LLM を用いたタカハトセンチメント付与タスクの検証

川原一修 [1]
[1]Japan Digital Design 株式会社
{takanobu.kawahara}@japan-d2.com

概要

中央銀行の要人発言にタカハトセンチメントを付与するタスクは, 金融実務家にとって重要でありながら, 商用利用可能なアノテーション付き学習データの提供はまだ行われていない. 本研究では, プロンプトエンジニアリングを用いて, 大規模言語モデル（LLM）を活用したセンチメント推定の精度向上を検証した. その結果, センチメント付与ガイドラインを含むプロンプトを用いることで, ファインチューニングされた小型モデルよりも高い精度を達成するケースが存在することが明らかになった.

1 はじめに

米国の連邦準備制度（Federal Reserve System, 以下「Fed」と略称）は, 2022 年 3 月に長らく 0%付近で維持されていた政策金利を引き上げ,2023 年 7 月までの約 1 年半で合計 5%の引き上げを行った. これに伴い, 米国の 10 年国債の指標金利も約半年間で 2%上昇し, FTSE 世界国債インデックスは最大 19%,S&P500 種指数は最大 17%下落し[1)]投資家に大きな影響を及ぼした. 中央銀行の金融政策が市場与える影響は大きく, その動向を分析することは市場参加者にとって重要な課題である. Fed 高官のスピーチや公開市場委員会（Federal Open Market Committee, 以下「FOMC」と略称）の議事録は, 政策決定者からの直接的なメッセージであり, 金融政策や経済環境に対する彼らの見解が反映されている. これらを詳細に分析することで Fed の政策がタカ派（緊縮的）か, ハト派（緩和的）かを判断することの価値は大きい.

自然言語処理技術を用いた Fed のテキストデータ分析は多く行われており, 例えば [1] や [2] は, LDA を用いたトピック抽出と極性辞書によるセンチメント分析を組み合わせ, 資産価格への影響を分析している. しかし,[3] が指摘するように Fed 要人の発言には婉曲表現が多く, テキスト外の情報を参照しないと意味が理解しにくいケースも多い. Fed 高官の難解な文章をさす Fedspeak[4] という表現すら存在する. 難解な Fedspeak をルールベースの極性推定だけで読み解くことは難しい可能性があり、実際 [5] や [6] は, 独自にタカ, ハトのアノテーションを付与したデータセットを用いて BERT や RoBERTa などの Transformer ベースのモデルでファインチューニングを行い, タカハトセンチメントの付与タスクにおいてルールベースのモデルを上回る精度を達成したと報告している. ただし, 前者のデータセットは公開されておらず, 後者は商用利用が制限されている. 金融実務家にとってタカハトセンチメントの付与は重要なタスクであるが, 現在利用できるアノテーション付きの学習データセットは限られている. ChatGPT のような大規模言語モデル（LLM）がファインチューニングなしで従来の自然言語処理タスクを高精度で処理できるという報告が増えている中 [7][8][9], プロンプトに様々な工夫を加えることで, 従来のアプローチでは解決が難しい問題に対しても新たな可能性が開ける. 金融政策に関連するテキストデータの分析においても LLM が優れた解釈能力を発揮し, より複雑なニュアンスやコンテキストを理解することが期待される. 上記のような背景から本研究では LLM を利用してタカハトセンチメントを効果的に識別する方法を検証する.

2 先行研究

言語モデルをタカハトセンチメントの付与タスクに直接適用した研究として [5][6] の他に [10] がある. [10] は金融領域に特化した LLM の開発に取り組み, LLaMa-65B[11] に独自データを用いてインストラクションファインチューニングを行ったところ, タカハトセンチメント付与タスクにおいてもファインチューニング前後で精度が向上したことを報告している.3 研究ともモデルのファインチューニングに焦点を当てている点が本稿とは異なる.

1） Bloomberg による

タカハトセンチメント付与タスクに対してプロンプトエンジニアリングに焦点をあてて検証した研究は見つからなかったが, プロンプトエンジニアリング自体には多くの先行研究があり, 例えばタスクに関連する情報をプロンプト内にとりこむインコンテキストラーニング [8][12] の報告事例は多い. 中でも [8] は, ChatGPT に Few-shot プロンプト [13] を活用して Zero-Shot では精度の上がりにくいタスクに対しても精度の改善が見られたことを報告している. また [14] は LLM が与えられたコンテキストを活用せず事前学習で得られた知識を優先するケースが有ることに注目し, Opnion プロンプトや automatic prompt engineering[15] を利用して作成した Instruction プロンプトを利用することでこの傾向が改善することを確認している.

本研究では, これらの先行研究を踏まえ, まず Few-Shot プロンプティングのタカハトセンチメント付与タスクへの有効性を検証する. 加えて, 難解な Fedspeak を Few-shot の事例のみから一般化して解釈することが難しい可能性に考慮し, センチメント付与の指針（ガイドライン）をプロンプト内に含め, センチメント付与時にガイドラインに従うよう指示したプロンプトの効果を検証する.

3 使用するデータセットとモデル

3.1 FOMC データセット

[5] で作成されたデータセットを著者らの Git Hub レポジトリ[2] から取得して使用する. データセットは Fed のホームページ[3] から FOMC 議事録 (MM), 記者会見のトランスクリプト (PC), Fed 高官のスピーチ (SP) の 3 種類のテキストデータを取得して作成されている. また,3 種類のデータセットについてそれぞれ 3 つのテストデータセットと訓練データセットの分割が準備されており, 本稿の検証ではそれらを使用した.

3.2 タカハトセンチメントインデックス

タカハトセンチメントインデックスの作成に Fed のホームページから FOMC 声明文を取得して利用した. 取得した声明文のテキストファイル数は 219 ファイルで総センテンス数は 9,591 センテンスあった. 2000 年 2 月から 2023 年 12 月までの声明文が含まれている. 取得したセンテンスから金融政策の先行きに関連する可能性の高いセンテンスを抽出するため [16] の単語リストを利用し, ルールベースでセンテンスの抽出を行った. 具体的には当該論文の Table2 の PanelA1 または PanelB1 に含まれる単語がセンテンス中に存在する, かつ PanelA2 または PanelB2 に含まれる単語が存在する, センテンスのみを利用し, 最終的に 1,942 センテンスを抽出した.

3.3 利用したモデル

メインの LLM として OpenAI 社の ChatGPT [4] を利用する, 具体的には次の 3 モデルを利用する, gpt-3.5-turbo-0613(GPT3.5-Turbo), gpt-4-0613(GPT4), gpt-4-1106-preview(GPT4-Turbo). また GPT 以外のモデルでもガイドライン付きプロンプトの効果があるか検証するため, Meta 社の Llama-2-7B/13B[17] と金融に特化した LLM である Finma-7B[18] も利用する. タカハトセンチメントインデックスの検証ではベンチマークとして FomcRoberta[5] を利用する. GPT 以外のモデルはすべて Hugging Face[5] から取得して利用した.

4 実験設定

4.1 Few-shot

判定対象のセンテンス類似したセンテンスを訓練データから取得し, プロンプトの中にコンテキストとして追加することで精度が向上するか検証した. テキストの類似度は sentence-transformers/all-MiniLM-L6-v2 を使用し得られたベクトルの cosine 類似度用いた。追加するセンテンスの数は 1,3,10 の 3 パターン, 使用する訓練データについても全量を使用する場合,50%,10%とサンプリングして使用する場合の 3 パターンで精度に差が出るか確認した. 本実験のモデルには GPT3.5-Turbo を利用した. プロンプトは Appendix に記載した Base プロンプトに例示用のセンテンスを類似度の高い順に追加し作成した. また一部訓練データとテストデータで重複するセンテンスが存在したため, 訓練データからそれらのセンテンスを削除した上で利用した.

2） https://github.com/gtfintechlab/fomc-hawkish-dovish
3） http://www.federalreserve.gov/
4） https://platform.openai.com/docs/models/continuous-model-upgrades
5） https://huggingface.co/

4.2 ガイドライン付きプロンプト

[5] に記載されているアノテーター用のアノテーションガイドラインを LLM 判定のガイドラインとして取り入れたプロンプトを作成して利用した. ベースラインのプロンプトとして [5] で使用されたプロンプトを利用した. 実際に使用したプロンプトは Appendix に記載する.

4.3 タカハトセンチメントインデックス

節 3.2 で作成したデータセットについて,GPT4 とガイドライン付きプロンプトを利用してセンテンス毎にタカハトセンチメントを判定し, 声明文毎にアグリゲートした上で過去 3 回の声明文のスコアの平均をとりタカハトセンチメントセンチメント指数を作成する.

$$HawkDovScore_i = \frac{\#Hawkish_i - \#Dovish_i}{\#Total_i}$$

$\#Hwakish_i$ は声明文 i 中で Hawkish と判定されたセンテンスの数,$\#Dovish_i$ は Dovish と判定されたセンテンスの数,$\#Total_i$ は声明文 i 中から抽出されたすべてのセンテンス数である.

5 実験結果

5.1 Few-shot

紙面の都合上, 図 1 に訓練データを 10%サンプルして利用した結果のみ報告する. 記者会見については Zero-Shot のケースと比べて最大で F1 値が 3%ポイント程度向上し, Few-shot を活用する効果が確認できた. しかし, スピーチデータについてはむしろ精度が低下していることに加えて, 議事録データについても F1 値の改善幅は最大で 0.7%ポイント程度であり, 改善幅は限定的であった. 全体としてタカハトセンチメントの付与タスクに関しては Few-shot プロンプトの効果は限定的であった.

5.2 ガイドライン付きプロンプト

表 1 に検証結果を記載する. 左 3 列はベースプロンプト, 右 3 列はガイドライン付きプロンプトを使用した結果である. 最下段の RoBERTa-large は Shah[5] らに報告されている最も精度の高い, ファインチューニングされた小型モデルの検証結果を転載している. 数値はすべてテストセットにおける weighted-F1 であり,3 シードの平均を報告している,

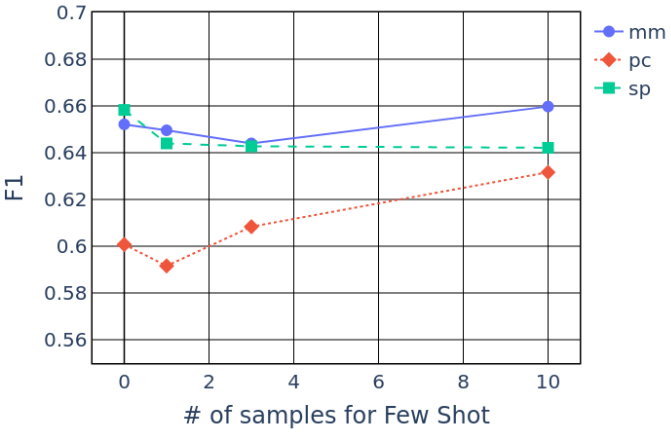

図 1 トレインデータから 10%サンプル

括弧内は標準偏差である.
以下の 3 点が特に重要だと考える.

- ガイドライン付きプロンプトでは 3 データセット中 2 データセットで,GPT4 と GPT4-Turbo の精度が RoBERTa-large を上回っておりタカハトセンチメントの付与タスクについては, プロンプトの工夫次第で LLM による Zero-Shot の推論がファインチューニングされた小型モデルを上回るケースが有ることが示された.
- モデル毎にガイドライン付きプロンプトの効果について確認すると,LLama2-13B, GPT4, GPT4-Turbo で精度が高まっている.LLama2-7B, Finma-7B については精度が低下しており, 比較的小型な LLM では長いプロンプトを有効活用できていない可能性がある. GPT3.5-Turbo についても精度が低下しており, これについては今後追加の検証が必要となる.
- Llama2-7B/13B と Finma-7B についてベースプロンプトの結果を比較すると Finma が自身よりサイズの大きいモデルに精度で上回っており, 金融に特化した知識の取得が本タスクに効果がある可能性がしめされた.

5.3 タカハトセンチメントインデックス

図 2 に作成したタカハトセンチメントインデックスの推移を表示する. 実線が GPT4, 破線が FomcRoberta よるセンチメントインデックスの推移, 細線は Federal Funds Effective Rate(以降 FF 金利と表

表 1　モデル, プロンプト毎の精度比較

	Base			Guideline		
	MM-S	PC-S	SP-S	MM-S	PC-S	SP-S
Llama2-7B	0.310(0.006)	0.281(0.105)	0.458(0.028)	0.380(0.012)	0.235(0.111)	0.356(0.025)
Finma-7B	0.451(0.043)	0.367(0.090)	0.570(0.059)	0.282(0.015)	0.268(0.095)	0.407(0.035)
Llama2-13B	0.435(0.011)	0.325(0.119)	0.524(0.057)	0.468(0.059)	0.452(0.182)	0.558(0.034)
GPT3.5-Turbo	0.652(0.039)	0.601(0.138)	0.658(0.026)	0.607(0.030)	0.550(0.119)	0.639(0.051)
GPT4-Turbo	0.659(0.015)	**0.674(0.074)**	0.656(0.018)	0.696(0.021)	**0.720(0.015)**	0.709(0.020)
GPT4	0.692(0.006)	0.670(0.085)	0.698(0.036)	0.706(0.015)	0.688(0.045)	**0.731(0.032)**
RoBERTa-large	**0.715(0.014)**	0.535(0.058)	**0.705(0.030)**	**0.715(0.014)**	0.535(0.058)	0.705(0.030)

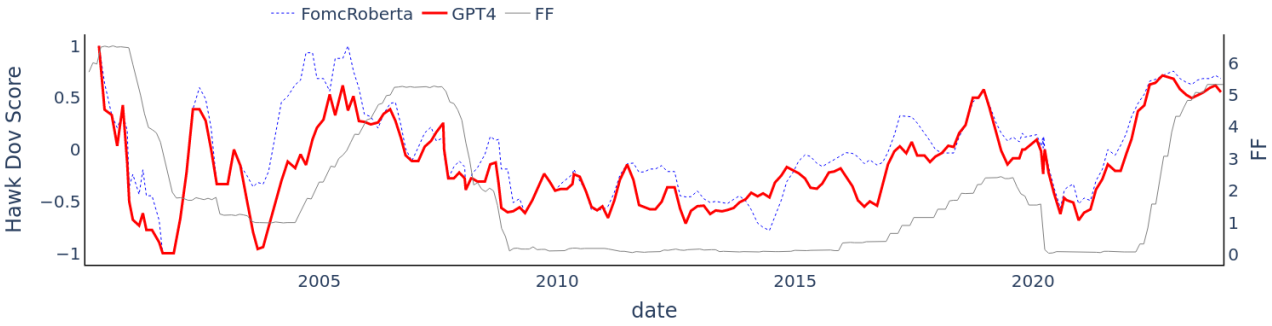

図 2　タカハトセンチメントインデックス

記) の月次データ[6)]の推移である. GPT4,FomcRoberta によるタカハトセンチメントインデックスともに FF 金利と概ね連動した動きをしかつ先行して推移している. 図 3 に FF 金利とタカハトセンチメントインデックスのラグ付き相関を計測した図を表示した. 相関が最大となるのは FomcRoberta でラグ 8 ヶ月で相関が 0.69, GPT4 で 6 ヶ月で 0.67 であり両者とも FF 金利に先行している. Fed が実際に FF 金利を引き上げる前に声明文のトーンに変化が出てきている点については, 市場との対話を重視し急激な態度の変更を避けるという観点で Fed の行動指針と合致しているといえる. これらの点から, 作成したタカハトセンチメントインデックスが声明文のトーンの変化を捉えているといえると考えた.

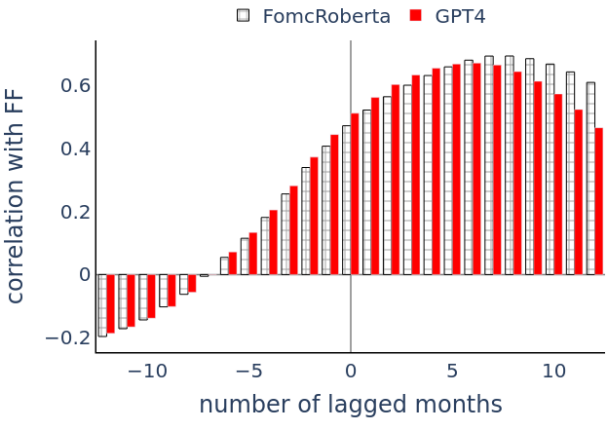

図 3　タカハトセンチメントインデックスと FF 金利のラグ付き相関

6　結言

本研究ではプロンプトエンジニアリングを通じて, 大規模言語モデル (LLM) によるタカハトセンチメントの推定精度を高める方法を検証した. 結果としてタカハトセンチメントの付与ガイドラインをプロンプトに加えたガイドライン付きプロンプトを利用することで,LLM がファインチューニングされた小型モデルを上回るケースが有ることを示した.

6)　FRED(https://fred.stlouisfed.org/) から取得した

参考文献

[1] Stephen Hansen and Michael McMahon. Shocking language: Understanding the macroeconomic effects of central bank communication. **Journal of International Economics**, Vol. 99, pp. S114–S133, 2016.

[2] Narasimhan Jegadeesh and Di Wu. Deciphering fedspeak: The information content of fomc meetings. **Monetary Economics: Central Banks–Policies & Impacts eJournal**, 2017.

[3] 高野海斗, 内藤麻人, 長谷川直弘, 中川慧. 中央銀行の要人発言に対するタカ・ハト極性付与タスクの検討. 言語処理学会 第 29 回年次大会 発表論文集 (2023 年 3 月), 2023.

[4] Alan S Blinder. **How do central banks talk?** Centre for Economic Policy Research, 2001.

[5] Agam Shah, Suvan Paturi, and Sudheer Chava. Trillion dollar words: A new financial dataset, task & market analysis. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)**, 2023.

[6] Anne Lundgaard Hansen and Sophia Kazinnik. Can chatgpt decipher fedspeak? **Available at SSRN**, 2023.

[7] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond, 2023.

[8] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check. **arXiv preprint arXiv:2305.15005**, 2023.

[9] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. **arXiv preprint arXiv:2305.16938**, 2023.

[10] Yi Yang, Yixuan Tang, and Kar Yan Tam. Investlm: A large language model for investment using financial domain instruction tuning. **arXiv preprint arXiv:2309.13064**, 2023.

[11] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[12] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. In **Proceedings of the Fourth ACM International Conference on AI in Finance**, pp. 349–356, 2023.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[14] Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. **arXiv preprint arXiv:2303.11315**, 2023.

[15] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. **arXiv preprint arXiv:2211.01910**, 2022.

[16] Federico M Ferrara, Donato Masciandaro, Manuela Moschella, and Davide Romelli. Political voice on monetary policy: Evidence from the parliamentary hearings of the european central bank. **European Journal of Political Economy**, Vol. 74, p. 102143, 2022.

[17] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv e-prints**, pp. arXiv–2307, 2023.

[18] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. **arXiv preprint arXiv:2306.05443**, 2023.

A 実験に使用したプロンプト

実験に使用したプロンプトを以下に示す.

Base Prompt

Discard all the previous instructions. Behave like you are an expert sentence classifier. Classify the following sentence from FOMC into 'HAWKISH', 'DOVISH', or 'NEUTRAL' class. Label 'HAWKISH' if it is corresponding to tightening of the monetary policy, 'DOVISH' if it is corresponding to easing of the monetary policy, or 'NEUTRAL' if the stance is neutral. Provide the label in the first line and provide a short explanation in the second line. The sentence:

Guideline Prompt

Discard all the previous instructions. Behave like you are an expert sentence classifier. Classify the following sentence from FOMC into 'HAWKISH', 'DOVISH', or 'NEUTRAL' class. The table below is the annotating guideline, refer to this guideline while labeling the sentence.

Category	**Dovish**	**Hawkish**	**Neutral**
Economic Status	when inflation decreases, when unemployment increases, when economic growth is projected as low	when inflation increases, when unemployment decreases when economic growth is projected high when economic output is higher than potential supply/actual output when economic slack falls	When unemployment rate or growth is unchanged, maintained, or sustained
Dollar Value Change	when the dollar appreciates	when the dollar depreciates	N/A
Energy/House Prices	when oil/energy prices decrease, when house prices decrease	when oil/energy prices increase, when house prices increase	N/A
Foreign Nations	when the US trade deficit decreases	when the US trade deficit increases	when relating to a foreign nation's economic or trade policy
Fed Expectations/Actions/Assets	Fed expects subpar inflation, Fed expecting disinflation, narrowing spreads of treasury bonds, decreases in treasury security yields, and reduction of bank reserves	Fed expects high inflation, widening spreads of treasury bonds, increase in treasury security yields, increase in TIPS value, increase bank reserves	N/A
Money Supply	money supply is low, M2 increases, increased demand for loans	money supply is high, increased demand for goods, low demand for loans	N/A
Key Words/Phrases	when the stance is "accommodative", indicating a focus on "maximum employment" and "price stability"	indicating a focus on "price stability" and "sustained growth"	use of phrases "mixed", "moderate", "affirmed"
Labor	when productivity increases	when productivity decreases	N/A

Provide the label in the first line and provide a short explanation in the second line. The sentence:

ガイドライン付きプロンプトの表は実際には Markdown 方式でプロンプトに記載したがここでは視認性を向上させるために表形式で記載している.