日本の SNS における有害な投稿と健全な投稿の比較分析

松帆愛 [1]　彌冨仁 [1]
[1] 法政大学理工学研究科応用情報工学専攻
ai.matsuho.3r@stu.hosei.ac.jp　iyatomi@hosei.ac.jp

概要

本研究は，日本語 SNS 上の有害な投稿と健全な投稿を比較して分析することで，オンライン環境の理解と健全なコミュニケーション環境の促進を目指している．X（旧 Twitter）上の 100 人のユーザーから収集されたデータを用いて語彙，それらの単語の組み合わせならびにトピックの解析を行ったところ，有害な投稿はネガティブな感情や攻撃性の高いトーンが顕著で，健全な投稿はポジティブなトピックや穏やかな言葉を含む傾向があった．本解析により，有害な投稿と健全な投稿の単語やトピックの違いが明らかにされ，特に有害なトピックの一つとして「育児」が挙げられる可能性が，日本独自の文化的背景に関連して示唆された．

1　はじめに

SNS 上の交流は社会生活に欠かせないものとなっているが，一方で有害な投稿が引き起こすネットトラブルが深刻な問題となっている [1][2]．特に個人への有害な投稿は精神的健康に深刻な影響を与えるため，オンライン空間を安全に保ち，健全な環境を作る方法を探ることが急務である．

有害な投稿の特性を理解するため，その背後にある「有害な投稿とは何か」という問題に焦点を当てることで，教育や啓発など有害な投稿を減少させるための包括的なアプローチを取ることが可能となる．

しかし，国や文化によって有害な投稿の原因やトピックが異なるにもかかわらず，これらの研究 [3][4] の多くは英語圏に集中している．日本では有害な投稿の検出を行う研究は，Twitter 上のネットいじめの自動検出 [5] や BERT を利用した煽りツイート検出 [6]，また SNS 上の攻撃的表現の検出と位置特定 [7] や誹謗中傷におけるトゲワード（罵詈雑言やネガティブな言葉）の検出 [8] などが挙げられる．一方，日本独自の有害な投稿の詳しい分析は十分に行われていない．

本研究では日本語における SNS 上の有害，および健全とみなせる投稿について，頻繁に出現する名詞やそれらの組み合わせ，またトピックの基礎的な分析を行う．これらの分析によって，有害な投稿について洞察を得ると同時に，日本の健全なオンライン空間の形成に貢献することを目指す．

2　関連研究

2.1　英語における有害な投稿の分析

Alshamrani ら [9] は，14,506 本の YouTube ニュース動画に投稿された 700 万件の YouTube コメントを分析し，有害なコメントの検出とトピックとの関連性の調査を行った．収集された動画の約 69%に有害性の高いコメントが含まれていること，またトピックとの関連性では宗教と暴力・犯罪関連のニュースが有害なコメントの割合が最も高く，経済関連のニュースが最も低いことが判明した．

Chong ら [10] は，シンガポールにおける SNS 上の有害コメントの特徴，また有害コメントのトリガーについて分析を行い，シンガポールにおける有害コメントのトリガーとなる 8 つの主要なカテゴリーを特定し，またそのトリガーは西洋と東洋では著しく異なる可能性があることを実証した．

2.2　日本語における有害な投稿の分析

植田ら [11] は，誹謗中傷行為に発生する投稿者の心理感情傾向について YouTube 上で炎上したコメント欄を題材に研究を行い，ネガティブ感情傾向よりも圧倒的にポジティブ感情傾向から誹謗中傷行為が発生する結果を報告した．

瀬川ら [12] は，日本語の X のサンプリングデータから，攻撃ポスト (旧ツイート) を抽出し，そのポストをしたユーザーの分析を行った．一見，世の中の様々な人間から攻撃されているように感じられるポストでも，実際それらのユーザーは，ほとんどが普

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 Toxic と Normal と判断された投稿の一例

投稿内容	有害度 †
死ねカス	0.939
食い気味でキレてきたババア気悪いな	0.812
ブスは臭い息吐くな	0.765
頑張って早く治しますね (´ ω `)	0.001
なるほどー！明日、スーパー見てこよ	0.002
俺も佐賀ラ行きたかったあぁぁぁぁ	0.002

† PerspectiveAPI による推定

段からインタラクションを行なっているユーザーであり，そうでない場合は普段からネガティブな投稿を多く行う特殊なユーザーからの攻撃であるとの結果が得られた．

これらは個々の炎上事例に基づくものや，投稿したユーザーの分析であり，有害な投稿内容の文面を解析するための分析は十分に行われていない．本研究では日本語の有害な投稿内容の分析を健全な投稿内容との比較から行う．

3 データ

日本の有害な投稿の詳しい分析を行うため，国内ユーザー数が 4500 万人を超える SNS である X で，ランダムに選択したユーザー 100 人の投稿を収集する．この収集では，公式アカウントの投稿や画像・リンクのみの投稿は除外する．対象期間は 2022 年 10 月 8 日から 2023 年 10 月 9 日までで，計 21,972 件の投稿を収集する．

これらに対し Perspective API[1] を活用して有害度を推定し，数値が高い 25%の投稿（5493 件）を Toxic(1) とし，一方で数値が低い 25%の投稿（5493 件）を Normal(0) とする．これらの例を表 1（上記の 3 つが Toxic，下記の 3 つが Normal）に示した．

4 分析内容

4.1 単語分析

Toxic と Normal のカテゴリ間で TF-IDF の差が 0.5 以上の名詞（固有名詞などの限定的な単語は除外）を抽出し，比較を行う．この比較により，Toxic と Normal で用いられる名詞の違いを明らかにし，それぞれの投稿の特性や傾向を分析する．

1） https://perspectiveapi.com/

表 2 Toxic と Normal の特徴的な単語（TF-IDF）

Toxic の特徴的な単語			Normal の特徴的な単語		
単語	Toxic	Normal	単語	Toxic	Normal
嫌がらせ	2.983	0.000	素敵	0.145	1.903
死	2.150	0.000	今週	0.200	1.886
アホ	2.093	0.000	ごはん	0.273	1.885
ゴミ	2.089	0.000	楽しみ	0.232	1.826
バカ	2.054	0.000	笑顔	0.381	1.667
カス	2.000	0.103	今日	0.612	1.854
バイト	2.000	0.158	休み	0.630	1.479
嫌	1.913	0.174	昨日	0.716	1.386
問題	1.527	0.545	仕事	0.802	1.372
無理	1.366	0.832			

4.2 単語の組み合わせ分析

両カテゴリにおいて，各投稿内に同時に出現する名詞の共起回数を計測する．各投稿内で同じ名詞が異なる位置に出現する場合も，同じ名詞同士の組み合わせとしてカウントする．例えば，投稿内容が「テストが嫌だな。嫌だ！」の場合，「テスト」と「嫌」のペアと「嫌」と「嫌」のペアが計測される．この分析により，単語間の関連性やパターンに関する洞察を得られる．

4.3 トピック分析

両カテゴリで LDA（潜在的ディリクレ配分法）を用いた，トピック（主題）分析を行う．トピック数は 4 に定め，それらを構成している重要度の高い単語の上位 10 個ずつを抽出する．重要度は，その単語が特定のトピックにおいて出現する確率として表される．この手法を通じて，トピック同士の関連性や文章の内包する意味を解析することができる．

5 分析結果と考察

5.1 単語比較の結果と考察

Toxic と Normal の投稿に含まれる名詞の TF-IDF の値から，カテゴリ間で差が大きい特徴的な単語（名詞）を抽出した結果を表 2 に示した．

Toxic はネガティブな感情や攻撃的なトーンを持ち，他の人や特定のトピックに対する不満や批判を表現する傾向があった．一方，Normal はポジティブなトピックや日常生活に関連した言葉を含み，幸福感や楽しさを共有しようとする傾向があった．これ

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 3　Toxic と Normal の特徴的な単語ペア

Toxic の単語ペア			Normal の単語ペア		
単語 1	単語 2	回数	単語 1	単語 2	回数
今日	今日	378	お願い	今日	206
鬱鬱	鬱鬱	190	今日	朝	150
弁当	弁当	140	おすすめ	話題	92
作り	弁当	108	朝	読書	90
無理	無理	91	お願い	ありがとう	87
授乳	駅	58	朝	活	78
弁当	生活	56	今日	明日	76
子ども	授乳	52	今日	仕事	72
子育て	授乳	52	今日	読書	72
嫌がらせ	授乳	51	今日	雨	72

らの特徴は，投稿者の感情や態度を反映しており，コミュニケーションのトーンや内容に影響を与えている．

5.2　単語の組み合わせ分析の結果と考察

単語の共起分析によって，Toxic と Normal の特徴的な単語の組み合わせが捉えられた．出現回数の多かった 10 組を表 3 に示した．また，両カテゴリの投稿における単語同士のつながりを可視化したものの一部を図 1，図 2 に示した．

Toxic における特徴的な単語の組み合わせにはいくつかの傾向が見られた．まず，同じ単語が繰り返されることが多かった．例えば，「今日」や「鬱鬱」などの単語が連続して出現した．さらに，「弁当」や「授乳」などの育児関連の単語の組み合わせが見られた．図 1 からも「嫌がらせ」という単語に「子ども」や「子育て」などの多くの育児関連の単語がつながっていることがわかった．また，育児関連の単語同士にもつながりが見られたため，育児に関連した投稿に「嫌がらせ」という単語が結びついていることがわかった．このことから，育児関連の単語が他の単語と関連付けられている場合，攻撃的なコメントが育児に関連した有害な発言である可能性が示唆される．また，ネガティブなトーンを示す単語も頻繁に出現した．「嫌がらせ」や「鬱鬱」などの感情的なネガティブな単語は，攻撃的な感情や不快感を表現するために使われることが多かった．

Normal における特有の単語の組み合わせにはいくつかの傾向が見られた．まず，礼儀正しい言葉が頻繁に使用されていた．「お願い」や「ありがとう」といった言葉がよく見られ，これは非攻撃的なコミュニケーションの特徴であり，他の人との協力やポジティブな対話を促す役割を果たしている．また，一般的なトピックに関連する単語の組み合わせ

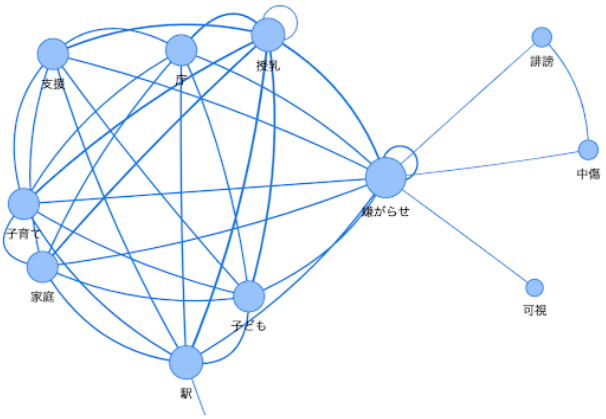

図 1　Toxic の単語同士のつながり（一部）

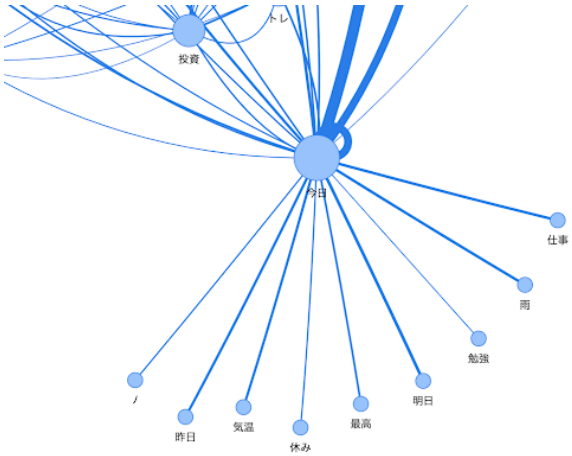

図 2　Normal の単語同士のつながり（一部）

も多く見られた．例えば，「今日」や「朝」といった日常的な話題に関連する単語が共に出現し，攻撃的な感情を表現するためのネガティブな単語は見当たらなかった．図 2 からも「今日」という単語に，「昨日」や「気温」などの多くの日常的な話題に関連する単語がつながっていることがわかった．さらに，特定のトピックに関連する単語の組み合わせも Normal の特徴であった．「今日」と「明日」，「おすすめ」と「話題」など，トピックに密接に関連した単語が結びついていた．これらの特徴は，健全なコミュニケーションにおける情報共有を示唆する．

よって，単語の組み合わせ分析から Toxic は感情的でネガティブなトーンを持ち，特定のトピックに焦点を当てて攻撃的な言葉が使われる傾向があることがわかった．一方，Normal は礼儀正しい言葉を使用し，一般的なトピックや情報共有に焦点を当てたコミュニケーションの特徴があることが判明した．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

5.3 トピック分析の結果と考察

Toxic と Normal のそれぞれのトピックを構成している重要度の高い単語上位 10 個を表 4 と表 5 に示した．

Toxic の Topic0 は「一般的な日常会話」に関連し，「今日」や「朝」といった一般的な言葉が含まれていた．特に有害な要素は見当たらず，このトピックは Normal と共通する要素を持っていることが判明した．Toxic の Topic1 は「育児に関する攻撃的なコメント」と推定できる．一見ポジティブな育児関連の単語（「弁当」，「授乳」，「家庭」，「子ども」など）が多いが，その中に攻撃的な要素（「嫌がらせ」，「無理」など）も含まれていた．よって，ネガティブな単語がポジティブな単語と組み合わさって攻撃的な文脈を持つ可能性がある．Toxic の Topic2 は，「文脈判断に基づく攻撃的なトピック」であった．「お前」や「ダメ」などの攻撃的な言葉が含まれており，他の単語も攻撃的な要素を持っていた．「人間」や「日本」といった非攻撃的な単語も含まれているが，文脈から判断して攻撃的な性質を持っていると言える．Toxic の Topic3 は，「有害な要素を含むコンテキスト」であった．「彼氏」や「息子」などの特定の相手を示す単語や，「地獄」や「嫌」などのネガティブな単語から，特定の相手に対する攻撃的な投稿であると考えられる．

Normal の Topic0 は「ポジティブな日常の楽しみ」であった．ポジティブな単語（「明日」，「楽しみ」，「美味」など）が含まれており，日常の楽しみや食事についての投稿が見られた．Normal の Topic1 は「ポジティブな感情とブログ更新」であった．ポジティブな単語（「素敵」，「大好き」，「幸せ」など）が含まれており，非攻撃的な感情を表していた．ブログ更新や日記に関する投稿が含まれている可能性が高い．Normal の Topic2 は，「穏やかな季節と日常の出来事」であった．穏やかな単語（「お過ごし」，「お願い」，「花」など）が含まれており，季節や日常の出来事に関する投稿が見られた．Normal の Topic3 は，「健康的な行動と生活の質向上」であった．健康的な行動に関連する単語（「朝」，「読書」，「活」など）が含まれており，生活の質の向上や健康に関する投稿が見られた．

トピック分析から，Toxic はネガティブな言葉や攻撃的なトーンを含み，Normal はポジティブな単語や穏やかなトーンを持っていることが特徴であることが判明した．Toxic は他人を攻撃したり傷つけたりする意図を持つ可能性が高い一方，Normal は情報共有やポジティブなコミュニケーションを目的としていることが把握できた．

表 4 Toxic のトピック

Topic0		Topic1		Topic2		Topic3	
今日	0.007	弁当	0.010	人間	0.005	今日	0.016
朝	0.005	好き	0.009	w	0.005	彼氏	0.002
友達	0.004	笑	0.006	僕	0.005	息子	0.002
明日	0.004	嫌がらせ	0.006	今日	0.004	相手	0.002
最高	0.004	授乳	0.005	仕事	0.004	意味	0.002
一緒	0.003	駅	0.004	お前	0.003	大丈夫	0.002
笑	0.003	家庭	0.004	親	0.003	昨日	0.002
お願い	0.003	普通	0.004	www	0.003	地獄	0.002
最近	0.003	無理	0.003	ダメ	0.003	僕	0.002
昨日	0.003	子供	0.003	日本	0.003	嫌	0.002

表 5 Normal のトピック

Topic0		Topic1		Topic2		Topic3	
明日	0.010	素敵	0.009	今日	0.035	今日	0.018
楽しみ	0.009	朝	0.008	お過ごし	0.014	朝	0.016
今日	0.008	今日	0.006	お願い	0.012	読書	0.012
お昼	0.007	幸せ	0.005	花	0.010	活	0.011
本	0.005	大好き	0.005	仕事	0.008	起床	0.009
2023	0.005	最高	0.005	夜	0.008	投資	0.008
ごはん	0.005	ブログ	0.005	素敵	0.006	脳	0.008
美味	0.004	ごはん	0.004	季節	0.005	ストレッチ	0.008
試合	0.003	更新	0.004	おすすめ	0.005	トレーニング	0.008
章	0.003	日記	0.004	健やか	0.005	改善	0.008

6 おわりに

本研究では日本語における SNS 上の有害な投稿と健全な投稿を比較し分析することで，有害な投稿についてより理解を深めることができた．3 つの分析結果から，有害な投稿はネガティブで攻撃的な言葉を含み，特定のトピックに焦点を当てた攻撃的な傾向があることがわかった．一方，健全な投稿はポジティブで礼儀正しく，一般的な情報共有に焦点を当てたコミュニケーションを示した．この違いは投稿者の感情やコミュニケーショントーンを反映しており，有害な投稿は攻撃的な意図が，健全な投稿は情報共有やポジティブなコミュニケーションが主な目的であることを示唆してる．また，有害な投稿の主要なトピックとして「育児」が挙げられるのは日本人特有の価値観であると考えられる．今後の展望としては，データ数（ユーザー数や投稿数など）を拡大しての分析，また日本語以外の有害な投稿との比較分析が挙げられる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Mondal, Mainack, Leandro Araújo Silva, and Fabrício Benevenuto. A measurement study of hate speech in social media. In **Proceedings of the 28th ACM Conference on Hypertext and Social Media**, 2017.

[2] 平野太一, 田中文英. Sns の誹謗中傷抑制に向けた投稿者の意識・行動変容を促す手法の検討. 2022 年度人工知能学会全国大会（第 36 回）, 2022.

[3] Friederike Schultz, Sonja Utz, and Anja Göritz. Is the medium the message? perceptions of and reactions to crisis communication via twitter, blogs and traditional media. **Public relations review**, Vol. 37, No. 1, pp. 20–27, 2011.

[4] Did you miss my comment or what?" understanding toxicity in open source discussions. Miller, courtney and cohen, sophie and klug, daniel and vasilescu, bogdan and kaustner, christian. In **Proceedings of the 44th International Conference on Software Engineering**, pp. 710–7222, 2022.

[5] 大友泰賀, 張建偉. 多特徴を用いた twitter 上のネットいじめの自動検出. Technical report, 情報処理学会東北支部研究報告, 2018.

[6] 松本典久, 上野史, 太田学. Bert を利用した煽りツイート検出の一手法. 第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM2021) 論文集, 2021.

[7] 牧元大悟, 徳永健伸. Sns 上の攻撃的表現の検出と位置特定. 言語処理学会第 28 回年次大会 (NLP2022) 発表論文集, pp. 1961–1965, 2022.

[8] 伊藤圭吾, 荒澤孔明, 服部峻. 誹謗中傷による被害を減らすためのツイートにおけるトゲワード検出. Technical Report 311, 信学技報, 2021.

[9] Sultan Alshamrani, Mohammed Abuhamad, Ahmed A. Abusnaina, and David A. Mohaisen. Investigating online toxicity in users interactions with the mainstream media channels on youtube. In **International Conference on Information and Knowledge Management**, 1988.

[10] Yun Yu Chong and Haewoon Kwak. Understanding toxicity triggers on reddit in the context of singapore. In **the Sixteenth International AAAI Conference on Web and Social Media**, 2022.

[11] 植田康孝, 梨琉生. 人工知能を用いた誹謗中傷行為の投稿者感情に関する解析〜「ユーチューバー 31 人宴会」炎上のケース〜. 江戸川大学紀要, No. 32, pp. 187–202, 2022.

[12] 瀬川友香, 浅谷公威, 坂田一郎. ユーザーに着目した sns 上の攻撃とそのメカニズムに関する分析. 人工知能学会全国大会論文集 JSAI2021, 2021.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).