

文章の長さが金融センチメントに与える影響の分析

高野 海斗¹ 平松 賢士²

¹ 野村アセットマネジメント株式会社 ² 株式会社アイフィスジャパン
takaito0423@gmail.com kenji.hiramatsu@ifis.co.jp

概要

本研究では、ブローカーレポートにおけるセンチメントの判定に文章の長さが与える影響（バイアス）を体系的に評価する。サンプリングした長文のレポートに対して、LLMにより要約した短文を生成し、原文（長文）と要約（短文）のペアを作成する。そして、各ペアに対して、LLMやBERTを用いて相対比較することで得られる結果を集計し、二項検定により評価する。分析の結果、LLMによる比較において、ポジティブな文章は長文の方が短文よりもポジティブであると判定される傾向が確認でき、ネガティブな文章も同様に、長文の方が短文よりもネガティブであると判定される傾向が確認できた。一方で、BERTによるセンチメントスコアの比較においては、LLMと逆の結果が確認できた。

1 はじめに

ブローカーレポートは、投資判断に資する情報が含まれる長文のテキスト情報であり、そのセンチメント（ポジティブ・ネガティブ）を安定的かつ妥当に推定することは、投資支援やレポート評価において重要である。センチメントの計算には、近年、大規模言語モデル（LLM）が用いられることが多く、先行研究も多く存在する [1, 2]。しかし、広く用いられている LLM には、入力文章の長さや提示形式に起因するバイアスが混入する可能性が示唆されている [3]。金融ドメインにおけるセンチメント推定において、企業名の与えるバイアスの調査は行われているが [4]、文章の長さによるバイアスや、相対比較における順序によるバイアスの調査は行われておらず、実務で使用するためのモデル理解のためには調査が不可欠である。

本研究では、これらのバイアスの調査を行う。具体的には、サンプリングしたポジティブ（ネガティブ）な長文章のブローカーレポートに対して、LLMにより要約した短文章を生成し、原文（長文）と要

約文（短文）のペアを作成する。そして、作成した文章のペアを LLM に入力し、原文と要約文のどちらがよりポジティブ（ネガティブ）であるかを判定する相対比較を行うことで、入力した二つの文章に優越をつける。引き分けを許容しない場合、原文と要約文は本来同じセンチメントであることが期待され、文章の長さによるバイアスが存在しないのであれば、原文が選択される確率は $p = 0.5$ となる。本研究では、帰無仮説 $p = 0.5$ 、対立仮説 $p \neq 0.5$ の二項検定を確認することで、文章の長さが金融ドメインにおけるセンチメント推定に与える影響を検証する。また、相対比較において選択肢の提示順序によるバイアスを調査するために、選択肢の順序を入れ替えての比較も合わせて実施する。

本研究の貢献は以下の通りである。

- 金融センチメント分析における文章の長さが判定に与える影響の確認
- 金融センチメント分析における選択肢の提示順序が判定に与える影響の確認

2 提案手法

本研究の目的は、金融ドメインにおける想定比較によるセンチメント推定において、文章の長さが推定に及ぼす影響を調査することである。そのためには、センチメントが本来同じである長さの異なる金融文書のペアが必要である。しかし、センチメントは主観的なものであり、真の値を見積もることは難しく、センチメントが同じテキストのペアをマッチングさせることは困難である。そこで本研究では、元々の文章量が多いブローカーレポートを対象に要約文を自動生成することで、センチメントが限りなく近く、文章の長さが大きく異なる文章のペアを作成することで、この問題を解消する。提案手法は大きく3つのステップで構成される。

- 文章のサンプリング
- LLMによる要約文の自動生成

3. LLM による原文と要約文の相対比較

2.1 文章のサンプリング

相対比較において、ポジティブ同士の文章の比較とネガティブ同士の文章の比較が同様に行われるとは限らない。したがって、文章を事前にポジティブな文章とネガティブな文章の分類を行う。ポジティブとネガティブの分類は、情報が付与されていないテキストの場合はファインチューニング済みのBERTモデルやLLMを用いる。

2.2 LLM による要約文の自動生成

LLMによって元の長い文章を要約する。原文と要約文のセンチメントをなるべく一致させることができるプロンプトが望ましい。また、どのくらいの長さに要約するかも重要な点であり、文字列、単語数を具体的に固定値で指定するか、元の文章の10%などの割合で指定する必要がある。その他、原文の文体を模倣させるなどの具体的な指示を必要に応じて入れることが期待される。

2.3 LLM による原文と要約文の相対比較

原文と要約文をペアとして、LLMに「どちらがよりポジティブ(ネガティブ)か」を相対比較させる。位置バイアスの調査も合わせて確認するために、選択肢の順序を入れ替えての比較も実施する。

ポジティブな文章で使用するプロンプトを以下に示す。なお、アンダーライン部分は文章の埋め込みを意味する。

ポジティブ文章で使用するプロンプト

入力した文章Aと文章Bを読み、センチメントがよりポジティブだと感じた文章を選んでください。最終的な出力は'A'または'B'のみを出力してください。
文章A: {原文}
文章B: {要約文}

3 実証分析

本研究では、OpenAIが公開している「gpt-oss-20b¹⁾」を使用する。ハイパーパラメータは、デフォルトの値を使用する。

1) <https://huggingface.co/openai/gpt-oss-20b>

3.1 データセット

テキスト情報は、各証券会社が発行しているブローカーレポートPDFに対して、簡易的なデータクレンジングを行ったものを使用する。2017年から2025年に公開されたレポートの中から、日本語で書かれた文章をサンプリングする。ポジティブとネガティブの判定に関しては、以下のプロンプトを使用して判定を行う。

センチメント判定プロンプト

入力された文章のセンチメントを判定してください。回答は'positive', 'neutral', 'negative'のいずれかでお願いします。
{原文}

判定を行ったブローカーレポートを対象に、ポジティブなブローカーレポート250件、ネガティブなブローカーレポート250件を本研究では使用する。

3.2 要約文生成の設定

要約文の生成方法によって傾向が変わる可能性もあるため、プロンプトは以下の2種類を使用する。

文章要約プロンプト1

入力した文章を要約してください。400文字以内に要約してください。
{原文}

文章要約プロンプト2

入力した文章を要約してください。元の文章の10%の文字列に収まるように要約してください。
{原文}

3.3 LLM による相対比較の設定

各ペアについて、順序を入れ替える形で2回相対比較を実施する。ポジティブな文章で使用するプロンプトは、2.3節のプロンプトを使用し、ネガティブな文章は以下のプロンプトを使用する。

ネガティブな文章で使用するプロンプト

入力した文章Aと文章Bを読み、センチメントがよりネガティブだと感じた文章を選んでください。最終的な出力は'A'または'B'のみを出力してください。
文章A: {原文}
文章B: {要約文}

これらのプロンプトに対して、原文と要約文の埋め

表1 統計量

統計量	センチメント	要約プロンプト	原文の文字数	要約文の文字数	圧縮率
平均値	ポジティブ	プロンプト 1	2301.68	269.01	0.157
	ネガティブ	プロンプト 1	2049.45	185.67	0.176
	ポジティブ	プロンプト 2	2301.68	274.27	0.093
	ネガティブ	プロンプト 2	2049.45	184.35	0.100
中央値	ポジティブ	プロンプト 1	1909.50	266.00	0.141
	ネガティブ	プロンプト 1	1501.00	263.00	0.169
	ポジティブ	プロンプト 2	1909.50	158.00	0.084
	ネガティブ	プロンプト 2	1501.00	144.50	0.093

表2 LLMを用いた実証分析の検定結果

センチメント	要約プロンプト	N	X	\hat{p}_x		N	A	\hat{p}_a	
ポジティブ	プロンプト 1	500	293	0.586	***	500	225	0.450	**
ネガティブ	プロンプト 1	500	429	0.858	***	500	249	0.498	
ポジティブ	プロンプト 2	500	294	0.588	***	500	242	0.484	
ネガティブ	プロンプト 2	500	399	0.798	***	500	221	0.442	**

注：*は10%水準，**は5%水準，***は1%水準で有意

込み位置を逆にすることで、選択肢の順序によるバイアスの調査も同時に行う。

3.4 BERTによる相対比較の設定

従来のBERTによって計算されたセンチメントスコアを用いた相対比較も行う。長文も扱うことが可能であるModern BERTをベースに金融センチメント分析に特化したファインチューニング済みのモデルである「modernbert-ja-310m-eco-watcher²⁾」を使用する。このモデルは入力された文章に対して、連続値のセンチメントスコアを返すモデルである。したがって、各文章に対してセンチメントスコアを計算した後で比較を行う。

3.5 集計・統計解析

まず、LLMによる相対比較について述べる。ポジティブ・ネガティブ(2通り)×使用する要約プロンプト(2通り)の合計4通りの実証分析結果が得られる。例えば、ポジティブなブローカーレポートを対象に、プロンプト2で要約を行い、原文と要約文の相対比較を行う。

ある設定の実証分析によって、250件の原文と要約文のペアに対して順序を入れ替えて相対比較を行うため、500回の比較結果が得られる。ある設定の実証分析によって得られた500回の相対比較の結果、

原文が選択された回数を X とする。相対比較の合計回数である500回を N とするとき、推定値 $p_x = X/N$ を対象に二項検定を行う。具体的には、帰無仮説 $p_x = 0.5$ 、対立仮説 $p_x \neq 0.5$ の検定を行い、文章の長さによる影響の有無を検証する。

同様に、文章Aが選択された回数を A とし、推定値 $p_a = A/N$ を対象に二項検定を行う。帰無仮説 $p_a = 0.5$ 、対立仮説 $p_a \neq 0.5$ の検定を行うことで、選択肢の順序によって生じるバイアスの有無を検証する。

次に、BERTによる相対比較では、各文章に対して計算したセンチメントスコアを用いて、相対比較を行うため、順序の影響は受けず、順序を入れ替える必要がない。したがって、各設定において250回の相対比較の結果が得られ、原文が選択された回数を Y として二項検定を行う。ここで、LLMと設定を揃えるため、ポジティブな文章であれば、よりスコアが高い方を選択し、ネガティブな文章であれば、よりスコアが低い方を選択する。

3.6 結果

文章の要約によって、文字列がどのくらい短くなっているかを確認するために統計情報を表1に示す。次に、LLMを用いた4通りの実証分析の検定結果を表2に示す。また、BERTを用いた4通りの実証分析の検定結果を表3に示す。

2) <https://huggingface.co/takaito/modernbert-ja-310m-eco-watcher>

表3 BERTを用いた実証分析の検定結果

センチメント	要約プロンプト	N	Y	\hat{p}_y	
ポジティブ	プロンプト 1	250	56	0.224	***
ネガティブ	プロンプト 1	250	58	0.232	***
ポジティブ	プロンプト 2	250	54	0.216	***
ネガティブ	プロンプト 2	250	54	0.216	***

注：* は 10%水準，** は 5%水準，*** は 1%水準で有意

3.7 考察

表1の結果から、要約プロンプト1よりも要約プロンプト2の方が、文章をより短く要約している。これは、原文の文字数の平均値が2000から2300文字程度であるのに対して、400文字以内での要約を要求していることに起因している。また、要約プロンプト2では、「元の文章の10%の文字列に収まるように」と指示を入れているが、圧縮率の通り、たいていは指示に従う結果が得られている。

表2の結果から、原文と要約文、つまり文章の長さによって、センチメントの判定にバイアスが生じることが確認できた。ポジティブな文章では、短文よりも長文の方がポジティブであると判定される傾向が確認でき、ネガティブな文章においても、短文よりも長文の方がネガティブであると判定される傾向が確認できた。また、その傾向はネガティブな文章でより顕著に確認でき、文章Aと文章Bの順序を入れ替えた実験を含む500回の比較において、429回も原文がネガティブであると判定されている。

選択肢の順序に関しては、一部で有意な差が確認できた。また、全体的な傾向として、後半の選択肢を選択する傾向が確認できた。ただし、選択肢の順序によって生じるバイアスは、順序を入れ替えて判定を行うことで緩和することができるため、この程度であれば神経質になる必要はないと考える。

表3の結果から、BERTによるセンチメントスコア比較においても、文章の長さによってバイアスが生じることが確認できた。ただし、LLMの結果と真逆の結果となっており、ポジティブな文章では、長文よりも短文の方がポジティブであると判定される傾向が確認でき、ネガティブな文章においても、長文よりも短文の方がネガティブであると判定される傾向が確認できた。

表2と表3の結果から、LLMによる文章の直接比較とBERTのセンチメントスコアによる比較の特徴について考察する。BERTのセンチメントスコア

は、センチメントに関係のない内容を無視することができず、それらをニュートラルな内容と判定してしまう傾向が確認できた。要約文はセンチメントスコアの計算に不要な内容が削られていることで、センチメントスコアの絶対値が大きくなる傾向が確認でき、原文は不要な内容によってニュートラル側にセンチメントスコアが引っ張られることで、センチメントスコアの絶対値が小さくなる傾向が確認できた。一方で、LLMは、センチメントの判定に不要な文章をある程度無視することができているのではないかと考えている。また、ポジティブ（ネガティブ）な文章センチメントの判定において、より詳細な説明が含まれていれば、ポジティブ（ネガティブ）度合いを加算していくといった加算式で判定を行っている可能性が高いことが示唆される結果が得られた。

4 おわりに

ペアワイズやリストワイズを用いた分析は、ポイントワイズの分析よりも良好な結果が得られることが先行研究で示唆されているが[5]、金融センチメント分析においては、文章の長さによるバイアスが強く生じており、課題があることが明らかになった。また、そのバイアスは使用する手法によって異なることが確認できた。実務において、企業間分析を行うことは非常に多いため、このようなバイアスの存在は課題である。

本研究では、各手法で使用しているモデルが限定的であるため、これがモデル固有の傾向なのか、モデル全般の傾向なのか、引き続き調査が必要である。また、モデル全般の傾向である場合、取得可能な文章をそのまま使用するのではなく、数値データを扱うようにテキストデータも標準化が必要であると考えており、バイアスが生じないようにするような標準化や手法の開発を進めたいと考えている。

参考文献

- [1] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In **Proceedings of the Fourth ACM International Conference on AI in Finance**, pp. 374—382, 2023.
- [2] Katsuhiko Okada, Moe Nakasuji, Yasutomo Tsukioka, and Takahiro Yamasaki. From words to returns: sentiment analysis of japanese 10-k reports using advanced large language models. **PeerJ Computer Science**, p. 11:e3349, 2025.
- [3] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 517–545, 2024.
- [4] Kei Nakagawa, Masanori Hirano, and Yugo Fujimoto. Evaluating company-specific biases in financial sentiment analysis using large language models. In **2024 IEEE International Conference on Big Data**, pp. 6614–6623, 2024.
- [5] 高野海斗. 金融テキストを対象とした強弱を捉えることができるセンチメントモデル開発のためのデータセット構築方法の検討および分析. 人工知能学会第二種研究会資料, pp. 98–108, 2025.