ソーシャルメディアにおけるヘイトスピーチ検出に向けた日本語データセット構築の試案

荒井 ひろみ [1,2]　和泉 悠 [3]　朱 喜哲 [4]　仲宗根 勝仁 [1]　谷中 瞳 [1]
[1] 理化学研究所　[2] 科学技術振興機構 さきがけ
[3] 南山大学　[4] 大阪大学
{hiromi.arai,katsuhito.nakasone,hitomi.yanaka}@riken.jp,
yuizumi@nanzan-u.ac.jp, nyarl905@gmai.com

1 はじめに

今日のデジタル社会が直面している主要な問題の1つに，ソーシャルメディアなどのネット上に蔓延する攻撃的・暴力的表現の取り扱いがある．特に社会的弱者を標的としたいわゆる「ヘイトスピーチ」は世界的な問題となっており，差別や暴動すらを煽動するこうした表現への対応が求められている．ヘイトスピーチには排外主義的なものも多く存在するが，特にコロナ禍において，排外主義に対する懸念が多くの組織や識者により表明されている[1)]．

ヘイトスピーチをどのように検出すればよいのかといった課題に取り組むため，現在までに多数の暴言やヘイトスピーチに関するデータセットが多様な言語及びメディア資源を基にして構築されてきた．例えば，英語だけでなく，ドイツ語，イタリア語 [1]，ロシア語 [2]，タイ語 [3]，インドネシア語 [4] などのデータセットがある．本研究では，日本語のヘイトスピーチのデータセットの構築に取り組んだ．

大規模にヘイトスピーチのデータセットを構築するためには，クラウドソーシングの活用など非専門家によるアノテーションが想定される．しかし，「ヘイト（スピーチ）」という言葉は個々人による見解が一致しないと想定され [5]，厳密な定義は不可能だという指摘もある [6]．そのため，SNS 上の投稿のような，発言の対象が曖昧であり複雑な様相を呈すデータの分類に際しては，アノテーターの主観的な判断に委ねる要素を極力減らすことが望ましい．

本研究の貢献は以下の通りである．まず，日本語におけるヘイトスピーチのデータセット作成に向けて，発言が話題にしている対象者とその対象者に対する攻撃性の種類を判断させるような手続き的な暴言分類のアノテーションガイドラインを設計した．特に日本独自の歴史・地域性を踏まえたヘイトスピーチ理解と，人文社会科学におけるヘイトスピーチに関する理論的知見に依拠したアノテーションガイドラインを設計した．さらに SNS 上のヘイトスピーチの候補となるツイートを収集し，クラウドソーシングを用いたアノテーションを試験的に実施し，得られたデータセットの特徴を考察した．

2 関連研究

ヘイトスピーチやその他ネット上の有害表現をもとにしたデータセット作成が急速に進められている．言語の種類，データの収集方法，データのアノテーション手法・ラベルの種類どれも多様である[2)]．他の法的概念と同様，「ヘイトスピーチ」に世界共通の定義があるわけではなく，各研究者が独自にラベルを規定して，試行錯誤を通じてアノテーション手法を検討している段階だと言える．例えば [8] は社会学における批判的人種理論の枠組みを利用し，「人種差別的」(racist) もしくは「性差別的」(sexist) というラベルをツイートに付与した．「憎悪的」(hateful) や「攻撃的」(offensive) といったラベルを何らかの形で特徴付け，アノテーターに直接的に判断させる事例も多い [9, 10, 11]．既存のアノテーション手法の限界として，黒人英語をより攻撃的なものと分類している，という人種差別的バイアスの存在も指摘されている [5]．多言語を使用してデータセットを構築することにより，南アジアでの女性蔑視的傾向の高さを測定したといった [12]，地域性を考慮した研究も増加している．本研究も，日本国内における排外主義（移民や外国人，在日コリアン

1） 移住連 (https://migrants.jp/news/voice/20200318.html) など

2） 例えば [7, Table 2] を参照されたい

など「日本人」以外を排斥しようとするもの）に焦点を当てた，地域性を持つものと言える．

より精度の高いアノテーションを目指して，ラベル設定の精緻化やタスクの分散化が図られている．攻撃対象，攻撃内容やその程度などを細分化するといった手法が採用されている．例えば，ヘイトスピーチの根拠となる属性として，人種・性別・障がいといったラベルを付与させるといったものである [4, 13, 14, 15]．本研究も，以下で見るように攻撃対象と攻撃内容を分けて細目化したアノテーション方法を提案している．

日本語を対象とした類似研究として特にネットいじめ (cyberbullying) の文脈における有害情報検出の手法を検討したものが多く存在する [16, 17, 18, 19]．本研究は，個人だけでなく特定属性の集団を標的とした排外主義的なヘイトスピーチに焦点を当て，有害な書き込みを別の角度から検討している．

コロナ禍における反アジアのヘイトスピーチについて，英語ツイートのデータセットが作成されている．COVID-19 のキーワードが含まれるツイートのうち，個人またはアジアの人々，組織，国，または政府のグループに向けられたヘイトスピーチ，反ヘイト，中立の 3 種類の分類 [20] 及び敵意や糾弾，カウンタースピーチなどの分類 [21] がアノテーションされている．本研究も同様にコロナ禍を考慮しているが，特に日本における排外主義的ヘイトスピーチに着目した分類を行っている．

3 構築手法

3.1 検索語リスト作成

日本語のツイートを対象に，排外主義的ツイートの収集を目的として検索語を合計 12 件選定した．検索語には，日本で特に顕著に見られる在日コリアン及び中国人に関連する差別語（例えば「在日」や「シナ」）と，差別語に対応する中立的な語（例えば「韓国人」や「中国」），コロナ禍における排外主義的な言動を収集するために「反日」や「売国」といった語を含めた．

3.2 データ収集

日本語のツイートを Twitter Search API を用い収集した．COVID-19 関係の暴言ツイート候補については，上記の検索語と，一般的に COVID-19 関係の話題に含まれるキーワードである「コロナ」を AND 検索して収集した．また，COVID-19 以外の暴言ツイート候補については，上記の検索語から「コロナ」を除外して検索し収集した．なお，今回収集した COVID-19 関係の暴言ツイート候補は全て目視で COVID-19 関連であることを確認した．検索においては URL や画像・動画を含むもの，リツイートを除外した．検索は 2020/12/1–2020/12/9 の期間の日本時間の各日に，各検索語の組に対して上限 100 ツイートまでランダム選択で収集した．さらに各検索語から期間中に収集されたツイートに対し，それぞれ 5 から 23 件を乱択しデータセットを作成した．

3.3 ガイドライン設計

ヘイトスピーチに関する人文社会科学分野の研究蓄積からも「ヘイト（スピーチ）」概念の定義については一致を見せておらず，なおのこと非専門家アノテーターに対して同概念の適用に関する判断を求めることは困難である．そのため，本研究ではこうした概念を使用せず，その典型的な特徴である「集団に対する攻撃」という側面に基づいて，より負荷の少ない判断を求める方法をとった．具体的には，ラベルを攻撃対象 (A) と攻撃内容 (B) に分け，各々を細分化したガイドラインを設計した．ラベル A のカテゴリはツイートが話題にしている対象者の分類である．ラベル B のカテゴリは A で選択された対象者に対する攻撃の種類の分類である．また，[22] の分類を参考に，文として成立しておらず意味が取れないもの，広告のみのもの，フィクションについてのツイートなどは分類の対象外となるラベル C とした．アノテーターには各対象ツイートの本文について A，B 各カテゴリから 1 つ以上の該当ラベルを選択するか，C を選択するよう指示した．ツイート内容に複数の主張が含まれるケースを考慮し，A と B それぞれのラベルについて複数選択可能にした．

ガイドラインには，簡単な説明と短い例文を添えたラベル及び，アノテーションについての補足説明とラベル付けの具体例を記載した．また，ある表現が差別語にあたるかどうかの判断は個々の知識に依存するため，今回は「中国ウイルス」や「在日特権」などを含む 27 項目の差別語からなるリストを作成し，アノテーターが参照できるよう配慮した．

表 1 にアノテーションに用いたラベルの概要を示す．A1 は人種やセクシャリティなどに基づくグループ，A2 はそれに限らない職業や思想などで定義されるグループとし，A3 は著名人や知人等であ

表 1　ラベルの概要.

カテゴリ A	対象
A1	容易に変更できないヒトの属性にもとづくグループ
A2	A1 には当てはまらない何らかのグループ
A3	個人
A4	対象がはっきりしないもの
カテゴリ B	**攻撃タイプ**
B1	コミュニティや地域からの排除を告知・扇動する
B2	生命を脅かす，精神的・身体的な危害を加える，名誉を傷つけるなどのことを告知・扇動する
B3	口汚くののしる，侮辱する
B4	不確かな根拠にもとづいた情報を言いふらす，拡散させる
B5	B1-B4 のどれにも当てはまらず，攻撃的でない

る．B1 と B2 は，日本のいわゆる「ヘイトスピーチ解消法」の施行に伴う啓発活動において用いられたヘイトスピーチの例示[3]，及び川崎市のヘイトスピーチ禁止条例解釈指針[4]を参考に作成した．B3 に該当するのは，例えば「ゴキブリ以下」など侮蔑的なことを言い誰かをおとしめる行為である．B4 に該当するのは，例えば「○○人が暴動を起こしている」など事実関係が不明なことを言いふらす・拡散させるなどして誰かをおとしめる行為である．

補足説明では，差別語や「バカ」や「チビ」といった不快語がツイートに含まれていることだけを根拠としてそのツイートを攻撃的だと判断しないよう指示した．また，「政権太郎は議員を辞めるべきである」といった内容を含むツイートはある意味で特定のコミュニティから個人を排除するよう主張するものだが，批判・論評の範囲に収まる内容の場合には攻撃的と判断しないよう指示した．A1 かつ B1，または A1 かつ B2 のツイートが明確なヘイトスピーチの事例と考えられる．A1 かつ B3，または A1 かつ B4 のツイートは保護の対象となるグループに対する攻撃的な内容を有し，法規制の対象外ではあるものの少なくともその一部には広義のヘイトスピーチに該当するものが含まれていると考えられる．

3）　http://www.moj.go.jp/JINKEN/jinken04_00108.html
4）　https://www.city.kawasaki.jp/templates/press/cmsfiles/contents/0000115/115983/0316houdou.pdf

3.4　アノテーション

本研究ではクラウドソーシングプラットフォームであるランサーズ[5]を通じてアノテーションを実施した．アノテーターは日本語のネイティブスピーカー 3 名とした．アノテーターは最初にガイドラインの読み込み及び 10 問のアノテーションテストを全問正解するまで繰り返し実施し，ガイドラインを十分理解するようにした．その後アノテーターはツイートのアノテーションを実施した．多数決によって 2 人以上一致したアノテーションを各ツイートのラベルとした．

4　データセットの概要

キーワード「コロナ」を含むツイート 230 件，含まないツイート 270 件のデータセットが得られた．それぞれ **CALD** (COVID-19-related abusive language dataset)，**NALD** (non-COVID-related abusive language dataset) とする．B1〜B4 の少なくとも一つのラベル（攻撃的ラベル）を得たデータは，CALD が 56.5%，NALD が 46.3%であった．本研究で無害と判定される B5 のラベルを得たデータは，CALD が 41.3%，NALD が 50%であった．各データセットにおける各ラベルの総数を表 2 に示す．

クラウドワーカー 3 人のアノテーションの一致度は，全体で CALD で 0.83，NALD で 0.80 であった．ラベルがつかなかったケースは，CALD で 15 件，NALD で 17 件あった．他の方法はアノテーションの前に候補ツイートのスクリーニング [4] などを利用しているが，本研究ではラベルなしの割合が 10% 以下となるため，ワーカーによる判断のゆらぎはある程度抑えられたと考えられる．

5　分析・考察

差別語を含むが無害な事例　3.3 節で述べたように，アノテーターには差別語が含まれていることだけを理由に攻撃的なツイートと判断しないよう指示した．その結果，差別語が含まれる発言を引用しその発言を非難するといったカウンタースピーチにあたる事例や，差別語が含まれているが誰かをおとしめる目的のない事例が B5 に分類されていた．

CALD と NALD の比較　CALD と NALD との比較を通じて，以下の仮説を得ることができた．まず，同期間で明らかなヘイトスピーチに該当する A1

5）　https://www.lancers.jp/

表 2　データセットのラベル数及びアノテーション一致度.

データセット	A1	A2	A3	A4	B1	B2	B3	B4	B5	C	all
CALD ラベル数	131	102	39	25	11	8	82	101	95	1	-
CALD 一致度	0.88	0.83	0.78	0.61	0.60	0.85	0.79	0.86	0.88	0.67	0.83
NALD ラベル数	156	93	48	32	4	3	86	89	135	5	-
NALD 一致度	0.85	0.76	0.76	0.63	0.45	0.47	0.75	0.82	0.89	0.41	0.80

と B1，または A1 と B2 の組み合わせの出現率は，CALD で 5.7%，NALD で 2.2%といずれも低水準であり，有意な差はみられなかった．他方，罵倒や侮辱などの暴言を指す B3 と A1 との組み合わせでは，NALD がやや多かった（CALD で 22.6%，NALD で 24.8%）．またフェイクニュースやデマに相当する B4 と A1 との組み合わせでは，CALD が顕著に多かった（CALD で 30.4%，NALD で 23.3%）．これより，COVID-19 の流行にかこつけた暴言に関しては，明らかなヘイトスピーチや罵倒・侮辱的発言よりも，まことしやかなデマ，フェイクニュースの形をとりがちである，という仮説が提出できる．これは今後の計量的な検証が必要なものの，今回のアノテーション方式を通じて可能になった示唆である．

アノテーションが揺れた事例　ラベルがつかなかったツイートについては，専門家かどうかに関わらず判断の難しい事例が多数見られた．例えば，「○○人の友人がいますが，表面的に付き合う分にはいいが社会的に付き合うなら警戒感が欠かせません」というツイートは，「黒人の友達がいる」論法[6)]の典型的な事例だが，この内容だけでは当のツイートが暴言かどうかを判断するのは難しい．このツイートは特定の人種に対する何らかの偏見に基づいており，ある種の有害さを含んでいると考えられるが，B1 や B2 に該当しないのはもちろんのこと，明確な侮辱でもなければ，具体的なデマや偏見を拡散させるものでもない．

個人差がある事例　他国の政権や政党に言及したものや在日米軍などの他国の組織への言及があるツイートについて，A2 だけをアノテーションする者と，A1 と A2 の両方にアノテーションする者に分かれる場合があった．例えば「中国さん」というフレーズを含むツイートは中国政府だけでなく中国人あるいは一般に中国に向けられたものと考えられるが，中国政府だけを標的とした攻撃的ツイートとして A2 のみアノテーションされた場合があった．

6)　被差別者の友人や知人がいることを前置きし，自身が差別主義者でないことを主張ないしは示唆すること．差別主義者の常套句として知られる．

6　おわりに

本研究のアノテーション設計は，「ヘイトスピーチ」のような高負荷な概念に頼ることなく，非専門家アノテーターによる一定水準の検出を可能にするものであった．とりわけガイドラインは，オンライン上の暴言・ヘイトスピーチについて，きめ細かな把握が可能である．これらはヘイトスピーチに関する人文社会科学の知見を生かしつつ，またそれらの分野から提出される仮説の検証と社会実装につながるデータセットの構築という学際的研究の進展に貢献するものである．

今後の課題として以下がある．まず，本研究は現段階ではアノテーション済みデータの総数が小さく，定量的な判断や仮説検証は難しい．同様に，アノテーター数が少ないために，現段階では偏りやバイアスの可能性を排除できない．今後，アノテーター数を増やしガイドラインを改良することによって，5 章で紹介したようなアノテーションの揺れや個人差の問題を解消できる可能性がある．

さらに，データセット構築の先に望まれる社会実装に向けた分野横断的な課題がある．言語処理の観点では，今回のアノテーションで C に分類されたノイズを事前に除去したり，B5 のような検索語を含むが非攻撃的なタイプのツイートを識別することなどの自動化を試みることが期待される．他方，人文社会科学の観点では，とりわけ国内法におけるヘイトスピーチ検出に向けて，現ガイドラインでは区別できない B3（罵倒表現）や B4（フェイクニュース，デマ）の悪さの「程度」について一定の判定を可能にするための理論的検討や，それに基づいたガイドライン整備などが求められる．

謝辞．本研究の一部は JST さきがけ JPMJPR1752，JSPS 科研費 18K12194，JP20K19868，T19K230050，2020 年度南山大学パッヘ研究奨励金 I-A-2 の助成を受けたものである．

参考文献

[1] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pp. 86–95, 2017.

[2] Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. Reducing unintended identity bias in Russian hate speech detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 65–69, 2020.

[3] Suppawong Tuarob and Jarernsri Mitrpanont. Automatic discovery of abusive thai language usages in social networks. In *International Conference on Asian Digital Libraries*, pp. 267–278, 2017.

[4] Muhammad Okky Ibrohim and Indra Budi. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 46–57, 2019.

[5] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 25–35, 2019.

[6] Alexander Brown. What is hate speech? part 1: The myth of hate. *Law and Philosophy*, Vol. 36, pp. 419–468, 2017.

[7] Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep learning models for multilingual hate speech detection. *arXiv:2004.06465*, 2020.

[8] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

[9] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pp. 145–153, 2016.

[10] 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*, Vol. 11, 2017.

[11] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12, 2018.

[12] Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 158–168, 2020.

[13] 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, 2019.

[14] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv:2012.10289*, 2020.

[15] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4675–4684, 2019.

[16] 松葉達明, 桝井文人, 河合敦夫, 井須尚紀. カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出. 言語処理学会第 17 回年次大会発表論文集, pp. 388–391, 2011.

[17] 新田大征, 桝井文人, Ptaszynski Michal, 木村泰知, Rzepka Rafal, 荒木健治. カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出. 人工知能学会全国大会論文集, Vol. JSAI2013, pp. 2039–2039, 2013.

[18] Michal Ptaszynski, Fumito Masui, Taisei Nitta, Suzuha Hatakeyama, Yasutomo Kimura, Rafal Rzepka, and Kenji Araki. Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization. *International Journal of Child-Computer Interaction*, Vol. 8, pp. 15 – 30, 2016.

[19] Michal E Ptaszynski and Fumito Masui. *Automatic Cyberbullying Detection: Emerging Research and Opportunities: Emerging Research and Opportunities*. IGI Global, 2018.

[20] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. Racism is a virus: Anti-Asian hate and counterhate in social media during the COVID-19 crisis. *arXiv:2005.12423*, 2020.

[21] Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 2020.

[22] 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. Association for Computational Linguistics, 2018.