

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

久田祥平 若宮翔子 荒牧英治

奈良先端科学技術大学院大学

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

概要

ネット上の誹謗中傷は大きな社会問題になっており、検出タスクやデータセット構築などの研究が行われている。既存研究は、誹謗中傷の主観的な性質のため、読み手の感性としての主観評価を受け入れた分類または、誹謗中傷をヘイトスピーチのようなより細かいサブタスクに分割することで、課題を具体化した、明確な基準による分類に基づいている。このため、ヘイトスピーチのような関心度の高いサブタスクを除き、社会問題に至る言語行為と、既存の誹謗中傷へのアプローチにはやや隔りがある。本研究では、誹謗中傷を人権侵害という観点から捉えることで、その専門家である法曹による判断を利用することができると考え、判例から、人権侵害が争われた言語表現のデータセットを作成し、<https://zenodo.org/record/7481459> で公開を開始している。これにより、人権侵害の言語的な特徴や、既存研究における誹謗中傷との相違点を明らかにすることが期待される。

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

インターネットが言論空間として台頭するにつれて、ソーシャルメディアや匿名掲示板を介した誹謗中傷は、様々な人権侵害を引き起こしている。この問題に対応するため、検出手法の提案や、そのリソースとして OLID [1] や HateSpeech18 [2] などのデータセットの構築が行われている。誹謗中傷は、主観的な性質の高いテキストである [3, 4] ために、一般化された客観的な基準を作成することが難しい [5]。そのため、既存のデータセットは、2つの方向性でアノテーションが実施されている。1つ目は、Toxicity [6] や一部のヘイトスピーチデータセッ

ト [7] のように、主観的な評価を認めるものである。2つ目は、誹謗中傷に関わる類型やサブタスクが整理された後 [4, 8]、ヘイトスピーチ [9, 2] や Offensive language の宛先 [1] といった、サブタスク毎に問題を具体化し、明確な基準を提案することで、検出モデルなどの応用面でも扱いやすいものになっている。

しかし、既存研究が定義してきた誹謗中傷と人権侵害のような社会で問題とされる言語表現行為の隔りがあると考えられ、その違いに対する理解には専門的な判断を利用、比較する必要がある。ロシアにおける侮辱に取り組んだ研究 [10] は、既存研究における侮辱とその法的規制の対象となる言動としての侮辱の言動には隔りがあることを指摘し、専門知を取り込んだ検出手法の提案した。ロシア連邦法 5.61 条に基づいて、侮辱を 1) 罵詈雑言や卑猥な語彙を含むこと、2) 発言の方向がわかること、3) 不道徳や公的に非難される行動を示すような行動の表現を含むことと定義し、法律の条文に基づく検出モデルを検証したが、データセット作成過程で 3) の形式化が難しく、アノテーター間での議論の余地が残ることを報告している。誹謗中傷対策のコンテンツモデレーションの基準にも、現地法やコミュニティーガイドラインなどの根拠、言語的な線引の明確化や、文化的背景を考慮することが求められている [11, 12] 中では、誹謗中傷に対する専門家の判断を理解することは重要である。

本研究は、ネット上での誹謗中傷を現実の人権侵害の問題として扱うことで、その専門家である法曹による判断である判例を利用する。インターネット上の人権侵害の判例上の定義について議論した後、具体的な言語的な特徴を理解することに向けた、インターネット上の人権侵害を争われた言語表現のデータセットを構築する。具体的には、判例から、人権侵害の有無が争われたテキスト、侵害の判断に利用された前後のテキスト、どの権利の侵害が主張されたか、その侵害は認められたか、といった 6 項目の情報を抽出する。このデータセットによ

り、言語表現における人権侵害の分離境界面の定量的な分析や、既存の誹謗中傷の類型論や自動検出比較を可能とし、誹謗中傷問題へのより深い理解が期待される。

2 人権侵害に至る誹謗中傷と基準

今回のデータセットは、人権侵害のうち、誹謗中傷による人格権の侵害、中でも主に名誉権・名誉感情の侵害を取り扱う。データセットの作成過程では、上記の以外的人格権侵害事例も含まれている。本章では、名誉権・名誉感情への侵害が、制度上どのように理解されているかを整理し、その質的な判断の基準について議論する。

2.1 名誉権

名誉権の侵害（名誉毀損）は、問題とされる表現が、人の品性、徳行、名声、信用等の社会から受ける客観的評価を低下させる、事実を摘示するもの、又は意見ないし論評を表明するもの [13] とされる。事実の摘示とは、特定の行為又は具体的事実を、叙述することである。それが公共性や公共目的、真実性といった違法性阻却事由がなく、人の社会的評価の低下をもたらす場合、或いは、ある真実を基礎とした意見ないし論評の表明であるが、人身攻撃に及ぶなど意見ないし論評としての域を逸脱したものの場合、名誉権侵害は成立するとされる。今回のデータセットでは、公共性や真実といった違法性阻却事由は、他の証拠などをもって証明されるものを扱うことができないため、どのような表現が、表 1 の事例のように、事実の摘示として人の社会的評価が下

表 1: 本データセットで扱う名誉権侵害事例と非侵害事例：前者は集団虐めを主導する人物であるとの事実の摘示が認められた一方、後者は具体的事実を摘示するものではなく、その批評も社会通念を超えた不適切なものでないと判断された。

名誉権侵害事例 [14]	大人なのに集団イジメのリーダーとかマジ痛いし、言ってることも中二病過ぎてキモい、親に甘んじられて実家の部屋で>>■悪口言ってたんだ、
名誉権非侵害事例 [15]	すぐ外誘ってお金を使わせようとする悪い癖があるよな 相手への思いやりや配慮が全くないから、別にもう会いたくない

げるか、そうでないなら、通常の見解や論評を逸脱したものかについて、判例にあるような「一般の閲覧者の普通の注意と読み方を基準とした」表現の基準を探ることが目的となる。

既存の Abusive language 研究との対比では、事実の摘示は、必ずしも侮辱語や汚い言葉を使っているとは限らないため、問題が重複しない部分が出てくる。その中でも、企業の評判を扱ったもの [16] や、“calumny”（中傷や誣告）をラベルとして扱った研究 [17] は名誉権侵害と一部問題が被っていると考えられる。

2.2 名誉感情

名誉感情とは、人が自己自身の人格的価値について有する主観的な評価である [18]。実際には、客観的な評価が求められ、その侵害の有無は「社会通念上許される限度を超える侮辱行為であると認められる」 [19] ことと判断され、文言自体の侮辱性の程度、根拠の示されていない意見ないし感想かどうか、侮辱文言の数、表現の明確性といった要因がその都度考慮されると指摘されている [20]。

既存研究 [22, 1] における侮辱 (insult) は、宛先のある罵倒的、下劣、または不快な言葉を含む表現と定義される。一方、名誉感情侵害では、表 2 の侵害事例のような差別的な表現だけでなく、表 1 の侵害

表 2: 本データセットで扱う名誉感情侵害事例と非侵害事例：前者は頭脳が健常ではないことを揶揄する表現は一種の差別的表現であり、侮辱と解された一方、後者は批判的又は消極的意見の一つとして、社会的通念上許される限度を超える侮辱行為とまでは認め難いと判断された。

名誉感情侵害事例 [14]	D の馬鹿さは異常だと思う 落ち着き始めた揉め事、わざわざ煽って大きくしなくてもw おまえの信者も燃えてるけど大丈夫？（頭）
名誉感情非侵害事例 [21]	個人的意見ですけど、P 4 さんお顔は綺麗というか好きな方なので露出方向に走ったり、汚いコスプレしなければもっとマシだったのにと残念な気持ちでいっぱいになります・・・その辺彼女さんの自由ですけど、もう少し人を不快にさせないような活動やキャラレイプしなければいいのに・・・

1:本件投稿	事実婚で籍は入っていないから言うたら破局ってとこかね。あんなに口うるさく重箱に隅をつつくような粗探しされたら嫌になるだろう。顔は般若性格は腹黒金に汚くいねには弱いそりゃ性交も子作り以外では無理ー！ってなるな。
2:文脈	あのフラッシュの対談読んだら、離婚秒読みかなーって思っちゃう。そして[P1]さんは自分の部屋が汚い自覚、まるでないね。汚って話をしているのに、ウォークインクローゼットが[P2]さんのものばかりとか話しはじめて論点ずらし？
3-1a:人格権の種類1	名誉感情
3-2b:類型1の判断	1(侵害が認められる)
3-2a:人格権の種類2	名誉権
3-2b:類型2の判断	0(侵害するとは認められない)
4:事件番号	東京地方裁判所令和3年(ワ)第25111号
5:文献番号	LEX/DB 文献番号25604944
6:プラットフォーム	ガールズちゃんねる

図 1: データセットラベルの一例

事例の「マジ痛い」「中二病過ぎてキモい」も強く生理的嫌悪感を表現するものとして、通常の社会生活において投げかけられることはめったにない強い侮辱侵害として判断された。これに対して、表 2 の非侵害事例のように、嫌悪感を伴う消極的意味の言葉の使用が、批判や消極的な意見の 1 つとして、社会通念上許される程度の範疇であると判断されているなど、単純に基準の解釈ができるものではない。「社会通念上許される限度」という基準を探り、侮辱表現の判断の傾向に対する理解支援が、本研究の目的の一つである。

3 日本語人権侵害表現データセット

本研究では、テキストによる誹謗中傷に対する裁判における人権侵害の有無の判断に基づき、日本語人権侵害表現データセットを作成する。図 1 は、その一例を示す。

3.1 作成手順と項目

日本語人権侵害表現データセットの作成手順を説明する。調査対象は、民事事件の発信者情報開示請求事件または、損害賠償等請求事件のうちインターネット上の投稿による事件である。アノテーターは、これらの裁判例から、次の 6 項目に該当する文章の抽出やラベルの付与を行う。

1: 本件投稿

本件投稿は、判例の投稿記事目録や事実・理由部分から、権利侵害の有無が争われた表現を匿名化してタグを付与し、書き起こしたテキストである。もとの投稿記事目録や事実・理由部分では、原告に関わるような、個人情報と考えられるテキストが、アルファベット、記号、黒塗り等で置き換えられている。この匿名化済みの記述に対して、個人の身元証明に関わる記述には [P1]、企業や団体の身元証明に関わる記述には [G1]、商品や制作物に関する記述には [M1]、住所など個人情報に関わる部分には [INFO1] のタグを、それぞれ同一人物や同一表現がわかるように番号分けて付与する。本件投稿中に URL が含まれる場合は、URL を [URL] タグに置換している。

2: 文脈 (本件投稿に関わる他の投稿)

本件投稿への判断に考慮されたスレッドタイトル・会話・前後の投稿を判例で引用している場合がある。本件投稿に関わる他の投稿は、そのようなテキストを書き起こしたものである。より問題の正しい理解や精度の高い学習や分類以外のタスクにおいても、利用可能な文脈情報を付与する。本件投稿と同様の方法で匿名化処理を施している。

3: 人格権侵害

本件投稿が、どの人格権の侵害を含んでいるか、問題が争われた権利の種類とその判断についてラベルで示したものである。

3-1a: 人格権の種類 1

人格権の種類 1 は、本件投稿がどの権利に対する侵害であるか、裁判で争われた点についてラベルで示したものである。裁判においては、問題とされる投稿によって侵害されている権利を原告側が訴えるものであり、裁判所は、それぞれ訴えについて判断を下す。1 つの本件投稿に対して、複数の権利侵害の訴えが存在する場合、侵害が認められた権利を含む最大 2 つを抽出する。判例文中で、本件投稿に対して「社会的評価」、「名誉権」といった文言を含む判断が行われている場合は名誉権、「社会通念上許容される限度を超えた侮辱」「名誉感情」という文言を含む判断が行われている場合は名誉感情、「プライバシー」という文言を含む判断が行われている場合はプライバシー権を扱い、上記の 3 項目に該当しないが「人格権」という文言を含む場合は、その他の人格権をラベルとして選択する。加えて裁判例の中には、本件投稿の表現は原告を標的としていると推認出来ないと判断されるものなど、人格権と

いった権利侵害についての判断が下される前に、主張が棄却された事例の場合、None を付与する。

3-1b: 類型1の判断

「原告の名誉感情を違法に侵害するものである。」のように、人格権等に対する侵害を認める表現が含まれる場合、本件投稿による類型1の権利侵害が認められた事例として1を、「社会通念上許される限度を超えて原告を侮辱するものとは評価できない。」のように本件投稿が人格権等に対する侵害が認められると言えないと判断される表現が含まれる場合、権利侵害が否定された事例として0を付与する。

3-2a, 2b: 人格権の類型2, 類型2の判断

複数の権利について裁判で争われた場合、類型1の場合と同様に2つ目の項目の類型を選択し、判断についてラベルを付与する。

4: 事件番号

本件投稿の引用元となった事件番号と争われた裁判所である。

5: 文献番号

本件投稿の引用元となった文献番号である。

6: プラットフォーム

事件の本件投稿が発生したコンテンツプラットフォームが明記されている場合、記載する。

3.2 統計量

日本語誹謗中傷判例データセットは、132件の事件に対する442件の投稿を抽出して構築された。ラベルの分布を付録の表3に示す。なお、判例集データへのアクセスライセンスの都合上、1名のアノテーターがアノテーションを実施した。

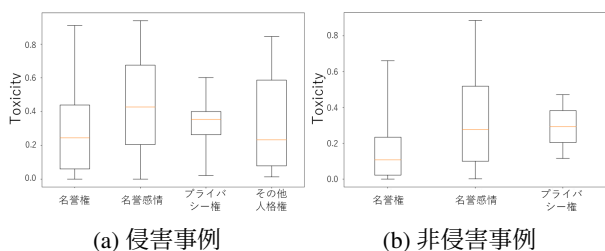


図2: Toxicity 値の箱ひげ図: 侵害事例における平均は名誉権 0.285, 名誉感情 0.429, プライバシー権 0.317, その他人格権 0.346. 非侵害事例における平均は名誉権 0.165, 名誉感情 0.306, プライバシー権 0.293.

3.3 Toxicity との比較

通常、類似のデータセットとの比較を行うのが一般的だが、誹謗中傷に関連する日本語で公開されているデータセットを認知していない。本稿では、構築したデータからラベル別に、誹謗中傷等の代表的な分類モデルの1つである Perspective API [23] による Toxicity と比較することで、既存の誹謗中傷検出関連モデルとの比較を行う。Toxicity はその定義より、判断を読み手に委ねる主観的な指標と考えられる。図2aは、判例において、ある投稿の権利侵害を認めた事例における Toxicity 値の箱ひげ図である。名誉感情侵害表現へのスコアが高く出る傾向は、テキストから受ける主観的な不快さのある表現と、重複する部分がある示されている一方で、名誉権侵害のように批判的な表現の意味への理解が必要な事例では検出の難しいことがわかる。一方、図2bが示すように、非侵害事例においては、平均的には侵害事例よりも低い Toxicity を示している。しかし、一部事例では、Toxicity の高い表現であっても、権利侵害とは認められない事例が存在する。既存研究の手法では、捉えきれない誹謗中傷の問題の存在が示唆される。事例を交えた Toxicity の比較は中傷表現を含むため付録にて示す。

4 おわりに

本研究では誹謗中傷について人権侵害という観点を取り入れて、専門家の判断を理解するためのデータセットを提案した。テキスト表現よりどのような人権侵害を引き起こすかについてや、その判断の分かれる表現の境界面についての、より詳細な分析を通じて、複雑な言語行為である誹謗中傷の深い理解が期待できる。一方で、今回は、その扱うテキストの問題で、1名のアノテーターによる442件の投稿へのアノテーションにとどまっている。また、判例中の判断内容をより紐解いていくと、同定可能性、違法性阻却事由、そして誹謗中傷を受けている人の属性といった項目も、実際では重要な要素であるが、今回十分には考慮できていない。言語的な特徴から問題性への判断を理解する上で、判例中よりこの論点を抽出できるようなアノテーションのスキームの提案も必要である。今後も量・質の面で判例を利用したデータセットを拡張することで、誹謗中傷について理解する上で、表現だけでなく、文脈を踏まえた意味といった点に注目できる。

謝辞

本研究の一部は、JSPS 科研費 JP22J23161, JP22K12041, JP19H01118 および JST SICORP JP-MJSC2107 の支援を受けたものである。

参考文献

- [1] 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 **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 1415–1420, Minneapolis, Minnesota, June 2019.
- [2] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate Speech Dataset from a White Supremacy Forum. In **Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)**, pp. 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics.
- [3] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. **PLoS One**, Vol. 15, No. 12, p. e0243300, December 2020.
- [4] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. In **Proceedings of the First Workshop on Abusive Language Online**, pp. 78–84, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics.
- [5] Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. Challenges and frontiers in abusive content detection. In **Proceedings of the Third Workshop on Abusive Language Online**, pp. 80–93, Florence, Italy, August 2019. Association for Computational Linguistics.
- [6] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In **Proceedings of the 26th International Conference on World Wide Web**, WWW '17, pp. 1391–1399, Republic and Canton of Geneva, CHE, April 2017. International World Wide Web Conferences Steering Committee.
- [7] Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. A hierarchically-labeled Portuguese hate speech dataset. In **Proceedings of the Third Workshop on Abusive Language Online**, pp. 94–104, Florence, Italy, August 2019. Association for Computational Linguistics.
- [8] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In **Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media**, pp. 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [9] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. **ICWSM**, Vol. 11, No. 1, pp. 512–515, May 2017.
- [10] Liliya Komalova, Anna Glazkova, Dmitry Morozov, Rostislav Epifanov, Leonid Motovskikh, and Ekaterina Mayorova. Automated classification of potentially insulting speech acts on social network sites. In **Digital Transformation and Global Society**, pp. 365–374, 2022.
- [11] Jillian C. York and Corynne Mcsherry. Content moderation is broken. let us count the ways. <https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways>. [accessed on December 31th, 2022].
- [12] The santa clara principles on transparency and accountability in content moderation. <https://santaclaraprinciples.org/>. [accessed on December 31th, 2022].
- [13] 最高裁判所第三小法廷平成 6 年 (オ) 第 978 号. LEX/DB 文献番号 28021760.
- [14] 東京地方裁判所令和 3 年 (ワ) 第 22400 号. LEX/DB 文献番号 25604264.
- [15] 東京地方裁判所令和 3 年 (ワ) 第 29589 号. LEX/DB 文献番号 25604756.
- [16] Nikolay Babakov, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. Detecting inappropriate messages on sensitive topics that could harm a company’s reputation. In **Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing**, pp. 26–36, Kiyv, Ukraine, April 2021. Association for Computational Linguistics.
- [17] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deeper attention to abusive user content moderation. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 1125–1135, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [18] 最高裁判所平成 21 年 (受) 第 609 号 民集第 64 卷 3 号 758 頁.
- [19] 最高裁判所第三小法廷平成 21 年 (受) 第 609 号. LEX/DB 文献番号 25442100.
- [20] 公益社団法人商事法務研究会. インターネット上の誹謗中傷をめぐる法的問題に関する有識者検討会. https://www.soumu.go.jp/main_content/000818861.pdf. [accessed on December 31th, 2022].
- [21] 東京高等裁判所令和 2 年 (ネ) 第 1555 号. LEX/DB 文献番号 25567236.
- [22] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. Detection and Fine-Grained classification of cyberbullying events. In **Proceedings of the International Conference Recent Advances in Natural Language Processing**, pp. 672–680, Hissar, Bulgaria, September 2015. INCOMA Ltd. Shoumen, BULGARIA.
- [23] Jigsaw. Perspective api. <https://perspectiveapi.com/>. [accessed on December 31th, 2022].

A 日本語データセットの分析

投稿コンテンツ

図3に、人権侵害と判断された事例における、本件投稿テキストに基づき作成されたワードクラウドを示す。名誉権侵害の場合、「自分」や「会社」といった出現頻度の高い表現の次には、犯罪行為に関連する単語が様々出現している。このような犯罪行為を糾弾することが、事実の摘示と判断される事件が起きる傾向を示している。名誉感情侵害の場合、「笑」や「www」と言った表現とともに、多様な侮辱語の表現が出現している。このように嘲笑する形で、他者を侮辱するような表現が事件として判断される傾向を示している。



(a) 名誉権侵害

(b) 名誉感情侵害

図3: 人権侵害表現におけるワードクラウド

データセット中の人権侵害表現

権利侵害あり

例1)「辞めたくても契約違反の条件が厳しくとことん追い込んでむしり取る。」フランチャイズに対する高圧的な契約内容の存在を掲示板に書き込んだ事例。契約関係からの脱退を申し出た加盟店に対し、追い込んで財産を奪い取るとの事実を摘示するものであり、社会的評価を下げると判断されるような表現である。このような、真贋は考慮しなくても、不正な事実描写に対しては Toxicity は上手に扱えておらず 0.00653 と非常に低い値である。このような、侮辱語の少ない事実の摘示に対しては、フェイクニュース分類なども分類手法の利用の可能性が示唆される。

権利侵害なし

例2)「店員じゃないならE? お前こそ心療内科行けよ どうせここで売れない女だろ? お前こそ嘘つき泥棒」という名誉感情侵害が争われた事例。「心療内科いけよ」「どうせここで売れない女」「嘘つき泥棒」という記載は、表現として適切とはいえないものの、社会通念に照らして許容できないほど著しく不適切なものとまではいえないと判断されている。一方、不快さは十分に高いと考えられ、Toxicity 値は 0.869 と高い値を示している。

事件数	132		
本件投稿数	442		
	侵害	非侵害	合計
名誉権	235	60	295
名誉感情	112	17	130
プライバシー権	14	2	16
その他人格権	7	0	7
None	0	17	17
合計	368	96	464

表3: 日本語人権侵害表現データセットにおいて取り扱った事件と本件投稿数、ラベルの分布。一部本件投稿には複数のラベルを含む場合があるので、テキスト数とラベル数の合計は一致しない。

例3)「タトゥーのことなんぞ知らねえよ! だかと呼んだことはないとかって発想にはならないけどな笑 なにを思うがお前の勝手にすりゃいいが、ワキガで臭いと感じたことある人間は俺だけじゃないことぐらいレス見返してみればわかるやろ? なにおそんなに書き誇りていいのか知らねえがワキガはワキガなんやから匂い嗅いで確かめて来いよ!」という発言で名誉権侵害が争われた事例。他の投稿も合わせて、原告から腋臭症のような臭いがするとの限度では、事実を摘示しているといえるが、違法性阻却事由があることがうかがわれること(文脈上、上記の投稿者は、実際に原告から性的サービスの提供を受けた上で、実体験として、腋臭症のような臭いを感じたことを主張していることが明らかである。)と、証拠だけでなく、掲示板の投稿内容から真実性が判断されたため、名誉権侵害は否定されている。こちらも、ワキガの指摘という不快感の高い発言に対して、Toxicity 値は 0.518 と高くなっている。