Market Sentiment Analysis using Low-Cost Hybrid Framework

Yahya Hassan, Krishi Manek, Ayush Panda, Samyak Jain

1 Introduction

Increasingly, the use of natural language processing (NLP) methods has played a key role in financial document analysis and interpretation. Developments in event extraction, generative capabilities such as summarization and dialogue generation (Zhao et al., 2023), and tools like Retrieval-Augmented Generation (RAG) and topic modeling methods (e.g., Latent Dirichlet Allocation, LDA) have uncovered new potential for converting unstructured financial data into actionable insights. This enables better informed data-driven decision making and market intelligence.

For the scope of this project, we define LLM models as "open" and "closed." By open, the intended meaning here is that the model is "reasonably reproducible" with publicly released information as defined in (Rogers et al., 2023). The field currently faces a growing split between open and closed LLMs. Open LLM models offer distinct advantages compared to closed models. With flexibility for developers to access and modify underlying architectures, open LLM models make it possible to fine-tune the models for specific domain requirements. The case for using open LLM models in niche applications is strengthened by their resource-efficiency and lack of licensing restrictions. The availability of diverse open-source datasets and platforms that offer GPU resources on a rental basis further lowers the barrier to entry. On the other end, closed commercial LLMs are available for use through APIs. Commercial APIs come with robust support and documentation that provide a method of access to state of the art models on a per-use basis. Commercial APIs play a valuable role in the development cycle: they can serve as baselines or sources of automated feedback, reducing reliance on expensive human annotation (Lee et al., 2023). Traditional Reinforcement Learning with Human Feedback (RLHF) is typically more time and resource costly due to need for human annotators and evaluators. High-quality AI models can deliver more consistent and scalable feedback quicker on model outputs at lower-cost with mitigatable drawbacks (Bhatia et al., 2024).

One use case of the open model RoBERTa-large, is the construction of a Hawkish measure:

$$\text{Measure}_i = \frac{\#\text{Hawkish}_i - \#\text{Dovish}_i}{\#\text{Total}_i} \quad (1)$$

where Measure_i is a document-level measure of the ratio of Hawkish sentences to Dovish sentences in a document (Shah et al., 2023). The data set subjected to this measure included text data (speech transcripts, press conference transcripts, and meeting minutes) from the FOMC. Each sentence is annotated using an altered method of sentiment analysis. Traditionally, sentiment analysis classifies text into positive vs negative. This classification does not extract policy stance. Sentences with "increase" could be dovish or hawkish without a clear connotation. Instead, introducing a new task to classify sentences as hawkish vs dovish as opposed to positive vs negative suits monetary policy texts. Figure (1) verifies the usefulness and causal relationship with economic performance of such a metric.

This paper presents a novel, cost-effective, and specialized approach for sentiment analysis and volatility modeling of FOMC texts with a focus on efficiency and adaptability. Our proposed solution is a multi-tiered, hierarchical framework of smaller, fine-tuned RoBERTa-large models. Each model is dedicated to a specific sentiment or stance (e.g., hawkishness, market confidence, volatility). Unlike traditional monolithic LLM solutions, our approach leverages “open” models and utilizes them in targeted, specialized ways, to mitigate the inefficiency and high costs associated with using very large, general-purpose LLMs.

Using domain-specific labeling through a closed

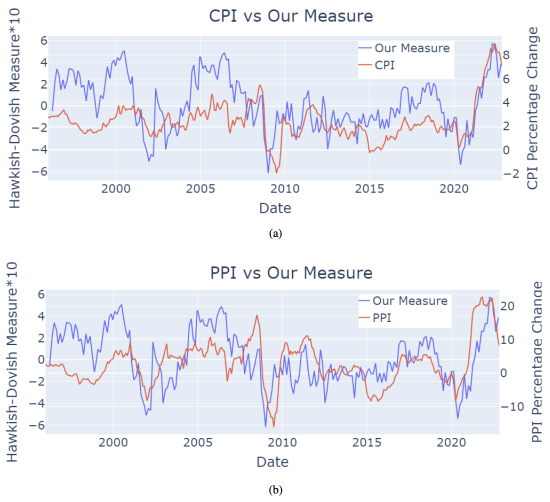

Figure 1: (a) Measure on meeting release date and 1-year change in CPI data on the first day of each month (b) Measure on meeting release date and 1-year change in PPI data on the first day of each month

commercial LLM at the zero-shot classification stage (the Gemini 1.5-flash model) we create an accurate, sentence-level labeled dataset of FOMC correspondences from 2000 to 2024. These labels are used to fine-tune multiple RoBERTa-large models. Each model is specialized in detecting distinct sentiment categories crucial to financial analysis. This process avoids the need for a large, commercial level, and "do-it-all" model. Additionally, using a closed model to label data instead of professional labelling of data decreases cost and supports efficient scaling to additional sentiment classes if necessary. Once trained using closed commercial models, these smaller specialized models can be rapidly deployed, reused, and integrated to capture nuanced sentiment dynamics far beyond simple positive/negative polarity. Thus, striking a balance between the use of closed and open models.

Our approach creates sentiment measures at a more granular, sentence-level scale, and aggregates results into interpretable metrics (e.g., a "hawkishness measure"), that can directly connect textual sentiment patterns to market indicators.

2 Proposed Research

The proposed research involves developing a novel sentiment analysis framework that utilizes multiple specialized hierarchical arranged LLMs. Each model is fine-tuned to d

2.1 Data Collection and Labeling

We utilize FOMC texts (speech transcripts, press conference transcripts, and meeting minutes) from 2000 to 2024. Instead of relying on costly, time-consuming expert labeling, we leverage a closed commercial model (Gemini 1.5-flash) for zero-shot sentiment classification. Sentences are classified into granular categories such as "hawkish," "dovish," "volatility," and "market confidence." By labeling at the sentence-level, errors from automated labeling are averaged out, resulting in a sufficiently accurate and scalable dataset.

2.2 Specialized Fine-Tuning

Using open LLM architectures like RoBERTa-large, we fine-tune separate models for each sentiment category. For example, one RoBERTa-large model is fine-tuned to detect hawkish sentiment, another to detect dovish sentiment, and another to assess signals of volatility or market confidence. This approach avoids developing a single large model trained on a broad and ambiguous set of tasks. It offers flexibility: if a new sentiment (e.g., "bullishness" or "bearishness") becomes relevant, another RoBERTa-large model can be fine-tuned using the same Gemini-powered labeling pipeline.

Secondly, this allows for a more nuanced and context-aware understanding of sentiment by considering combinations of sentiments rather than evaluating them in isolation. By integrating various sentiment analyses at different levels, the model can capture complex emotional dynamics and provide more accurate and comprehensive sentiment interpretations. Different transformer architectures and attention mechanisms are more well-suited for specific use cases. Effectively, this new framework relieves developers from having to develop a one size fits all model.

3 Implementation Details

Figure 2 illustrates the overall pipeline for generating sentiment-specific datasets, training specialized RoBERTa models, and applying them to FOMC data. Each phase of the process is designed to minimize manual labelling, ensure scalability, and maintain accuracy in sentiment classification tasks for monetary policy communications.

1. **Data Acquisition:** The starting point is a compiled dataset of FOMC communications, including meeting minutes, speeches, and press

Figure 2: Workflow Diagram: From Zero-Shot Labeling to Specialized RoBERTa Sentiment Models and Downstream Label Prediction Tasks

conference transcripts spanning from 2000 to 2024. This unified corpus will serve as both training and testing material for subsequent sentiment analysis tasks.

2. **Zero-Shot Sentiment Labeling:** The entire FOMC dataset is passed through a commercial LLM service (Gemini-1.5-flash) configured in a zero-shot mode. Without any prior fine-tuning, Gemini assigns preliminary sentiment labels at the sentence level. Using a model trained on extensive and diverse data, this step produces annotated datasets.

3. **Creation of Specialized Datasets:** Using Gemini's initial labels, the corpus is divided into three distinct one-hot encoded labeled datasets, each corresponding to a critical dimension of monetary policy sentiment:

   - *"Volatility" Dataset:* Sentences indicative of market volatility sentiment.
   - *"Market Confidence" Dataset:* Sentences reflecting confidence levels in market stability and growth.
   - *"Hawkish vs. Dovish" Dataset:* Sentences classified as either "hawkish" (favoring tighter, more restrictive monetary policy) or "dovish" (favoring looser, more accommodative monetary policy).

4. **Model Specialization with RoBERTa:** For each of the three sentiment dimensions, a separate RoBERTa-large model is fine-tuned:

   - *"Volatility" RoBERTa:* Trained exclusively on the Volatility Dataset, this model learns to identify sentences related to increased or decreased volatility.
   - *"Market Confidence" RoBERTa:* Fine-tuned on the Market Confidence Dataset, to distinguish between sentiments expressing high or low confidence in market conditions.
   - *"Hawkish vs. Dovish" RoBERTa:* Focused on the Hawkish vs. Dovish Dataset, the model is specialized on monetary policy stance.

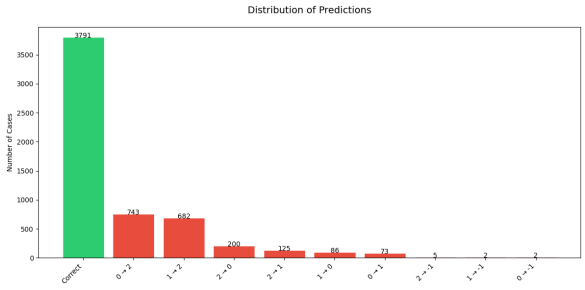

Figure 3: Predictions Distribution

By training dedicated models for each sentiment dimension, we achieve higher accuracy and faster inference than a single, general-purpose model.

5. **Prediction and Analysis:** Once fine-tuned, the three specialized RoBERTa models are deployed to classify new FOMC texts. For each incoming document sentences are fed into the respective specialized models to predict volatility, market confidence, and hawkish/dovish sentiments.

   The resulting predictions can then be aggregated and analyzed to construct metrics like the Hawkishness Measure, evaluate the level of market confidence, or assess volatility risks. These quantitative measures enable scholars and practitioners to correlate language from policy communications with subsequent market movements, economic indicators, or investor sentiment.

This structured, step-by-step approach ensures transparency, reproducibility, and adaptability, making it possible to continuously refine sentiment classification models and align them closely with evolving policy communication trends and market responses.

4 Performance Analysis on Test Data

Upon running our model on the test data set, we were able to predict hawkish/ dovish sentiments from FOMC meeting minutes with 66.39%. We also plotted the distribution of predictions (correct and incorrect mappings) as a bar chart to visualize which mistakes the model was making the most. This is shown in figure 3. The prediction distribution chart shows that while 3791 of the total 5710 predictions were correct, the model's most common mistakes occurred when classifying hawkish or dovish sentiments as neutral (743 and 682 times

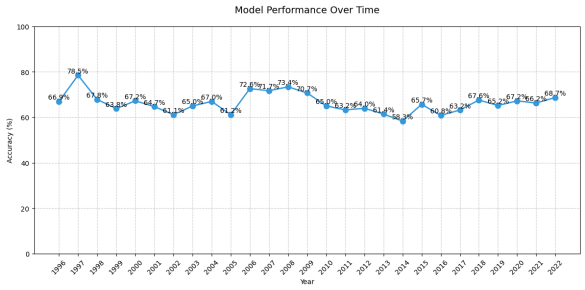

Figure 4: Performance Over Time

Class	Precision	Recall	F1-Score	Support
0	0.71	0.46	0.55	1504
1	0.77	0.46	0.58	1432
2	0.63	0.88	0.74	2773
Accuracy			0.66	5709
Macro Avg	0.53	0.45	0.47	5709
Weighted Avg	0.69	0.66	0.65	5709

Table 1: Detailed Classification Report

respectively). We also visualized the model's accuracy on the meeting minutes each year as shown in figure 4 which shows a relatively high performance between 60 to 80% year on year.

The performance of our model is summarized in the classification table (Table 1) below:
Furthermore, we used the three predicted sentiments from our model (hawkish/ dovish, stability/ volatility, and trust/ mistrust) to create a custom economic measure that mirrors the trend in 10-year real interest rate levels from 2000 to 2024. We obtained this data from the Federal Reserve Bank of St. Louis' website.
We calculated our measure as denoted by the below formula:

$$\text{measure}_i = \frac{\text{hawkish\_weight} \cdot \text{hawkish}_i}{\text{total}_i} + \frac{\text{trust\_weight} \cdot \text{trust}_i}{\text{total}_i} - \frac{\text{volatile\_weight} \cdot \text{volatile}_i}{\text{total}_i}$$

, wherein we started with initial weights of [1.0, 1.0, 1.0] and used a root mean squared objective function to be minimized to find the final weights. We found the optimized weights to be [1.31169443, 4.03913879, 0.] and plotted the measure with these weights on the same graph as the 10-year real interest rate data from 2000 to 2024 to obtain the graph shown in figure 2.

To further analyze the performance of our economic measure, we calculated its root mean squared error and mean absolute error values to obtain values of 1.184 and 0.939 respectively which are indicative of a very close correlation between

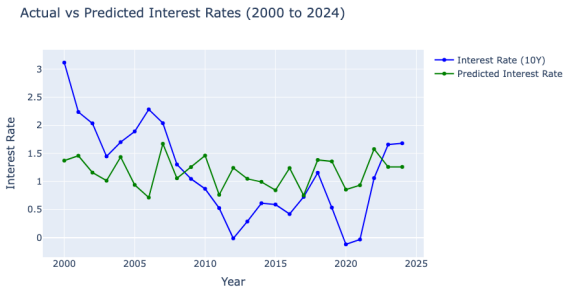

Figure 5: Our measure vs. 10-Yr Real Interest Rates (2000-2024)

our measure and the interest rates.

Lastly, we also performed residual analysis and found the mean residual value to be 0.722 which is very close to 0, indicating that the model has very low bias.

References

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A family of gpt-4 level multimodal financial large language models.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Anna Rogers, Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Luccioni, Maarten Sap, Roy Schwartz, Noah A. Smith, and Emma Strubell. 2023. Closed ai models make bad baselines.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task market analysis.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Contribution Matrix

Section/Task	Yahya	Krishi	Ayush	Samyak
Writing: Literature Review and Introduction	✓			
Writing: Propose Research	✓			
Writing: Implementation Details	✓			
Writing: Performance Analysis on Test Data		✓	✓	✓
Methodology Design (Hierarchy of Specialized LLMs)	✓			
Data Collection (FOMC Texts: 2000–2024)	✓			
Zero-Shot Labeling (Gemini 1.5-flash Integration)	✓	✓		
Dataset Creation (Volatility, Market Confidence, etc.)	✓	✓		
Model Fine-Tuning - Volatility (RoBERTa-large)	✓			
Model Fine-Tuning - Stability (RoBERTa-large)	✓			
Model Fine-Tuning - Trust (RoBERTa-large)	✓			
Baseline GPT-4o Labelling			✓	✓
Evaluation and Validation		✓		✓
Performance Analysis		✓	✓	✓
Economic Measure Calculation	✓	✓		✓
Residual Analysis and Correlation with Interest Rates		✓	✓	✓