GPT-4o による感情推定を用いた攻撃的投稿判定

草野雅也 [1]　佐久間拓人 [1]　谷文 [1]　加藤昇平 [1]
[1] 名古屋工業大学大学院 工学部 工学研究科
{kusano,sakuma,gu,shohey}@katolab.nitech.ac.jp

概要

近年，SNS 上で他者に対して攻撃的な投稿（攻撃的投稿）が増加しており，その自動検出に関する研究が実施されている．しかし，皮肉表現や文脈依存を含む投稿に対して誤判定する課題がある．本研究では，高い言語処理能力を持つ大規模言語モデルと感情情報を用いた，SNS 上の攻撃的投稿判定手法を提案する．提案手法では，GPT-4o を用いて喜びや悲しみなどの 8 種類の感情に対する推定確率を出力し，投稿と出力された推定確率の両方を参照して攻撃性判定する．正答率，F1 スコアおよび混同行列を用いて，感情情報および例示情報の有効性を調査した結果，LLM による攻撃性判定において情報の付加が有効であることが確認された．

1　はじめに

近年，世界中でソーシャル・ネットワーキング・サービス（SNS）の利用者数が増加傾向にあり，今後も利用者は増加することが推測されている [1]．しかし利用者の増加とともに，誹謗中傷やヘイトスピーチなど，他者に対して攻撃的な内容の投稿（攻撃的投稿）が増加し，社会的な問題となっている [2]．一方，各種 SNS において不適切な投稿を判定する AI が導入されているが，皮肉表現や前後の文脈を考慮する必要のある投稿に対して，攻撃的と誤検知しやすいという課題が存在する [3, 4]．我々は攻撃的投稿による問題を解決するには，攻撃的投稿に対する判定精度が重要だと考え，大規模言語モデル（LLM）と感情情報を用いて攻撃性判定する手法を提案した [5]．その結果，感情情報の有効性とプロンプト設計の重要性を示したが，学習コストが高いという課題が存在する．

そこで本研究では，LLM のみを用いて感情情報を活用した攻撃的投稿判定手法を提案する．感情情報を活用することで，投稿の感情的特徴を補助的な情報として利用して，判定の難しい投稿に対しても安定した攻撃性判定が期待される．提案手法では，判定対象の投稿から 8 種類の各感情の推定確率を，8 要素の配列データとして出力し，それらの感情情報と投稿の両方を参照して攻撃性判定する．LLM 単独の手法により，学習コストを抑えつつ LLM の言語処理能力を用いて，より安定した攻撃性判定を目指す．

性能評価実験では，従来手法，感情情報を利用しない手法，および感情情報を利用する提案手法を対象として，例示を与えない設定と例示を与える設定の両条件で比較し，正答率と F1 スコア，さらに混同行列を用いて評価した．その結果，提案手法の有効性が示され，LLM の攻撃的判定では感情情報や例示が必要であることが示唆された．

2　関連研究

2.1　SNS 上の攻撃性のある投稿の検知

SNS 上の攻撃的投稿の検知を目的とした研究は，数多く報告されている．これまでに新たな手法を提案した研究と攻撃性の定義やラベル設計に着目した研究が実施されている．新たな手法を提案した研究では，攻撃性タスクに加えて感情情報を補助タスクとして導入した手法 [6, 7] や，CNN と GRU を統合した手法 [8] を提案している．また攻撃性の定義やラベル設計に着目した研究では，攻撃性の有無だけでなく，攻撃の種類や標的に着目した研究 [4] や，攻撃性判定が難しい投稿をグレーゾーンというラベルで新たに定義した研究 [9] が実施されている．

これらの研究では，新たな手法や攻撃性の定義やラベル設計によって，SNS 投稿の攻撃性判定に有効になる可能性が報告されているが，皮肉表現や前後の文脈を考慮する必要のある投稿に対して，誤判定しやすいという課題が指摘されている．そこで本研究では，LLM を用いて投稿と感情情報を統合的に扱うことで，従来手法では誤判定されやすい皮肉表現や前後の文脈を考慮する必要のある投稿に対する

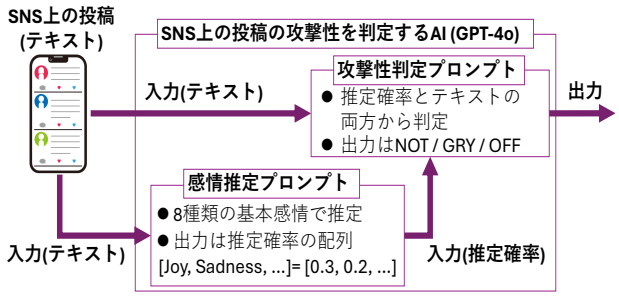


図 1: 提案手法

判定精度の向上を目指す．

2.2 LLM を用いた SNS 投稿分析

テキスト分析に LLM を用いた研究も実施されている．Rathje ら [10] は LLM を用いて，多言語のテキストから感情や攻撃性などの心理的特徴を推定できるかを検証した．その結果，LLM は辞書ベースの手法よりも高い性能を示し，データ数の少ない言語に対しても一定の有効性が確認された．しかし，従来手法の機械学習モデルよりも性能が低いタスクも存在したことが報告されている．また竹中ら [11] は，日本語感情分析におけるモデル比較として，BERT 系モデルと GPT-4o を検証した結果，GPT-4o の性能は BERT 系モデルに劣っていたが，プロンプト設計などによる改善の余地があることを指摘している．

以上の研究により，LLM を用いたテキスト分析は多言語対応や学習コストの低さより，LLM のテキスト分析への応用可能性が報告されている．そこで本研究では，竹中らが提案した感情推定プロンプトを参考にして，感情情報を明示的に確率分布として出力し，その結果を用いて攻撃性判定する手法を構築する．LLM の言語処理能力を利用しながら感情情報を介して攻撃性判定することで，従来手法では誤判定されやすかった文脈依存的または曖昧な投稿に対する判定精度の向上を目指す．

3 提案手法

図 1 に提案手法を示す．本研究では OpenAI 社が提供する GPT-4o を使用し，攻撃性判定を使用するデータセットに対応した 3 値分類タスクを対象とする．本研究で使用したプロンプトは，竹中らの研究で提案されたプロンプト [11] を参考に，「SNS 上の投稿の攻撃性を判定する AI」の役割を指示して，感情推定タスクを実施してから攻撃性判定タスクを実施することで攻撃性判定するプロンプトを設計した．付録 A に本研究で使用したプロンプト全文を付する．

3.1 感情推定プロンプト

感情推定タスクを実施するプロンプトは，判定対象の投稿を入力として，プルチックの基本 8 感情（喜び, 悲しみ, 期待, 驚き, 怒り, 恐れ, 嫌悪, 信頼）の推定確率を出力するように設計した．プロンプト設計にあたっては，竹中らの研究 [11] を参考に，各感情の評価基準をプロンプト内に明示し，各感情の推定確率を，0 から 1 までの連続値で出力するように設計した．評価基準を以下に示す．

- Joy（喜び）：幸福感、満足感、ポジティブな気持ちが表現されている
- Sadness（悲しみ）：悲嘆、喪失感、絶望感を示す表現がある
- Anticipation（期待）：未来への期待や希望、予測が示されている
- Surprise（驚き）：予想外の出来事に対する驚きや戸惑いがある
- Anger（怒り）：怒りや強い反発心を示す言葉がある
- Fear（恐怖）：危険や不安、恐怖心を表す表現がある
- Disgust（嫌悪）：嫌悪感や拒絶感、強い不快感を表す言葉がある
- Trust（信頼）：他者への信頼、安心感、信用を示す表現がある

また一貫した確率分布を出力するため，8 感情の推定確率の総和が 1 となる制約をプロンプトに追加した．これにより，投稿に内在する感情情報を確率分布として扱い，後に実施する攻撃性判定において利用できるように設計した．

3.2 攻撃性判定プロンプト

攻撃性判定を実施するプロンプトは，感情推定プロンプトによって出力した各感情の推定確率分布と，元の投稿テキストを入力として，投稿の攻撃性を判定するように設計した．

なお，本稿では性能評価実験で使用するデータセットに基づき，NOT(非攻撃的)，GRY(グレーゾーン)，OFF(攻撃的) の 3 値分類を実施するように設計した．藤原らの研究 [9] を参考に，各攻撃性における評価基準をプロンプト内で明示した．各攻撃性の評価基準を以下に示す．

- NOT（非攻撃的）：他者に対して攻撃的な表現が無い
- GRY（グレーゾーン）：読み手や文脈によって攻撃的とも非攻撃的とも受け取られる可能性がある
- OFF（攻撃的）：ヘイトスピーチや誹謗中傷など、他者に対して攻撃的な表現がある

4 性能評価実験

4.1 使用したデータセット

本研究では，藤原らが構築した攻撃性推定用データセット [9] を使用した．このデータセットは X(旧 Twitter) から投稿を 800 件収集し，大学生 3 名による攻撃性評価によって NOT(非攻撃的)，GRY(グレーゾーン)，OFF(攻撃的) の 3 種類のラベルが付与されている．攻撃性ラベルは「攻撃性の可能性」と，攻撃的だと仮定したときの「攻撃性の強さ」の 2 種類の評価指標を用いて付与された．どちらの評価指標も低い場合は NOT，高い場合は OFF，片方の指標のみが高い場合は GRY とラベルが付与されている．各ラベルの分布は NOT が 334 件，GRY が 320 件，OFF が 126 件である．本研究では，このデータセットを用いて提案手法の評価を実施した．

4.2 実験内容

本研究では，感情情報の有無および例示の有無が攻撃性判定性能に与える影響を検証するため，以下の 5 つの手法を比較する性能評価実験を実施した．また，LLM を用いた全ての手法で temperature パラメータを 0 と設定し，例示については評価基準とは別に，定義が曖昧で判断が難しい投稿に対する解釈の補助を目的として，各攻撃性ラベルに対応する例を 1 件用意し，プロンプト内に付与した．

Baseline :	藤原らが提案した手法 [9]
Zero-shot :	感情情報を利用せずに攻撃性判定する手法（例示無し）
Few-shot :	感情情報を利用せずに攻撃性判定する手法（例示あり）
Proposed (Zero) :	感情情報を利用して攻撃性判定する手法（例示無し）
Proposed (Few) :	感情情報を利用して攻撃性判定する手法（例示あり）

性能評価は，藤原らの研究で使用されていた正答率（Acc.），各クラスの F1 スコア (F1) とマクロ平均 F1 スコア (Macro F1)，さらに正解ラベルと予測ラベルの混同行列を用いた．これらの指標に基づいて各手法を比較することで，攻撃的投稿の検知における LLM の有効性と，LLM での攻撃性判定における感情推定と例示の有効性について検証した．

4.3 実験結果

表 1 に各手法における正答率と各クラスの F1 スコアを示す．提案手法である Proposed(Zero) および Proposed(Few) が，Baseline と比較して全体的に高い性能であった．Proposed(Few) は，正答率，GRY クラスの F1 スコア，マクロ平均 F1 スコアで最も高い性能であることが確認され，Proposed(Zero) は OFF クラスの F1 スコアで最も高い性能であることが確認された．これにより，感情情報を利用した提案手法は特に GRY クラスや OFF クラスの判定において有効であることが示された．次に Baseline と Zero-shot を比較すると，Zero-shot は NOT クラスの F1 スコア以外では Baseline を下回った．特に GRY クラスの F1 スコアで，0.555 から 0.476 まで下回った．しかし，Few-shot と Proposed(Zero) では Baseline と比較して，全ての指標で同等の性能か高い性能を示した．これにより，LLM を用いた攻撃性判定では例示情報や感情情報が必要であることが示された．

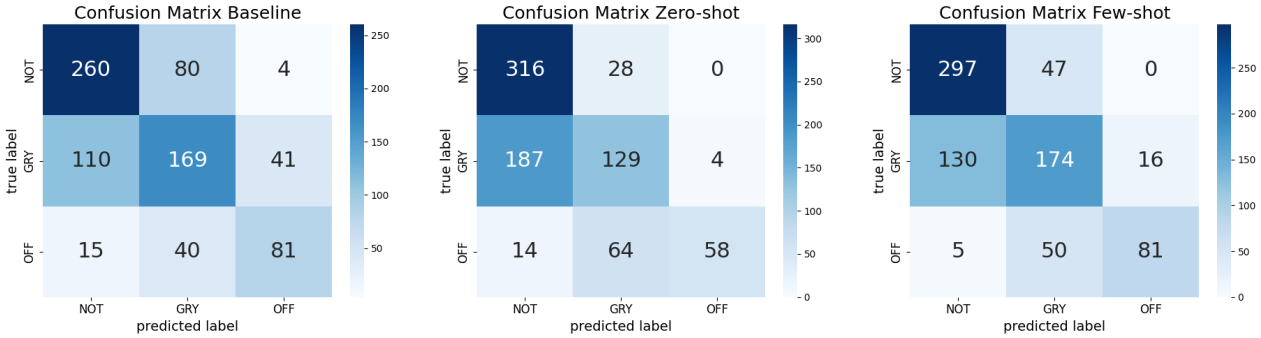
図 2 に各手法における正解ラベルと予測ラベルの混同行列を示す．すべての混同行列において，縦軸が正解ラベル，横軸が予測ラベルである．Baseline と Proposed(Few) を比較すると，Proposed(Few) では全てのクラスで正しく分類した件数が増加したことが確認された．次に Baseline と Zero-shot を比較すると，Zero-shot では GRY クラスを NOT クラスと誤判定する件数の増加したことが確認された．しかし，Few-shot と Proposed(Zero) では Baseline と比較して，GRY クラスと OFF クラスで誤判定する件数が減少したことが確認された．これにより，LLM を用いた攻撃性判定では例示情報や感情情報を付与することで，GRY クラスと OFF クラスの判定向上に寄与して性能向上することが示された．

5 考察

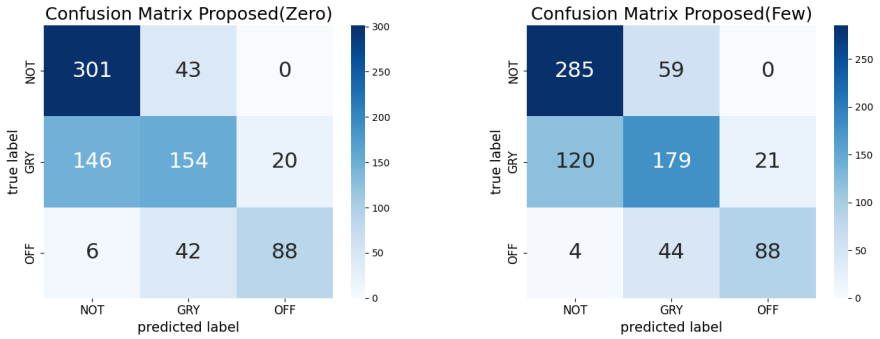
実験結果より，実験結果より，LLM を用いた攻撃性判定においては，感情情報や例示情報を付与することで，Baseline と比較して性能が向上することが確認された．これは LLM が高い言語処理能力を持つ一方で，本研究における GRY クラスのように判定基準が曖昧なクラスに対しては，判定の指針となる情報がないため，適切な判断が困難となるためであると考えられる．例示情報を提示した手法の方が感情情報よりも誤判定の減少に貢献した理由としては，攻撃性判定の基準を直接指示できたことが考えられる．感情情報は投稿に含まれる感情状態を数値として表現できるが，この情報をどのように扱うかは LLM に依存している．これに対して例示は，各クラスに該当する投稿例を通じて判定基準を明示す

表 1: 各手法の正答率と F1 スコア

比較手法	例示情報	感情情報	Acc.	F1(NOT)	F1(GRY)	F1(OFF)	Macro F1
Baseline	-	-	0.638	0.713	0.555	0.618	0.629
Zero-shot			0.629	0.734	0.476	0.586	0.599
Few-shot	✓		**0.690**	**0.765**	0.589	0.695	0.683
Proposed(Zero)		✓	0.679	0.755	0.551	**0.721**	0.676
Proposed(Few)	✓	✓	**0.690**	0.757	**0.595**	0.718	**0.690**



(a) Baseline (b) Zero-shot (c) Few-shot



(d) Proposed (Zero) (e) Proposed (Few)

図 2: 各手法の正解ラベル-予測ラベルの混同行列

ることで，判定基準を明確化できたと考えられる．さらに LLM を用いた提案手法は，感情情報や例示をプロンプトとして付与するのみで判定性能を向上させられるので，各種 SNS ごとに異なる攻撃性の定義や許容範囲に応じてプロンプトを調整することによって，柔軟かつ安定した攻撃性判定が可能と考えられる．

以上より，LLM を用いて SNS 上の攻撃的投稿を判定するためには，判定例や評価基準といった指標となる情報も指示することが重要であると示唆される．一方で，本研究で用いたデータセットでは GRY クラスという判定基準が曖昧なクラスが性能に影響している可能性も考えられる．そのため，今後は GRY クラスが存在しない他データセットを用いて検証する必要がある．

6 おわりに

本研究では，LLM のみを用いて感情情報を活用した攻撃的投稿判定手法を提案し，感情情報の有無および例示の有無が判定性能に与える影響を検証した．藤原らの攻撃性推定用データセット [9] を用いた性能評価の結果，投稿テキストのみを入力する LLM の直接判定は従来手法に対して優位でない一方で，感情情報や例示をプロンプトとして付与することで，GRY クラスにおける誤判定が抑制されることを確認され，感情情報や例示などの追加情報によって判定基準を補助することが必要であることが示唆された．今後は，他データセットでの検証を通じて提案手法の汎用性を評価しつつ，判定精度の向上を目指して手法を改良する．

謝辞

本研究は，一部，文部科学省科学研究費補助金（課題番号 JP24H00741，JP24K20900），ならびに，国立研究開発法人情報通信研究機構委託研究の助成により行われた．

参考文献

[1] 総務省. 令和 6 年版情報通信白書, 2024. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/html/nd217100.html.

[2] 総務省. 令和 5 年度 インターネット上の違法・有害情報対応 相談業務等請負業務 報告書, 2024. https://www.soumu.go.jp/main_content/000946765.pdf.

[3] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In **Proceedings of the international AAAI conference on web and social media**, 2017.

[4] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In **Association for Computational Linguistics**, 2019.

[5] 草野雅也, 佐久間拓人, 谷文, 加藤昇平. 感情推定モデルと LLM 出力の統合による SNS 上の攻撃的投稿検知の検証. HCG シンポジウム 2025 pp.P-2-46 (4-pages), 2025.

[6] Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. Hate speech and offensive language detection using an emotion-aware shared encoder. In **ICC 2023-IEEE International Conference on Communications**, pp. 2852–2857. IEEE, 2023.

[7] 加藤大造, Le-Minh NGUYEN. 方言コーパスを用いた感情分析モデルの構築と炎上・ネットいじめ検知手法の提案. 言語処理学会第 30 回年次大会発表論文集, pp. 581–586, 2024.

[8] Kevin Usmayadhy Wijaya and Erwin Budi Setiawan. Hate speech detection using convolutional neural network and gated recurrent unit with fasttext feature expansion on twitter. **Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)**, Vol. 9, No. 3, pp. 619–631, 2023.

[9] 藤原知樹, 伊藤彰則, 能勢隆. ソーシャルメディア上の発話の攻撃性推定と会話補助. 言語処理学会第 30 回年次大会発表論文集, pp. 505–510, 2024.

[10] Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E Robertson, and Jay J Van Bavel. GPT is an effective tool for multilingual psychological text analysis. **Proceedings of the National Academy of Sciences**, Vol. 121, No. 34, p. e2308950121, 2024.

[11] Yoichi Takenaka. Performance evaluation of emotion classification in japanese using RoBERTa and DeBERTa, 2025.

A　本研究で使用したプロンプト全文

本研究で使用したプロンプトを図 3 に示す．感情情報を経由しない手法では，図 3 に示すプロンプトから「1. 感情推定」に関する記述および感情分布の出力を削除し，例示情報を用いない手法では「例（攻撃性ラベルの目安）」に関する記述を削除し，攻撃性判定のみを実施するように設定した．

タスク
1. 感情推定以下の短文に含まれる 8 つの基本感情（Joy, Sadness, Anticipation, Surprise, Anger, Fear, Disgust, Trust）について、それぞれの 推定確率 を 0〜1 の連続値 で出力してください。
評価基準（各感情の判断基準）：
- Joy（喜び）: 幸福感、満足感、ポジティブな気持ちが表現されている
- Sadness（悲しみ）: 悲嘆、喪失感、絶望感を示す表現がある
- Anticipation（期待）: 未来への期待や希望、予測が示されている
- Surprise（驚き）: 予想外の出来事に対する驚きや戸惑いがある
- Anger（怒り）: 怒りや強い反発心を示す言葉がある
- Fear（恐怖）: 危険や不安、恐怖心を表す表現がある
- Disgust（嫌悪）: 嫌悪感や拒絶感、強い不快感を表す言葉がある
- Trust（信頼）: 他者への信頼、安心感、信用を示す表現がある
制約：- 8 つの値の合計は必ず 1 にしてください。
—
2. 攻撃性判定上記で推定した感情分布も参考にして、投稿の攻撃性を NOT / GRY / OFF のいずれかで判定してください。
評価基準（攻撃性）：
- NOT（非攻撃的）：他者に対して攻撃的な表現が無い
- GRY（グレーゾーン）：読み手や文脈によって攻撃的とも非攻撃的とも受け取られる可能性がある
- OFF（攻撃的）：ヘイトスピーチや誹謗中傷など、他者に対して攻撃的な表現がある
—
重要- 出力は必ず JSON 形式のみ としてください- "aggressiveness" の値は NOT / GRY / OFF のいずれか 1 つのみを出力してください- 以下のフォーマットを厳守してください出力フォーマット (JSON のみ)： "probabilities": [0.12, 0.05, 0.18, 0.10, 0.30, 0.15, 0.10, 0.00], "aggressiveness": "OFF"
—
例（攻撃性ラベルの目安）：
入力短文：「今日のご飯おいしかった！」出力例 (JSON のみ): "probabilities": [0.50, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.20], "aggressiveness": "NOT"
入力短文：「いい加減にしてくれ。」出力例 (JSON のみ): "probabilities": [0.05, 0.20, 0.05, 0.50, 0.05, 0.05, 0.05, 0.05], "aggressiveness": "GRY"
入力短文：「こいつは本当に役に立たない。」出力例 (JSON のみ): "probabilities": [0.05, 0.05, 0.05, 0.05, 0.20, 0.05, 0.50, 0.05], "aggressiveness": "OFF"
— ここまで例 —
入力短文：「text」

図 3: プロンプト全文