LLM を用いた文脈考慮による攻撃性検出性能の改善

中野 雄斗 [1]　佐藤 志貴 [1]　赤間 怜奈 [1,2]
[1] 東北大学　[2] 理化学研究所
nakano.yuto.t2@dc.tohoku.ac.jp, {shiki.sato.d1, akama}@tohoku.ac.jp

概要

本稿では，長期文脈を扱える最新の大規模言語モデルは，SNS 投稿の攻撃性検出において適切な文脈考慮が可能か検証する．そのために，検出対象の投稿単体で攻撃性を評価するベースラインモデルと対話履歴を考慮して攻撃性を評価するモデルの比較実験を行った．実験の結果，対話履歴が特に必要とされる攻撃性を含まない投稿の検出に関しては，人手による検出と同様の傾向を示すなど，適切な文脈考慮ができている可能性があることがわかった．一方で，本来攻撃性のある投稿まで攻撃性が無くなったと誤った判断をしてしまうという課題が残った．

1　はじめに

昨今 SNS 上は誹謗中傷などの攻撃的表現で溢れており，攻撃対象になったユーザだけでなく，その投稿を目にしたユーザの心身にも負担を与えている [1, 2, 3, 4, 5, 6, 7, 8]．ユーザの負担を軽減するために，攻撃性検出や [3, 9, 10, 11]，攻撃性除去など [4, 5, 12, 13, 14, 15] を用いた第三者による投稿監視 [2, 16] の必要性が叫ばれている．

攻撃性検出に関する既存研究や実際に使われている商用検出モデルでは，対象の投稿単体で攻撃性を判断することが多い [3, 9, 10, 11]．しかし実際には，ユーザは対象の投稿だけでなく，周りのユーザ情報 [6, 17] や記事のタイトル [18]，そして対話履歴 [1, 7, 19, 20, 21, 22] などの文脈に目を通しながら対象の投稿に出会っている．以下，本稿では対話履歴の意で文脈と表現する．また投稿の直上にあたる１つの文脈を親投稿と呼ぶ．先行研究によると，人手の検出では文脈考慮により分類傾向が大きく変化し [20]，攻撃性が無い投稿をより適切に分類できるようになることが示されており [19, 23]，考慮する親投稿の数が多いほどその傾向が顕著になると報告されている [22]．また，3 節の分析から，検出対象の投稿が攻撃的表現を含む場合は攻撃性検出に文脈が不要な割合が高く，攻撃的表現を含まない投稿の検出には文脈が必要な割合が高いことがわかった．以上から人手による攻撃性検出の判断材料としては投稿単体だと不十分であり，特に攻撃性を含まない投稿に関しては文脈を考慮することによって，より適切な判断ができるようになる．

攻撃性の自動検出における文脈を考慮した先行研究では，１つ前の親投稿 [5, 12, 18, 19, 20, 24]，先行する複数個の投稿 [7, 21]，全文脈 [1, 22] を考慮したが，ほとんどはモデル化に失敗しており，文脈が逆にノイズとなり精度が下がるという結果になった．

そこで我々は大規模言語モデル (LLM) が長期文脈を扱える点に注目した．最新の LLM は，様々なタスクで高い性能を出しており，[8, 25]，特に ChatGPT においては，生成文の品質がクラウドワーカーの注釈よりも優れていると報告されている [26]．長期文脈については，全ての情報を適切に考慮できるわけではないという先行研究もあるが [27]，タスクによっては性能の向上も報告されている [28, 29]．

本研究では，長期文脈を扱える最新の LLM は，攻撃性検出において適切な文脈考慮が可能か検証するために，検出対象の投稿単体で攻撃性を評価するベースラインモデルと文脈を考慮するモデルの比較実験を行った[1)]．検出モデルには，現時点の最新モデルである GPT-3.5-turbo [30] と GPT-4-turbo [31] を用いた．実験の結果，文脈が特に必要とされる攻撃性を含まない投稿の検出に関しては，人手による検出と同様の傾向が得られているため，適切な文脈考慮ができている可能性があることがわかった．一方で，攻撃性のある投稿まで攻撃性が無くなったと誤った判断をするという課題が残った．

2　関連研究

攻撃性検出において文脈を考慮するべきか事前に判断する文脈感応度推定の研究が行われている [24, 32]．先行研究ではデータセットを攻撃性検出

1)　警告：本稿には気分を害する表現の例が含まれています．

表 1: CAD データセットのラベル分布.

ラベル	細分類	投稿数
Abusive	Affiliation directed	243
	Identity directed	513
	Person directed	237
	小計	906
Non-Abusive	Counter Speech	4,342
	Neutral	59
	小計	4,401
合計		5,307

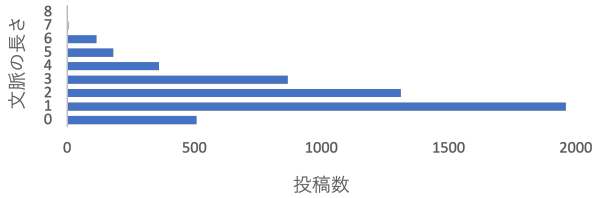

図 1: CAD データセットの文脈の長さの分布.

に文脈が必要な投稿と不要な投稿に人手で分割し，文脈考慮を行うモデルと行わないモデルを用意している．それぞれの投稿に対して 2 つのモデルで実験を行った結果，事前の判断に基づいて必要な場合のみ文脈を考慮すると，検出精度が向上すると報告されている．しかし，文脈考慮必要性の判断に人手評価を用いるとコストが増加する一方，文脈感応度推定を自動で行っても二段階の分類過程を経ることで誤分類が増加するという問題がある．我々は文脈考慮の必要性を検出器自身が判断するシステムを実現する必要があると考え，今回の検証を実施する．

3 データセット

文脈を考慮した攻撃性検出の検証を行うために，攻撃性検出に関する文脈を考慮した人手の注釈が行われており，文脈が復元可能なデータセットが必要である．そこで，Reddit 投稿をもとに構築された Contextual Abuse Dataset (CAD データセット) [1] のテストデータを用いた．CAD データセットは，文脈が考慮された人手の注釈が行われている．表 1 のように，攻撃性が有るか無いか (Abusive/Non-Abusive) かの二値ラベルが付与されており，Non-Abusive 投稿の割合が多いという特徴がある．また，細かく分類した 5 種類のラベルも付与されている．本実験では前者の二値分類を採用し，後者の細分類は節 4 で述べるように Few-shot を抽出する際に用いた．

文脈の復元方法． CAD データセットでは，1 つ前の親投稿の ID の注釈が行われている．文脈を構築するために，検出対象の投稿から親投稿の ID を参照し，親投稿から次の親投稿の ID を参照することを繰り返した．ここで構築した対象の投稿がもつ一連の親投稿の数を文脈の長さと呼ぶ[2)]．文脈の長さの分布を図 1 に示す．

データセットの分析． また，CAD データセットには人手の攻撃性検出の際に文脈が必要であったどうかを示すラベルが注釈されている．Abusive 投稿は，777 件が不要，317 件が必要であり，文脈が不要だと判断される割合が高かった．一方 Non-Abusive 投稿は 17 件が不要，59 件が必要であり，文脈が必要だと判断される割合が高かった．以上から，人手評価では，文脈考慮によって投稿の攻撃性が新たに表出することは比較的少ないということがわかった．

```
You are an expert assistant in the field of moderation in social media.
Your task is to help workers to moderate in social media.
Your task is to classify the user-written target text in order to eliminate
toxic, offensive and abusive text.
In order to help the worker, you MUST respond with the number and the name
one of the following classes you know for the following "Target comment".

The classes are:
0 Non-Abusive
1 Abusive

Here are some examples of questions and their classes:
{few-shotで与える 6 つの例}

Target comment is:
{対象の投稿}
```

(a) ベースラインモデルのプロンプト.

```
You are an expert assistant in the field of moderation in social media.
Your task is to help workers to moderate in social media.
Your task is to classify the user-written target text in order to eliminate
toxic, offensive and abusive text.
In order to help the worker, you MUST respond with the number and the name one
of the following classes you know for the following "Target comment", taking
into account the accompanying dialogue history.
Note that when classifying, the dialogue history itself is not considered
Abusive or Non-abusive, but is only used as a context for the "Target comment".

The classes are:
0 Non-Abusive
1 Abusive

Here are some examples of questions and their classes:
{few-shotで与える 6 つの例}

Accompanying dialogue history:
{対話履歴}

Target comment from User{番号} is:
{対象の投稿}
```

(b) 文脈考慮モデルのプロンプト.

図 2: 本実験の攻撃性検出モデルに入力したプロンプト.

4 攻撃性検出モデル

LLM が文脈を考慮した適切な攻撃性検出が可能か検証するために，対象の投稿単体で攻撃性を評価するモデル (ベースラインモデル) と，対象の投稿と対話履歴を入力するモデル (文脈考慮モデル) の比較実験を行った．両モデルとも GPT-3.5-turbo [30] と GPT-4-turbo [31] を採用し，計 4 つのモデルで実験を行った．両モデルのプロンプトを図 2 に示す．形式は Loukas らのプロンプトを参考にした [33].

2) ただし，文脈の長さには対象の投稿自体を含めず，文脈を含まない投稿については，文脈の長さが 0 として扱う.

ベースラインモデルはプロンプトにタスクの説明，Few-shot の 6 つの例，対象の投稿を入力し，文脈考慮モデルは文脈を追加で入力した．それぞれの検出モデルは，対象の投稿の攻撃性が有るか無いか (Abusive/Non-Abusive) の二値のラベルを出力する．

Few-shot の抽出方法． Few-shot の 6 つの例は，CAD データセットの訓練データから Abusive 投稿と Non-Abusive 投稿を 3 つずつ抽出し用意した．ベースラインモデルと文脈考慮モデルで同じ例を用いており，文脈は追加せずに対象の投稿のみを例示した．具体的には，異なる傾向を持つ投稿を例示するために，まず，表 1 の 5 つの細分類から 1 つずつランダムに選んだ．続いて，二値ラベルの数を等しくするために，最も投稿数の多い Non-Abusive に属する Neutral からランダムに 1 つの投稿を選んだ．

5 実験

5.1 評価方法

モデルの評価には正答率，適合率，再現率，F1 値を用いた．CAD データセットは表 1 のように Non-Abusive 投稿よりも Abusive 投稿が少なく不均衡である．そこで，データ数の多い Non-Abusive の投稿数をランダムに削減し，Abusive 投稿の数と揃えて正答率を測定した．Non-Abusive 投稿のダウンサンプリングを 100 回行い，Abusive 投稿と合わせ，求めた正答率の平均をバランス正答率と定義する．

また，文脈の長さに注目して評価を行った．人手評価では，文脈が長いほど攻撃性の無い投稿をより適切に分類できるようになる [22]．一方 LLM では，全ての入力情報を適切に考慮できるわけではないことが報告されているため [27]，文脈の長さが検出精度と直結するとは限らない．そこで，LLM における文脈の長さと攻撃性検出の精度の関係性を検証するために，上記のそれぞれの評価指標について，文脈の長さごとに投稿を分類して評価を行った．文脈の長さによる評価値の変化が文脈考慮によるものであることを確かめるために，文脈考慮を行わないベースラインモデルでも文脈の長さごとの評価を行った．ただしベースラインモデルでは実際に文脈を考慮しているというわけではない．

5.2 実験結果

ベースラインモデルと文脈考慮モデルの比較． 各モデルの実験結果を表 2 に示す．GPT-3.5-turbo と GPT-4-turbo を比較すると，両モデルの文脈考慮の有無による変化の傾向は類似しており，ほとんどの評価指標において GPT-4-turbo が上回っていた．続いて，ベースラインモデルと文脈考慮モデルを比較すると，文脈考慮による F1 値の大きな変化はみられなかった．一方その他の評価指標の差に注目すると，分類の特徴が異なることがわかった．具体的には文脈考慮モデルのほうが，正答率 (ACC) がやや高く，バランス正答率 (Balanced ACC) はやや低くなった．CAD データセットの Non-Abusive 投稿の割合が多い中，文脈考慮モデルが Non-Abusive と分類することが多かったため，正答率のみ高くなったと考えられる．また，GPT-4-turbo におけるベースラインモデルと文脈考慮モデルを比較すると，文脈考慮モデルのほうが，Abusive 検出における適合率が高くなり再現率は低くなること，逆に Non-Abusive 検出における適合率は低くなり再現率が高くなることがわかった．以上から，LLM は文脈考慮が特に必要とされる Non-Abusive 検出においては，文脈考慮によって適切な分類ができるという人手による検出 [19, 23] と同様の傾向が得られた．一方攻撃性のある投稿まで Non-Abusive と誤って分類し，Abusive 検出の再現率が下がってしまうという課題が残った．この場合の定性的なエラー分析を行った．具体例を 3 に示す．対象の投稿単体では明確な悪意が感じられるが．文脈を踏まえると，攻撃性は弱くなるが捉え方によっては攻撃的になる，人間でも判断が難しい場合が多かった．詳細は付録 A を参照されたい．

文脈の長さごとの評価． 文脈の長さごとに評価したグラフを図 3 に示す．Abusive 検出においては，文脈の長さによって評価値が上下しており，この実験においては長さと精度に大きな相関が無いことがわかった．一方 Non-Abusive 検出では，再現率のグラフの傾きに注目すると，GPT-4-turbo の文脈考慮モデルのほうがベースラインモデルよりも傾きが大きいことが確認できた．この傾きの差が有意であることを確かめるために検定を行った．それぞれのモデルについて，正解ラベルが Non-Abusive である投稿の分類結果が正解であった投稿の集合と，不正解であった投稿の集合に分割し，2 つの集合の文脈の長さの平均を求めた．有意水準 1%の片側 t 検定に基づき，2 つの集合間で文脈の長さの平均が有意に異なる場合，文脈が長いほど適切な Non-Abusive 検出に寄与する傾向にあるとみなした．t 検定の結果，ベースラインモデルでは統計量は 2.4，p 値は 0.017

表 2: 攻撃性検出モデルにおける文脈考慮の有無を比較した GPT-3.5-turbo と GPT-4 の評価結果．左から正答率，バランス正答率，Non-Abusive 検出における適合率，再現率，F1 値，Abusive 検出における適合率，再現率，F1 値．

				Non-Abusive			Abusive		
		ACC	Balanced ACC	P	R	F1	P	R	F1
GPT-3.5-turbo	ベースラインモデル	78.9	74.0	92.2	81.5	86.5	42.5	66.4	51.9
	文脈考慮モデル	79.2	72.0	91.3	82.8	86.8	42.4	61.5	50.2
	文脈考慮による差分	+1.2	-2.0	-0.9	+1.3	+0.3	-0.1	-4.9	-1.7
GPT-4	ベースラインモデル	81.4	77.3	93.3	83.5	88.2	47.0	71.1	56.6
	文脈考慮モデル	83.8	75.7	92.1	88.0	90.0	52.1	63.5	57.2
	文脈考慮による差分	+2.4	-1.6	-1.2	+4.5	+1.8	+5.1	-7.6	+0.6

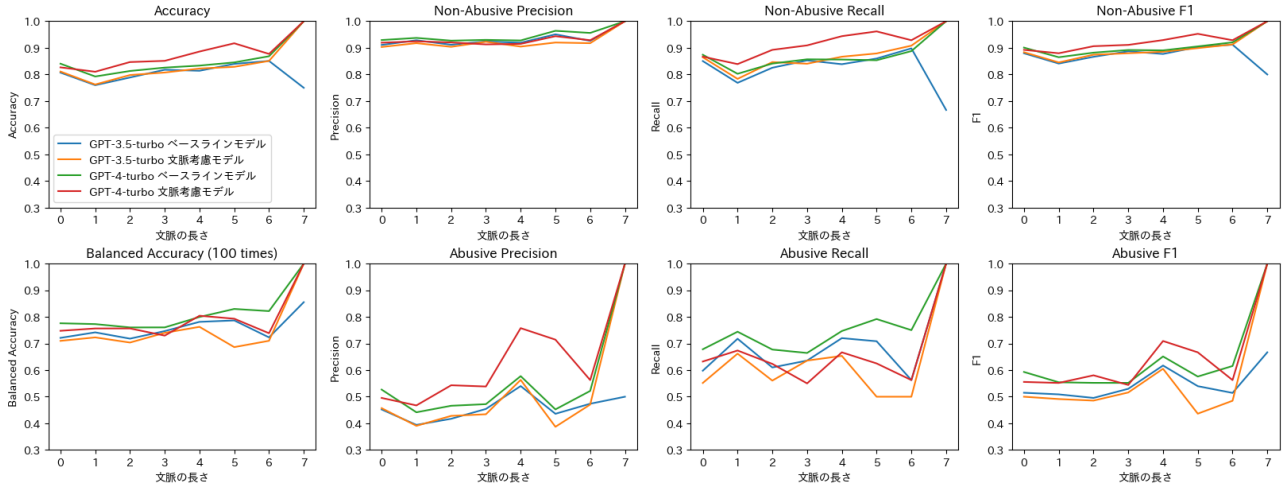

図 3: 文脈の長さに注目した実験結果．左 1 列は上から正答率とバランス正答率．右 3 列は上段が Non-Abusive 検出，下段が Abusive 検出．それぞれ左から適合率，再現率，F1 値．縦軸の描画範囲は 0.3 から 1.0 であることに注意．

User2: Loophole: move to Mexico and become a legal citizen. Illegally return to United States. Live in Cali. Collect free healthcare.
User3: Just start speaking Spanish, and go down and get free stuff like everybody else

表 3: 定性的なエラー分析の具体例．User2 が文脈，User3 が検出対象の投稿．詳細は付録 A を参照．

となり有意な差がなく，文脈考慮モデルでは統計量は 6.6，p 値は 3.6×10^{-11} となり有意な差があった．以上から考慮する文脈の長さが増加すると，攻撃性がない投稿をより適切に分類することができるようになることがわかった．この結果は先行研究の人手評価 [22] と一致しており，本実験においては，文脈の長さの観点でも，人手評価と GPT-4-turbo の傾向が類似していることが確認できた．

6 おわりに

本稿では，長期文脈を扱える最新の LLM は，攻撃性検出において適切な文脈考慮が可能であるかを検証するために，ベースラインモデルと文脈考慮モデルの比較実験を行った．実験の結果，文脈が特に必要とされる Abusive 投稿の検出に関しては，文脈考慮によって適切な分類ができるという人手による検出と同様の傾向が得られたが，本来攻撃性がある投稿まで Non-Abusive と誤った判断をしてしまうという課題が残った．また文脈の長さの分析から，GPT-4-turbo は考慮する文脈の長さが増加すると攻撃性が無い投稿をより適切に分類できるようになることがわかり，文脈の長さの観点でも，人手評価と GPT-4-turbo の傾向が類似していることが確認できた．しかし，本実験では 1 つのプロンプトと 1 つのデータセットでしか実験を行っていない．プロンプトエンジニアリングや，他の文脈が付与可能なデータセットで実験を行うことで，一般性の確認ができる．今後の研究課題として，他の投稿監視の手段である説明生成や攻撃性除去タスクと文脈考慮の関係性を検証することが挙げられる．攻撃性検出における文脈考慮のさらなる実践的なタスクとして，対象の投稿に付随する文脈だけでなく，投稿のタイムラインに注目してユーザが目にする他の投稿も考慮することが考えられる．

謝辞

本研究は JSPS 科研費 JP21K21343，JP22H00524，JP22K17943,JP21J22383 の支援を受けたものです.

参考文献

[1]Vidgen Bertie, Nguyen Dong, Margetts Helen, Rossini Patricia, and Tromble Rebekah. Introducing CAD: the Contextual Abuse Dataset. In **Proceedings of the 2021 Conference of the NAACL: Human Language Technologies**, pp. 2289–2303, Online, June 2021. ACL.

[2]Enrico Bunde. AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators – A Design Science Approach. In **Proceedings of the 54th Hawaii International Conference on System Sciences**, 2021.

[3]Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In **Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on NLP (Volume 1: Long Papers)**, 2021.

[4]John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. From the Detection of Toxic Spans in Online Discussions to the Analysis of Toxic-to-Civil Transfer. In **Proceedings of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)**, 2022.

[5]Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. KOLD: Korean offensive language dataset. In **In Proceedings of the 2022 Conference on EMNLP**, 2022.

[6]Saveski Martin, Roy Brandon, and Roy Deb. The Structure of Toxic Conversations on Twitter. In **Proceedings of The Web Conference 2021**, WWW '21. ACM, 2021.

[7]Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In **EMNLP-IJCNLP**, 2019.

[8]Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. HOT ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. In **arXiv:2304.10619**, 2023.

[9]Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In **Proceedings of the 17th Conference of the EACL**, pp. 1997–2013, Dubrovnik, Croatia, May 2023. ACL.

[10]Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social Bias Frames: Reasoning about Social and Power Implications of Language. In **proceedings of the 58th Annual Meeting of the ACL**, 2020.

[11]Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. I Feel Offended, Don't be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, 2020.

[12]Akhila Yerukola, Xuhui Zhou, Elizabeth Clark, and Maarten Sap. Don't Take This Out of Context!: On the Need for Contextual Models and Evaluations for Stylistic Rewriting. In **Proceedings of the 2023 Conference on EMNLP**, 2023.

[13]Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study. In **Findings of ACL**, 2022.

[14]Francesco Ventura, Salvatore Greco, Daniele Apiletti, and Tania Cerquitelli. Explaining the Deep Natural Language Processing by Mining Textual Interpretable Features. In **arXiv:2106.06697**, 2021.

[15]Yiming Zhang, Sravani Nanduri, Liwei Jiang, Tongshuang Wu, and Maarten Sap. BiasX: "Thinking Slow" in Toxic Content Moderation with Explanations of Implied Social Biases. In **Proceedings of the 2023 Conference on EMNLP**, 2023.

[16]Zhou Xuhui, Zhu Hao, Yerukola Akhila, Davidson Thomas, Hwang Jena D., Swayamdipta Swabha, and Sap Maarten. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. In **Findings of ACL**, 2023.

[17]Victoria Zayats and Mari Ostendorf. Conversation Modeling on Reddit Using a Graph-Structured LSTM. In **TACL 6**, 2018.

[18]Lei Gao and Ruihong Huang. Detecting Online Hate Speech Using Context Aware Models. In **RANLP 2017**, 2017.

[19]John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity Detection: Does Context Really Matter? In **Proceedings of the 58th Annual Meeting of the ACL**, 2020.

[20]Xinchen Yu, Eduardo Blanco, and Lingzi Hong. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In **Proceedings of the 2022 Conference of the NAACL: Human Language Technologies**, 2022.

[21]Hiren Madhu, Shrey Satapara, Sandip Modha, Thomas Mandl, and Prasenjit Majumder. Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. In **Expert Systems with Applications, 215:119342**, 2023.

[22]Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. Abuse is Contextual, What about NLP? The Role of Context in Abusive Language Annotation and Detection. In **arXiv:2103.14916**, 2021.

[23]Atijit Anuchitanukul, Julia Ive, and Lucia Specia. Revisiting Contextual Toxicity Detection in Conversations. In **arXiv:2111.12447**, 2021.

[24]Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. Context Sensitivity Estimation in Toxicity Detection. In **WOAH 2021**, 2021.

[25]Flor Miriam Plaza del arco, Debora Nozza, and Dirk Hovy. Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech. In **WOAH 2023**, 2023.

[26]Fan Huang, Haewoon Kwak, and Jisun An. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In **arXiv:2302.07736**, 2023.

[27]**Lost in the Middle: How Language Models Use Long Contexts**, 2023.

[28]Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling Large Language Models to Generate Text with Citations. In **Proceedings of the 2023 Conference on EMNLP**, 2023.

[29]Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. ZeroSCROLLS: A Zero-Shot Benchmark for Long Text Understanding. In **Findings of the ACL: EMNLP 2023**, 2023.

[30]OpenAI. ChatGPT: Optimizing Language Model for Dialogue. In **OpenAI blog (2022)**, 2022.

[31]OpenAI. GPT-4 Technical Report. In **arXiv:2303.08774**, 2023.

[32]Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. Towards Identifying Social Bias in Dialog Systems: Framework, Dataset, and Benchmark. In **Findings of EMNLP**, 2022.

[33]Loukas Lefteris, Stogiannidis Ilias, Malakasiotis Prodromos, and Vassos Stavros. Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance. In **Proceedings of the Fifth Workshop on Financial Technology and NLP and the Second Multimodal AI For Financial Forecasting**, 2023.

表 4: 文脈考慮の有無で分類したエラーの具体例．左から行数，検出対象の投稿，付随する対話履歴，人手の正解ラベル，ベースラインモデルの推測ラベル，文脈考慮モデルの推測ラベル．

	検出対象の投稿	対話履歴	正解	ベースライン	文脈考慮
1	User4: *probably they don't know what the n-word is but they don't want to say it. so to avoid saying the n-word they just remove all Ns from words*	User1: *Weird flex but okkk* User2: *Why are they using "m" instead of "n"? I guess every generation has to go to that stage where they add or change letters when they type on the internet, lol.* User3: *I was gonna say "it was probably just a typo" but nope, they literally replaced every 'n' with an 'm'*	0: Non-Abusive	1: Abusive	0: Non-Abusive
2	User3: *Im going to say the m word*	User1: *Weird flex but okkk* User2: *Note: The M thing is an inside joke with this facebook group. N is designated "forbidden letter" for no reason.*	1: Abusive	0: Non-Abusive	1: Abusive
3	User2: *They don't even need to shoot to kill us anymore*	User1: *Killer 5G warning: Expert warns superfast broadband could cause cancer in humans*	0: Non-Abusive	0: Non-Abusive	1: Abusive
4	User3: *Just start speaking Spanish, and go down and get free stuff like everybody else.*	User1: *While California can't help its own citizens with healthcare, it is now going to pay for illegals healthcare* User1: *SS - No laws that I'm aware of paying for Native Californian's healthcare, but illegals are now going to be covered. There is nothing more this country hates than its own people. It can't possibly find the funds to do anything for its own citizens, but seems to have open checkbooks when it comes to illegal citizens.* User2: *Loophole: move to Mexico and become a legal citizen. Illegally return to United States. Live in Cali. Collect free healthcare.*	1: Abusive	1: Abusive	0: Non-Abusive

表 5: 文脈考慮の有無に注目した GPT-4 のエラー分布．

正解	ベースライン	文脈考慮	件数
0: Non-Abusive	1: Abusive	0: Non-Abusive	316
1: Abusive	0: Non-Abusive	1: Abusive	45
0: Non-Abusive	0: Non-Abusive	1: Abusive	120
1: Abusive	1: Abusive	0: Non-Abusive	114

A　エラー分析

GPT-4-turbo のベースラインモデルと文脈考慮モデルのエラー分布を表 5，誤分類の具体例を表 4 に示す．誤分類の定性分析では，ベースラインモデルと文脈考慮モデルについて，片方のモデルが正しく判別できたが，もう一方のモデルが誤った分類（偽陽性または偽陰性）をした４つに場合分けをした．文脈考慮モデルのみが正しく判別できたときには文脈が対象の投稿の情報不足を補えており，ベースラインモデルのみが正しく判別できたときには人間でも判断が難しい投稿が多かった．

A.1　ベースラインモデルの偽陽性

表 5 のように誤分類が最も多かったのは，ベースラインモデルでは偽陽性であったが，文脈考慮モデルでは正しく Non-Abusive と分類できた場合である．表 4 の 1 行目は，このときの具体例である．背景知識として *n-word* が黒人差別の言葉の集合を指すことを踏まえて，対象の投稿単体で判断すると，全ての単語から N を取り除くことは極端な行動であり，皮肉を含む攻撃的な表現と言える．一方文脈を考慮すると，他の投稿の特定の単語内の*"n"*が*"m"*に置き換わっていることがテーマであることが判明し，対象の投稿がその置き換えの理由を説明していると推測できる．このように，対象の投稿のみでは情報不足であり攻撃的に見えるが，文脈を考慮すると単なる応答や説明だと判明することが多い．

A.2　ベースラインモデルの偽陰性

表 4 の 2 行目は，ベースラインモデルでは偽陰性であったが，文脈考慮モデルでは正しく Abusive と分類できた例である．対象の投稿だけでは *m-word* の意味が不明であり，攻撃的表現かどうか判断しきれない．一方，文脈を考慮すると *n-word* の代わりであることがわかり，それを進んで言おうとするのは不謹慎だと判明する．節 A.1 と同様に，情報不足が解消されることによって文脈考慮モデルでは成功している．

A.3　文脈考慮モデルの偽陽性

表 4 の 3 行目は，ベースラインモデルでは正しく Non-Abusive と分類したが，文脈考慮モデルでは偽陽性であった例である．対象の投稿は，「人を殺すのに銃は必要ない」と文脈によって攻撃性の有無が変化する内容である．文脈では，「5G の人間に対する危険性」が訴えられており，これを踏まえると対象の投稿は，5G に対する不安を表していることがわかる．一方で *Killer, cancer, kill* は攻撃性を持ち得る表現であり，文脈考慮モデルが誤った攻撃性を検出した可能性が考えられる．

A.4　文脈考慮モデルの偽陰性

表 4 の 4 行目は，ベースラインモデルでは正しく Abusive と分類できたが，文脈考慮モデルでは偽陰性の例である．対象の投稿単体では，「スペイン語を話し始めるだけで無料でものを入手できる」と，スペイン語を中傷していると捉えられ，悪意が感じられる．文脈を考慮すると，自国の民には厳しく不法入国民には優しいという政府がテーマだと判明する．すると対象の投稿は，政府や不法入国民に対する怒りと取れるが，悲しい現状に対する嘆きや自虐とも取れる．このように判断が難しい場合が多く見られた．