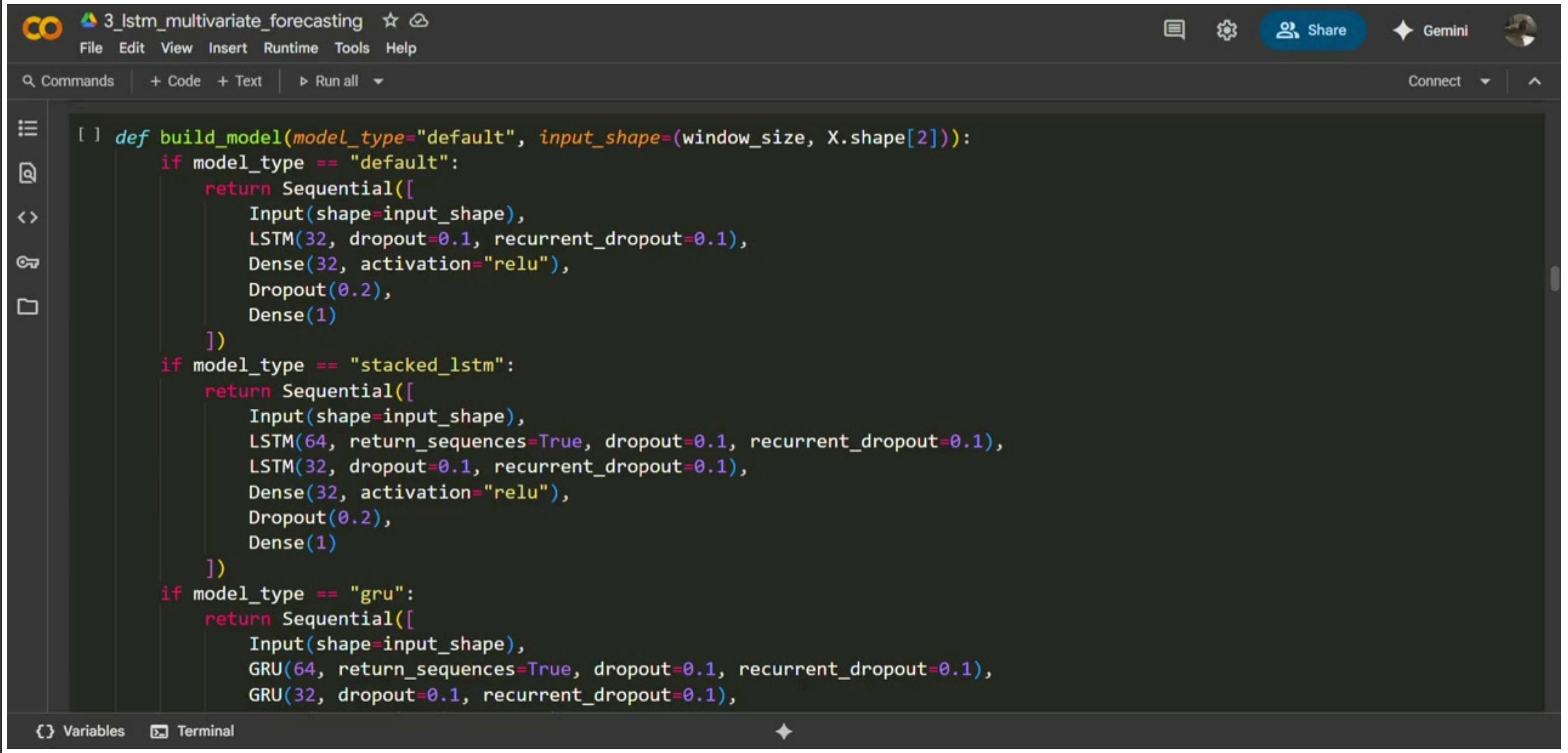


Multivariate Stock Price Forecasting

Hyperparameter Tuning

- Tested: window size, scalers, loss, batch size, epochs
- Auto-logged run metadata
- Identified best configs by SMAPE, R^2

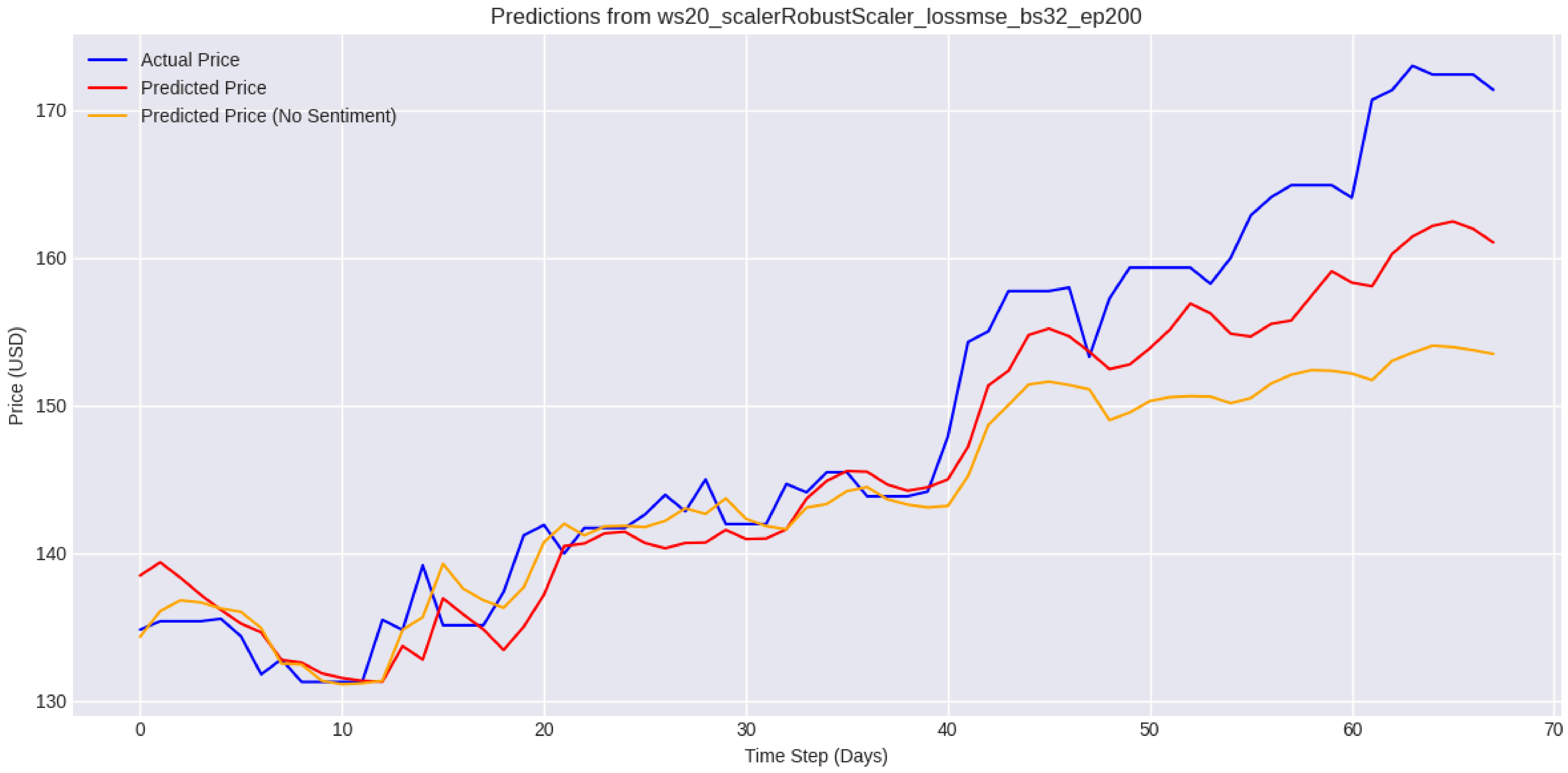




Multivariate Stock Price Forecasting – Results

Model Type	SMAPE (Non-Baseline)	SMAPE (Baseline)	ΔSMAPE
bi_lstm	1.9983	2.3994	-0.4011
gru	2.7892	2.8275	-0.0383
stacked_lstm	2.3116	3.4404	-1.1288
lstm	2.4128	3.4996	-1.0868
lstm_attention	2.4613	3.8524	-1.3911
lstm_gru	4.7335	5.2200	-0.4865
bilstm_attention	4.1467	5.4239	-1.2772
transformer	9.1348	11.0814	-1.9466

Multivariate Stock Price Forecasting – Results



Limitations & Future Work

Limitations:

- Fine-Tuning Cost Constraints
  - Reinforcement Learning Fine-Tuning (RLFT) costs ~$100 per 1M tokens—too expensive to fine-tune o4-mini.
  - Supervised Fine-Tuning (SFT) was used instead at ~$1.70 per 1M tokens.
- Azure Developer Tier Restrictions
  - Used Developer Tier for GPT-4.1-mini fine-tuning and deployment (no hourly cost).
  - Instances expire every 24 hours and require manual re-deployment.
- spaCy Model Deployment Challenges
  - spaCy's Linux wheel exceeds AWS Lambda's 250 MB size limit (313 MB).
  - Made deployment of the spaCy TextCategorizer model infeasible via Lambda.
- Stock Forecasting Imperfections
  - LSTM forecasts improved with sentiment but still lacked precision.
  - Could benefit from tuning, other architectures (like Transformers), and more diverse inputs.

Future Work:

- Incorporate RLFT & o4-mini Fine-Tuning
  - If resources permit, apply reinforcement learning to fine-tune reasoning-focused LLMs like o4-mini.
- Better Sentiment Data Sources
  - Include real-time Twitter/X or Stocktwits sentiment to enrich the news-based scores.
- Improve Forecasting Models
  - Explore hybrid or ensemble architectures; refine preprocessing; add external indicators (e.g., macroeconomic data).
- Use Retrieval-Augmented Generation (RAG)
  - Dynamically retrieve relevant articles or filings and feed into LLM prompts to improve both classification and interpretability.

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 zero-shot 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-scores across most datasets 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, deliver state-of-the-art performance in financial sentiment tasks. Additionally, their outputs offer real utility in financial time series forecasting.
Future work may extend to real-time trading signal generation, multi-lingual datasets, and integration with hybrid models such as reinforcement learning or RAG-enhanced systems.

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] 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.
[4] 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.
[5] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.
[6] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models, 2023.
[7] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models, 2023.
[8] Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Good debt or bad debt: Detecting semantic orientations in economic texts, 2013.
[9] 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.
[10] Neural Magic. Twitter financial news sentiment. http://precog.iiitd.edu.in/people/anupama, 2022.

Thank you!

Siddhant Maji
siddhant.maji@gmail.com
https://github.com/sidmaji/
https://finsa-utd.web.app/