

不法行為としての誹謗中傷検出と検出理由の説明可能性の検証

久田祥平 矢田竣太郎 若宮翔子 荒牧英治

奈良先端科学技術大学院大学

{s-hisada, s-yada, wakamiya, aramaki}@is.naist.jp

概要

オンライン上での誹謗中傷の問題に取り組む自動検出研究において、社会実装の観点で検出理由の説明が求められている。しかし、テキスト生成により説明するアプローチは、説明の論理性に課題があり、社会科学の専門家の介入がなく人間や社会にとって有益な説明を行えているか人手評価が不十分である。本研究は、裁判例に基づく誹謗中傷データセットを用いて、不法行為の観点から、誹謗中傷の分類タスクと、その理由の生成タスクに取り組んだ。そして、生成された理由に、裁判における判断が反映されているか法律の専門家が評価した。結果として、言語モデルを用いた、裁判例に基づく分類タスクの精度には課題があることや、理由の生成では論理的整合性のある回答を得られるが、モデルの学習データセットに含まれる価値観の影響を受け、法律や裁判での判断と異なる解釈を示す事例多くが確認された。本研究は、オンライン上の誹謗中傷の自動検出の社会実装において重要な、法的観点の導入の端緒となる。

本論文では、誹謗中傷の事例を扱う性質上、不快な表現が含まれることにご注意下さい。

1 はじめに

オンライン上での誹謗中傷の問題に取り組む自動検出研究において、コンテンツモデレーションの透明化 [1, 2] や社会的バイアス [3] の軽減のために、検出理由の説明が重要視される。人工知能における説明は、機械学習の推論や動作面について人にとって理解しやすい用語で表現 [4] するタスクとして取り組まれている。その中でも、推論面から誹謗中傷検出の説明に取り組む研究 [5] では、人手による分類における特徴語選択と、分類モデルにとって重要な特徴量との間の整合性について検証している。動作

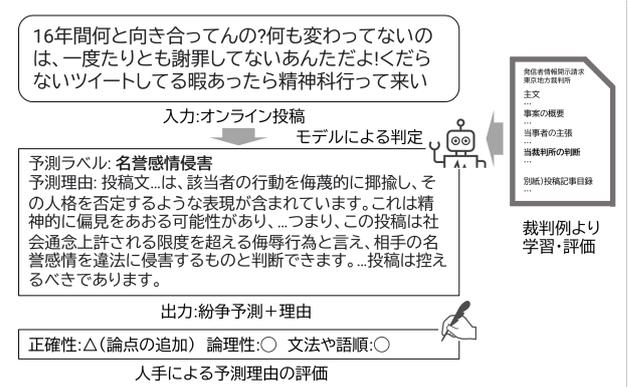


図 1: 研究の概要: 裁判例をもとに、オンライン投稿の不法行為を基準とした誹謗中傷の分類タスクと分類理由の生成タスクに取り組む、法律の専門家による人手評価を行う。

面に着目すると、モデルを分類と同時にその理由説明を生成するように学習する手法 [6, 7] がある。これを利用し、誹謗中傷に存在する社会的バイアスやステレオタイプの説明を目指した研究 [8, 9] では、差別や中傷発言の含意に対する自由記述コーパスを作成し、テキスト生成により誹謗中傷検出の説明を試みている。

しかし、これらの生成アプローチには、社会実装の観点から課題がある。具体的には、誹謗中傷を検出し対処するためには、社会における基準 (日本における違法情報 [10] や EU の Digital Services Act [11]) による誹謗中傷検出と、検出理由の詳細な説明をすべきである。ところが、クラウドソーシングで収集された自由記述コーパスを用いて学習されたモデルの生成文には、説明として論理の飛躍や常識からの隔たりが指摘 [12] されている。加えて、生成された説明の評価には、人手評価と必ずしも相関しないことが議論 [13, 14] されている自動評価 [8, 9, 12] か、専門的観点が必要な評価 [15] しか用いられていない。言い換えれば、社会問題にも関わらず社会科学の専門家の介入がなく、人間・社会の価値観との整合性に関する評価項目の議論や人手評

価が十分でないと考えられる。

本研究は、誹謗中傷による不法行為の分類タスクと、その理由の生成タスクに取り組む。このために、裁判例から作成された誹謗中傷データセットを用いる。実験により、言語モデルが、不法行為を基準とする誹謗中傷の検出（分類）をどの程度行えるか、生成された理由がどの程度論理的で、不法行為の観点に沿った議論を行えているかを法律の専門家が人手で評価する。以上を通じて、本研究では、法的な観点からの紛争予防、さらには紛争解決に向けた、言語モデルによる判断と説明についての可能性と課題を議論する。

2 本研究のアプローチ

2.1 誹謗中傷データセット

裁判例には、オンライン上の発言による誹謗中傷が、どのような権利を侵害する不法行為に該当するかの判断が記載されており、権利侵害の明白性の判断部分で、法律や判例をもとに、権利侵害の内容が論理的に述べられている。現実の問題に近い誹謗中傷検出のタスク設定にするために、裁判例より作成した誹謗中傷データセット [16] から、各事例の本件投稿記事、文脈、権利の種類ラベルを使用する。本件投稿記事と文脈を入力テキストとする。なお、入力テキストで匿名化加工されている箇所は、一律に [IDinfo] に変更した。入力テキストに対するラベルとして、不法行為としての誹謗中傷によって脅かされる権利の種類より、名誉権・名誉感情・営業権と（その他の）人格権・人格的利益を利用する。

誹謗中傷にどのような問題性があるかを理由として利用するために、各本件投稿記事に対応する裁判例より、裁判所による権利侵害の判断部分で、投稿記事による不法行為を議論している文章を抽出する。この作業により、1名のアナテータが、投稿記事、誹謗中傷ラベル、理由の三つ組からなる516件のデータセットを作成した。このうち150件を2名のアナテータが追加検証した結果、理由抽出のアナテータ3者間のペアワイズの一致率の平均は、完全一致率が0.198、ジャカード係数0.674であった。作成したデータセットのラベル分布は、名誉権280件、名誉感情171件、人格権・人格的利益11件、営業権17件（重複あり）である。

本研究は、オンライン上の誹謗中傷の検出を目的とするため、裁判例に含まれていない、裁判に至ら

ないような発言を含めて判定する必要がある。そこで、検出タスクの現実性を高めるため、誹謗中傷データセットにおいて出現回数が多かったプラットフォームの投稿をネガティブサンプルとして用いることとした。具体的には、Twitter、5ちゃんねる、爆サイ（それぞれ事件数は46件、34件、27件）から、それぞれ450件、340件、340件の投稿をランダムサンプリングした。そして、匿名化された誹謗中傷データセットに合わせるために、GiNZAライブラリ [17] で固有表現抽出を行い、人名、住所、都市名、組織名に該当するスパンを [IDinfo] に置換した。

2.2 GPT-4 による要約

解釈の余地を排除するために、裁判例のテキストには冗長な表現が多用されている。学習モデルの入力制限を考慮し、裁判例から抜粋した理由の代わりにGPT-4の要約を学習および評価用の説明として採用する。なお、名誉権の侵害が争われた事例については、抜粋した理由に裁判で議論されるような違法性阻却事由の存在を補足した。付録A.3.1に要約に用いたプロンプトと名誉権の侵害事例の補足を示す。この結果、入力テキストは平均395.8文字から平均103.1文字に短縮された。表4に元の理由の抜粋とその要約を示す。要約の忠実性に関する人手による評価結果を表3に示す。正確さに課題はあるものの、概ね元の裁判例の抜粋に忠実な要約を実現している。この評価項目については、生成した分類理由の評価と共に後述（2.4節）する。

2.3 実験設定

本研究では、大規模言語モデルを用いて、4つの誹謗中傷ラベルに対する分類タスクと、分類の理由の生成タスクに取り組む。

このタスクに、2つのLlama2モデルyouri-7b-instruction(以下、Youri7B) [18]とSwallow-13b-instruct-hf(以下、Swallow13B)とGPTモデルllm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0(以下、LLMJP13B)を用いる。モデルはPEFTパッケージ [19]のLoRA [20]によりファインチューニングする。ベースラインとして、BERTモデルのcl-tohoku/bert-base-japanese-v3による線形分類とT5モデルsonoisa/t5-base-japaneseによる分類と理由の生成により、分類タスクにおける性能を比較する。評価指標には、Accuracy, Macro-F1とMicro-F1を用いる。

表 1: 誹謗中傷ラベルの分類精度：各モデルの生成による分類の 3 回平均。

Model	Ave Score				ラベル別 F1			
	Accuracy	Macro-F1	Micro-F1	Micro-F1 SD	名誉権	名誉感情	人格権	営業権
BERT	0.858	0.537	0.673	-	0.711	0.634	0.222	0.580
T5	0.850	0.450	0.697	-	0.734	0.666	0.0	0.400
Swallow13B	0.843	0.440	0.640	± 0.0019	0.739	0.520	0.0	0.500
Youri7B	0.841	0.433	0.655	± 0.0106	0.717	0.594	0.0	0.422
LLMJ13B	0.811	0.429	0.621	± 0.0160	0.687	0.577	0.133	0.317

表 2: Youri7B による分類と分類の理由：アノテータは、正確さ「1: 反映しているが誤解を生む可能性」、流暢さと論理性「0: 問題なし」、正確さのエラーは「不要な追加・省略の問題」「モデル由来の不正確さ」と評価

本件投稿記事	16 年間何と向き合ってるの?何も変わってないのは、一度たりとも謝罪してないあんただよ!くだらないツイートしてる暇あったら精神科行って来い
分類	<名誉感情の侵害>
分類の理由	投稿文「16 年間何と向き合ってるの?何も変わってないのは、一度たりとも謝罪してないあんただよ!くだらないツイートしてる暇あったら精神科行って来い」は、該当者の行動を侮蔑的に揶揄し、その人格を否定するような表現が含まれています。これは精神的に弱っている人に対する無理解や偏見をおおる可能性があり、その結果、その人の名誉感情を侵害する可能性があります。つまり、この投稿は社会通念上許される限度を超える侮辱行為と言え、相手の名誉感情を違法に侵害するものと判断できます。したがって、投稿者は他人の名誉感情に配慮するべきであり、このような侮辱的な投稿は控えるべきであります。

2.4 生成文の評価

モデルが生成した理由について、原文の裁判例の判断を反映しているか、論理的な整合性があるかを評価する。そのために、テキスト生成タスク [13, 21] で重視される正確さと流暢さに論理性を加えた 3 つの評価項目を用いる。

正確さ 裁判例の抜粋と要約 (GPT-4 の場合) や生成された分類の理由を比較し、裁判で議論されるような誹謗中傷の不法行為による問題点を、適切な用語を用いて指摘できているか評価する。具体的には、「裁判例の判断を正確に反映していますか?」という質問に対して、「0: 問題なし」「1: 反映しているが少し誤解を生む可能性がある」「2: 反映していない」「3: 反映していない上に、致命的な誤解を生む可能性がある」という 4 段階のリッカート尺度を用いて回答する。また、不正確の場合、次の 4 つのエラータイプに該当するかをチェックする。

不要な追加・省略の問題 不要・冗長なテキストが含まれている。または、裁判例中の重要な論点が欠落している。

出典の不正確さ 裁判や法律、裁判例中の用語

や概念が誤って表現されている。

モデル由来の不正確さ 裁判例に示されていない内容や、勝手な論点の追加など無関係なテキストがあり、裁判例と事実が異なる。

結論の矛盾 裁判例と結論が矛盾している。

流暢さ 語順、文法や単語など日本語として自然さを「0: 問題なし」から「3: 文章として重大な破綻している」の 4 段階のリッカート尺度で評価する。

論理性 データセットの分量や事前学習データの制約により、裁判例の判断を反映した正確な理由を述べることは、容易ではないと考えられる。モデルが生成した理由は、予測したラベルに対する、論理的な評価をできているか「0: 問題なし」から「3: 論理的な指摘を行ってない」の 4 段階のリッカート尺度の評価とそのエラータイプのチェックを行う。

入力の不認識 入力となる投稿テキストに対する、問題点を指摘できていない。

タスクの不認識 理由を述べる文としての体をなしていない。

予測の矛盾 予測したラベルに対して、予測を正当化する理由を述べていない。

論理の不整合 取り上げている問題の原因と、

表 3: 生成した分類の理由の評価：2 名のアノテータによる 31 件ずつの評価で、Ave は 0~3 のリッカート尺度の評価の平均、エラータイプは 31 件中の該当エラー数を示す。ただし、IAA では、Ave は相関を、エラータイプは一致率を示す。

評価項目	正確さの問題					流暢さの問題		論理性の問題			
	Ave↓	追加等↓	出典↓	モデル↓	結論↓	Ave↓	Ave↓	入力↓	タスク↓	予測↓	論理↓
Swallow13B	1.355	21	3	18.5	7	0.355	0.129	0	0	0	3
Youri7B	1.661	22.5	5	20.5	8.5	0.774	0.403	1	0.5	4	5
LLMJ13B	1.339	23.5	7.5	18.5	6	0.129	0.290	1.5	1	3.5	5
GPT-4 要約	0.419	7.5	2.5	7.5	0	0.065	0.016	0.5	0	0	0
IAA	0.716	0.653	0.775	0.565	0.880	0.588	0.343	0.968	0.790	0.992	0.927

指摘する理由が論理的に整合していない。

法科大学院の学生 2 名に要約・生成した分類の理由をそれぞれ 31 件の評価を依頼した。合計の作業時間は 5 時間で、法学の専門知識を前提とする作業のため、謝礼として時給 2,000 円を設定した。アノテータ間合意度 (IAA; Inter-Annotator Agreement) を表 3 に示す。正確さ、流暢さ、論理性の全体評価には相関を、エラータイプ別の評価には一致率を用いる。専門的観点について尋ねる正確さの全体評価では、高い相関を示す一方、日本語の評価や論理性の全体評価には相違が見られた。

3 結果と考察

3.1 分類性能

表 1 は分類タスクの結果を示す。分類の評価にはラベルの生成を 3 回行った平均を用い、分類の理由の評価には各モデルの Micro-F1 が中間の事例の生成結果を利用した。LoRA で学習したモデルはいずれも、フルパラメータを更新する教師学習ありモデルの性能を超えることができていない。本タスクは、様々な要素を考慮した上で判断される裁判例からの事例を用いたことや、サンプル数の少ない人権侵害の類型もラベルとして取り入れたことで、モデルによる判別が難しいタスクになっている。今回の設定における Micro-F1 の結果で比較すると、IAA で Offensive language の分類を評価した先行研究 [22] が示す、IAA の低い事例における分類性能 (F1-score=0.656) と近い結果となった。

3.2 生成された理由の文の評価

生成された理由に対する人手評価の結果を表 3 に示す。裁判では議論されないような観点の言及が多

さにより、モデル由来の不正確さのエラーの数に反映されている。日本語訳した FLAN [23] や Human preference data about helpfulness and harmlessness [24] のような、不快な発言を避けるようなデータセットを用いてインストラクションチューニングが行われたことで、表 2 の分類の理由の例のように、不快な生成を避ける価値観が、学習した誹謗中傷に対する裁判の判断と競合している部分が存在することが示唆される。正確さについては課題もある一方で、与えられたタスクに対して、論理的な解釈を示す能力が示された。正確さを欠くことで、hallucination の危険性を同時に孕んでいるものの、論理的な判断能力が示されたことで、学習用の論理的なフレームワークを入力として用いることができれば、大規模言語モデルは正確に法律用語や判断枠組みを理解できる可能性がある。

4 おわりに

本研究は、裁判例をもとに、誹謗中傷の分類タスクと裁判所の判断を述べたテキストを用いて、分類の理由の生成タスクに取り組み、性能の評価を行った。その結果、分類タスクでは、タスク設定の難しさによって分類精度に課題を残す結果となった。分類の理由の生成においては、論理的な能力を見せる一方で、インストラクションチューニングデータセットに含まれる価値観の影響を受けることで、裁判における判断やそれに付随する法律の枠組みを正確に反映することの難しさを示した。

この課題の克服には、利用可能な裁判例の拡充や、裁判例や法学における推論過程を、忠実に再現した学習・評価用のフレームワークの作成が必須であり、人工知能研究者と法学研究者の協業は不可欠と考えられる。

謝辞

本研究は、JSPS 科研費 JP22J23161, JP22K12041 および JST SICORP JPMJSC2107, 戦略的イノベーション創造プログラム (SIP3) の支援を受けたものである。

参考文献

- [1] Eric Goldman Daphne Keller Olivier Sylvain Evelyn Mary Aswad, Danielle Citron. Free Expression, Harmful Speech and Censorship in a Digital World, 2020. [Accessed: 11 January, 2024].
- [2] Chiara Drolsbach and Nicolas Pröllochs. Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database. **arXiv preprint arXiv:2312.04431**, 2023.
- [3] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In **Proc. of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 1668–1678, 2019.
- [4] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. **ACM Comput. Surv.**, Vol. 55, No. 13s, July 2023.
- [5] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In **Proc. of the AACL Conference on Artificial Intelligence**, Vol. 35, pp. 14867–14875, 2021.
- [6] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing Neural Predictions. In **Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 107–117, 2016.
- [7] Hui Liu, Qingyu Yin, and William Yang Wang. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. In **Proc. of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5570–5581, 2019.
- [8] 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 **Proc. of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5477–5490, 2020.
- [9] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In **Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 345–363, 2021.
- [10] 資料2 ネット上の違法・有害情報に対する総務省の取組【総務省】. [Accessed: 11 January, 2024].
- [11] Overview of the digital services act. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en. [Accessed: 11 January, 2024].
- [12] Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 5490–5505, 2023.
- [13] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Kraemer. Best practices for the human evaluation of automatically generated text. In **Proc. of the 12th International Conference on Natural Language Generation**, pp. 355–368, 2019.
- [14] Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. How to do human evaluation: A brief introduction to user studies in NLP. **Natural Language Engineering**, Vol. 29, No. 5, p. 1199–1222, 2023.
- [15] Fan Huang, Haewoon Kwak, and Jisun An. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In **Companion Proceedings of the ACM Web Conference 2023, WWW ’23**. ACM, April 2023.
- [16] 久田祥平, 若宮翔子, 荒牧英治. 権利侵害と不快さの間: 日本語人権侵害表現データセット. 言語処理学会第29回年次大会, 2023.
- [17] 松田寛. GiNZA - Universal Dependencies による実用的日本語解析. 自然言語処理, Vol. 27, No. 3, pp. 695–701, 2020.
- [18] Tianyu Zhao and Kei Sawada. rinna/youri-7b-instruction.
- [19] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. <https://github.com/huggingface/peft>, 2022.
- [20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **Proc. of the International Conference on Learning Representations**, 2022.
- [21] Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. What have we achieved on text summarization? In **Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 446–469, 2020.
- [22] Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators’ Disagreement. In **Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 10528–10539, 2021.
- [23] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In **Proc. of the International Conference on Learning Representations**, 2022.
- [24] Yuntao Bai et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. **arXiv preprint arXiv:2204.05862**, 2022.

A 実験の詳細設定

A.1 学習時

`lora_config` `r=64, lora_alpha=64, target_module=q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, down_proj, lora_dropout=0.05, bias=None` `training_hyperparameter` `lr=5e-5, scheduler=linear, optimizer=adamW, epoch=50, batch_size=2`

A.2 生成時のハイパーパラメータ

Transformers の `AutomodelForCausalLLM` の Generation メソッドを利用。 `do_sample=True, temperature=0.9, top_p=0.95, max_new_tokens=400`

A.3 プロンプト

A.3.1 GPT4の要約に用いたプロンプト

ポジティブサンプル（裁判例から作成した事例）

インターネット“投稿”に対する人権の観点からの問題“ラベル”と、その“理由文”を与えます。投稿文中の問題点について、その問題となる表現スパンを「」で引用しつつ、理由文をもとに2文程度で議論してください。なお、[IDinfo] は人名や企業名、詳細な地名に対するマスク処理を行っています。投稿：「`{text}`」ラベル：`{label}`理由：「`{reason}`」

名誉権の侵害事例における理由への補足文

しかし、名誉権の侵害（名誉毀損）については、公共の利害に関する事実に係り、公益を図る目的を有するものであって、かつ、摘示事実がその重要な部分について真実であることが証明されたや摘示事実を真実と信ずるについて相当の理由があったとき、違法性が阻却される場合がある。

ネガティブサンプル（収集したオンライン投稿）

インターネット“投稿”に対する人権の観点からの問題“ラベル”と、その“理由文”を与えます。投稿文中の問題点について、理由文をもとに1文程度で教えてください。なお [IDinfo] は人名や企業名、詳細な地名に対するマスク処理を行っています。投稿：「`{text}`」ラベル：`{label}`理由：「人権侵害を認められた事例ではないので、一般論として問題ある表現かどうか1文程度で議論してください。」

A.3.2 ファインチューニングに用いたプロンプト

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。
\\n\\n### 指示:\\n 今から本文とその文脈を与えます。本文が誹謗中傷に当たるなら、ラベル <名誉権><名誉感情><人格権・人格的利益><営業権> から2つまで選び、そうでないならラベル <侵害なし> と選んだ上で、その理由を教えてください。尚、[IDinfo] は匿名化加工トークンです。
\\n\\n ### 入力: 本文 `{text}` 文脈:`{context}`\\n\\n ### 応答: ラベル `{label}` 理由:`{summarizedreason}` \\n

表 4: 裁判所の判断の抜粋と GPT4 による要約: 令和1年(ワ)第34145号本件記事1-☆☆☆より、事例は名誉権の侵害を認めているが、違法性阻却事由の存在より違法性は認められていない。

投稿記事	[IDinfo] は点滴もケチって輸液量を半分にしたんだって。こんやつはどぎゃんもこがやんもならん男ばい
------	--

裁判所の判断の抜粋	本件記事は、「[IDinfo] は点滴もケチって輸液量を半分にしていたんだって。こんやつはどぎゃんもこがやんもならん男ばい。」という内容である。これを一般の読者の普通の注意と読み方を基準として読むと、本件記事1-☆☆☆は、原告法人の理事長であり、原告法人が経営する [IDinfo] 医院の院長である原告 q1 が、費用削減のため、患者に対する輸液につき、本来必要な量の半分の輸液剤しか輸液しなかった旨の事実を摘示し、原告 q1 について、どうしようもない男であるという意見ないし論評をするものと認めるのが相当であり、これらの摘示事実及び意見ないし論評が少なくとも原告 q1 の、ひいては原告 q1 が理事長を務める原告法人の社会的評価を低下させるものであることは、明らかである。（後略）
-----------	--

GPT-4 による要約	投稿は、特定の人物が費用削減のために患者への輸液を適切な量よりも少なく施しているという事実を述べ、その人物を否定的に評価する内容であり、これはその人物の名誉を傷つけるものであるといえるため、名誉権の侵害が問題となります。ただし、投稿の内容に公益などのある目的がある場合や、事実が真実であることが証明された場合は、名誉権侵害とは認められない場合もあります。
-------------	---
