不適切投稿自動検出システムの構築と放送禁止用語による検証

近藤昌也 [1]　狩野芳伸 [2]
[1]Septeni Japan 株式会社　[2] 静岡大学
masaya.kondo@septeni.co.jp　kano@inf.shizuoka.ac.jp

概要

偏見や誹謗中傷といった不適切な表現を検知するモデルの構築には従来アノテーションを必要としてきたが，本研究はアノテーション無しに半自動的に訓練データを構成するシステムを提案する．不適切な単語リストを予め用意し，不適切な単語を含むSNS 投稿を頻繁に行うユーザとそうでないユーザをルールベースで分類し，機械的にアノテーションに相当する訓練データを用意する．この訓練データで構築されたモデルは，特定の不適切表現に基づくアノテーションデータで学習されたモデルに比べて，背後にある潜在的なユーザ属性を捉え様々な不適切表現をより有効に検知できる可能性を示す．

1　はじめに

近年，SNS 上の誹謗中傷書き込みが社会問題となっている．その背景のひとつには，コロナ禍により社会全体が不安に包まれている中，SNS の利用時間が増えネガティブな情報に触れる機会が多くなったことが挙げられる [1]．このような状況の下，SNS 上の偏見や誹謗中傷といった不適切な書き込みを検知するシステムの開発は社会的に重要であり，様々な研究が行われてきた [2, 3, 4, 5]．いずれの先行研究においても SNS 等のインターネット上の文章に対し不適切かどうかのラベルを人手で付与し，モデルの学習を行っており，人的コストがかかる．日本語の公開学習済みモデルで関連するものとしては，ネットいじめのデータで学習したモデルがある [4][1)] が，いずれの先行研究でも公開アノテーションデータをみつけることができず，報告されたモデルの再現が難しい．また，学習データが特定ドメインや小規模であり未知語に弱く，対象の特徴である新出単語の多さや不適切な表現の幅広さに対応し難い．

我々は，不適切な SNS 投稿の検知モデル構築のために，不適切かどうかの人手アノテーション作業なしに，モデル構築のためのデータを半自動的に収集するシステムを提案する．本研究のデータ収集システムは最初に不適切な単語のリストを定義する必要があるが，それ以外は機械的に実行できる．

本研究では，不適切な表現をより高頻度に使用するユーザは，最初に種として与えた単語リストにない不適切な表現をも含む投稿が多い，と仮定する．そのうえで，提案システムではこの不適切な単語のリストを用いて，不適切な投稿をしがちな SNS ユーザ群（不適切群）とそうでない一般の SNS ユーザ群（一般群）をルールベースで分類し，それぞれのユーザ群における投稿を大規模収集することで，不適切群かどうかの二値分類ラベルが付与された投稿データセットを構築する．単に半自動的構築ができるのみならず，潜在的なユーザ属性とさまざまな表現の広がりを捕捉できると期待する．この二値ラベル付き投稿データを用いて，SNS などの投稿が適切かどうかを分類し，荒井ら [6] が作成した日本語ヘイトスピーチコーパスと，自前で作成した偏見アノテーション付き投稿コーパスとでその性能をそれぞれ検証し，提案システムが有効に動作することを示す．

2　関連研究

松葉ら [2] は，学校非公式サイトに投稿された書き込みに対し，検知すべき有害情報の定義を行った．有害・無害情報を含む実際の書き込み 500 件に対し，6 人のアノテーターで有害（削除），有害（審議），無害の 3 つのラベルを付与し，アノテーションラベルの一致度を確認し有害情報の定義に問題がないことを確認した上で 2,998 件の書き込みに対しても同様のラベル付けを行い，教師データを作成した．このデータを用い，書き込み文章の品詞の素性や TF-IDF などのスコアを特徴量として SVM[7] で有害・無害を分類するモデルを構築した．

石坂ら [3] は，誹謗中傷のような「悪口文」は「悪口単語」の有無で十分に判断可能という仮説のも

1） https://huggingface.co/ptaszynski/yacis-electra-small-japanese-cyberbullying

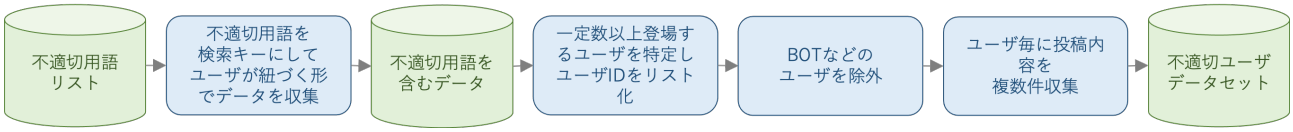

図 1　不適切群ユーザのデータ収集の流れ

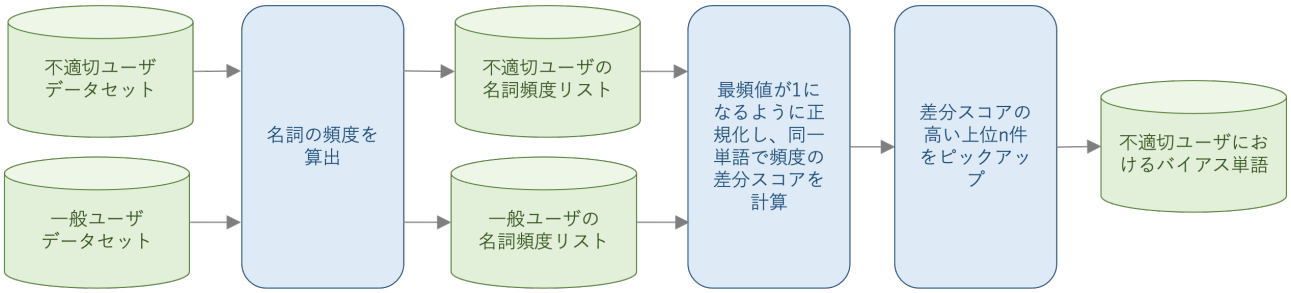

図 2　不適切群ユーザのバイアス単語抽出方法

と，単語の「悪口度」を用いて素性選択した SVM で，2 ちゃんねるの書き込みから誹謗中傷かどうかを分類するモデルを構築した．モデルを構築する際には 2 ちゃんねるから悪口文・非悪口文をそれぞれ 1,400 文ずつ収集している．

柴田ら [4] は，口語的な文章の処理に焦点を当てた ELECTRA[8] による事前学習済み言語モデルを構築し，応用例として口語の理解が必要なネットいじめ検出を扱っている．松葉ら [2] の学校非公式サイトの書き込みデータに加え，Twitter から収集した有害ツイートデータセットを組み合わせてネットいじめ検出データセットを構築している．

松本ら [5] では，BERT[9] を使って Twitter に投稿されたリプライツイートがリプライ元のツイートを煽っているかどうかを分類するモデルを構築している．リプライツイートとリプライ元ツイートの組を収集し，ツイートの組に対し，人手で煽り・非煽りのラベルを付与し，煽りツイートの組 2,134 件，非煽りツイートの組 7,786 件を構築している．

いずれの研究においても，人手でラベル付けしたデータを使ってモデルを構築しているが，データが公開されておらず再現が難しい．柴田ら [4] は HuggingFace 上で学習済みモデルを公開している．

3　提案手法

本研究では，SNS 投稿が不適切かどうかを分類するための訓練データを，人手のアノテーション作業を行わずに構成するシステムを提案する．提案システムはたとえば Twitter のように，投稿を取得する API が公開されており，かつそのプラットフォーム上のユーザとユーザの投稿内容が紐づく SNS からのデータ収集を想定している．

最初に「種」として予め不適切な単語のリストを用意し，不適切な単語を含む投稿をより高頻度に行うユーザを不適切群とみなして抽出する（図 1）．全投稿に共通する形態素列が多いユーザや全投稿に URL を含むユーザは，BOT や商品宣伝用のアカウントの可能性が高いため除外した．こうして得られた不適切群の各ユーザについて，投稿を収集する．さらに，不適切群以外から無作為抽出したユーザを一般ユーザとみなし，ユーザを 2 群に自動分類する．こうして得られた 2 群の投稿セットを用いて分類を学習し，不適切投稿推測モデルを構築する．

4　実験

4.1　放送禁止用語を利用した不適切 Twitter ユーザ群とその投稿の収集

Twitter 投稿を対象に，種となる不適切な単語リストとして放送禁止用語[2)]を採用し，不適切群ユーザのツイートデータを 2022 年 3 月 24 日から 2022 年 7 月 12 日の間で収集した．5 ツイート以上投稿履歴があるユーザの中から，収集期間中に放送禁止用語を含むツイートを 3 回以上行ったユーザを不適切群ユーザとして，1,243 人の不適切群ユーザ，計 1,209,138 件の投稿を集めることができた．mention や hashtag の文字列はツイートテキスト中から削除した．文字数が少ないツイートや，放送禁止用語自体を含むツイートも削除した．

実際に収集された不適切群ユーザのツイートを分析したところ，社会情勢（ロシアとウクライナの戦争や選挙，新型コロナウイルス関連）に関する時

2)　放送禁止用語一覧 http://monoroch.net/kinshi/で紹介されている見出し語

事的なツイートが相当数含まれていた．時事的なトピックのツイートかどうかで容易に分類ができてしまうと，必ずしも本研究の目的にそぐわないため，これらのトピックの代表的な特徴語を抽出し（図2），抽出された単語を含むツイートを不適切群ユーザのツイートから除外する処理を準備した．実際に抽出された単語は「日本，ロシア，ウクライナ，国民，中国，日本人，ワクチン，戦争，自民党，コロナ」であった．

4.2 不適切投稿分類の学習

分類モデルは事前学習済みの日本語 BERT モデル[3]の最終層に全結合線形層を 1 層追加しファインチューニングを行って構築した．一般群ユーザから不適切群と同数のユーザをランダムピックアップした上で，一般群ユーザ，不適切群ユーザそれぞれを 7:1.5:1.5 で三つにランダム分割し，学習データ，検証データ，テストデータに分けた．

前節で説明した手法により，時事単語を含むツイートを不適切群ユーザのツイートから除外したデータを使って学習したモデル（***_notopical**）も用意し比較した．**_notopical** がモデル名末尾にないものはこの時事単語の除外を適用していない．

学習時に与えるデータの単位として，1 ツイート（**1tweet_***），5 ツイート（**5tweet_***），モデルの入力最大長である 512 トークン（**maxlength_***）の 3 種類を比較した．すなわち，学習時のエポック毎に対象ユーザを割り振り，そのユーザのツイートをランダムサンプリングして，1 件，5 件ないし最大長に至るまで結合して入力とした．

学習は HuggingFace の Trainer クラス[4]を使って行い，検証データの ROC-AUC が最大になる時点のモデルを保存した．学習に使用した主な設定情報は表5 の通りである．

5 評価

前章で説明した提案モデル 6 種類（1 ツイート単位, 5 ツイート単位, 最大長の 3 種類と，時事トピックを除外するしないの 2 種類の組合せ）に，ベースラインとしてネットいじめデータで訓練した柴田ら [4] の公開学習済みモデル（yacis-electra-small-japanese-cyberbullying, 以下 **cyberbullying** と呼ぶ）を加えた計 7 種類のモデルを対象に評価を行った．

3） https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
4） https://huggingface.co/docs/transformers/main_classes/trainer

表 1 検証・テストデータにおける精度 (ROC-AUC)

	検証データ	テストデータ
1tweet	0.927	0.896
5tweet	0.984	0.986
maxlength	0.992	0.993
1tweet_notopical	0.926	0.785
5tweet_notopical	0.973	0.888
maxlength_notopical	0.984	**0.970**

評価データとして，前述の半自動的に構築した不適切・一般の二値分類データセットに加え，荒井ら [6] の日本語ヘイトスピーチコーパスと，我々が独自に作成した 5 ちゃんねるの偏見アノテーション付き投稿コーパスとの 3 種類を用いた．

5.1 不適切群ユーザ群投稿分類

不適切かどうかの二値分類データにおける ROC-AUC スコアを表 1 に示す．スコアは入力長が長いほど向上し，時事単語を除くと全体に数ポイント下落したが，**maxlength_notopical** のテストデータ評価値は 0.97 であり，提案手法により収集した一般群と不適切群ユーザの投稿は BERT モデルにより高い精度で分類可能であることがわかった．

5.2 日本語ヘイトスピーチコーパス

日本語ヘイトスピーチコーパスでは，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 件の合計 500 件のツイートに対し，ヘイトスピーチの分類を攻撃対象 (A カテゴリ) と攻撃タイプ (B カテゴリ) の 2 軸で分け，各々を細分化したガイドラインを設計している．また，3 名のアノテーターが各カテゴリにおけるラベルに投票しており，各カテゴリについて全員が一致した割合を計算したところ，平均 70.2%（43.2%〜97.0%）であった．

実験の評価値として，A1 から B5 までの各カテゴリについて，各ツイートに対する 3 名のアノテーターの投票数と，各モデルが不適切群ユーザと予測した確率の相関係数を算出した（表 2）．

5.3 5 ちゃんねる偏見コーパス

5 ちゃんねるの「私の中の偏見...」「【愚痴・悪口】ネガティブ専用チラシの裏...【誹謗中傷】」スレッド[5] における書き込みから 1,000 件を抽出し，3 名

5） http://medaka.5ch.net/test/read.cgi/kankon/1637901301, https://medaka.5ch.net/test/read.cgi/wmotenai/1655780116, https://medaka.5ch.net/test/read.cgi/wmotenai/1626458070

表 2　日本語ヘイトスピーチコーパスの各カテゴリにおける相関係数

	A1	A2	A3	A4	B1	B2	B3	B4	B5
1tweet	0.127	0.175	-0.014	-0.165	0.068	0.059	0.119	0.213	-0.151
5tweet	0.112	0.198	-0.014	-0.203	0.041	0.064	0.240	0.313	-0.317
maxlength	0.056	0.077	-0.061	-0.046	0.073	0.140	0.217	0.223	-0.267
1tweet_notopical	0.139	0.350	-0.100	-0.265	0.132	0.127	0.283	0.456	**-0.413**
5tweet_notopical	0.098	0.308	-0.078	-0.228	0.115	0.142	0.270	0.458	**-0.414**
maxlength_notopical	-0.016	0.027	0.002	0.068	0.018	-0.008	-0.006	-0.043	-0.011
cyberbullying	-0.086	-0.088	0.160	0.041	0.050	0.067	0.109	-0.081	-0.023

表 3　5 ちゃんねる偏見コーパスにおける評価スコア

	相関係数	ROC-AUC
1tweet	0.205	0.682
5tweet	0.128	0.636
maxlength	0.161	0.627
1tweet_notopical	0.292	**0.707**
5tweet_notopical	**0.375**	**0.702**
maxlength_notopical	**0.295**	0.620
cyberbullying	0.088	0.596

のアノテーターにより，5 ちゃんねるの書き込みそれぞれに対し偏見に当てはまるかを 0（当てはまらない），1（少し当てはまる），2（概ね当てはまる），3（強く当てはまる）の 4 段階でラベル付与を行った．Accuracy によるアノテーター間のラベル一致率は，全員一致の場合で 29.8%，2 名一致の場合で 87.2%であった．ラベルが 0 かそうでないかの二値分類とみなすと，全員一致が 52.5%であった．

5 ちゃんねる偏見コーパスについても同様に，3 名のアノテーターの評価ラベルの平均値と不適切群ユーザに属する予測確率値の相関係数を算出した．また，1 名でも 1 以上のラベルをつけていた場合は偏見ラベル，全員が 0 とラベル付していた場合は正常ラベルとする二値分類タスクの ROC-AUC スコアの算出も行った（表 3）．

6　考察

表 4　不適切群ツイートのイメージ

この党の議員は全員落選して 潰れたほうがいいと思う
知能の低い貧乏人ほどわかってない
戦争美化する嘘つき野郎に騙されて結婚とか
円安煽って軍事支援とか信者だけだろ道化

日本語ヘイトスピーチコーパスの B5 カテゴリは「誹謗中傷に該当しない」である．B5 に着目した二値分類を実行すると，ROC-AUC スコアは 0.7 程度であった．不適切群ユーザで投稿と予測した確率とは，**5tweet_notopical** と **1tweet_notopical** が負の相関を示し，いじめ問題で学習された **cyberbullying** よりも負の相関が強い．提案手法はヘイトスピーチを判別する性能があることが伺える．5 ちゃんねる偏見コーパスの評価スコアも同様の結果となった．

日本語ヘイトスピーチコーパスや 5 ちゃんねる偏見コーパスの評価スコアに比べて，不適切群ユーザ群投稿の分類ははるかに評価スコアが高かった．前述のように，それぞれのコーパスのアノテーション一致率は二値分類タスクで 70%程度であり，比較的難易度の高いタスク設計になっている．

また，提案手法により機械的に収集した訓練データのほうが，特定の有害情報を元に人手のアノテーションで作成された訓練データよりも広く一般的な不適切表現の特徴を多く含んでいた可能性がある．表 4 は実際に提案手法によって収集された不適切群投稿の例（原文の一部を改変）であるが，社会情勢に関する投稿の他に，偏見や皮肉，誹謗中傷とも取れる投稿が見受けられた．

7　おわりに

本研究では，アノテーション作業なしに半自動的に不適切な表現に関するデータを収集してモデルを構築するシステムを提案しその有効性を検証した．今後の課題として，種として用意する放送禁止用語を別の不適切な単語群に置き換えても同様の結果が示せるのか調査したい．また今後の応用例として，全く別の観点における単語リスト（例えばポジティブな表現の単語リストなど）を用意し，同様のデータ収集を行うことで，どのような属性のデータが収集できるのか，また収集したデータを使って機械学習モデルを構築したとき，どのようなタスクに応用できるのか検討したい．

謝辞

日本語ヘイトスピーチデータセットを提供いただいた理化学研究所の荒井先生に感謝いたします．また本研究を進めるにあたって，Septeni Japan 株式会社の勢〆弘幸氏と勝谷龍一氏から日々助言をいただき感謝いたします．

参考文献

[1] 山口真一. わが国における誹謗中傷・フェイクニュースの実態と社会的対処. Technical report, プラットフォームサービスに関する研究会, 2021.

[2] 松葉達明, 桝井文人, 河合敦夫, 井須尚紀. 学校非公式サイトにおける有害情報検出. 言語処理学会 第 16 回年次大会 発表論文集, pp. 383–386, 2010.

[3] 石坂達也, 山本和英. Web 上の誹謗中傷を表す文の自動検出. 言語処理学会 第 17 回年次大会 発表論文集, pp. 131–134, 2011.

[4] 柴田祥伍, プタシンスキミハウ, エロネンユーソ, ノヴァコフスキカロル, 桝井文人. 日本語大規模ブログコーパス yacis に基づいた electra 事前学習済み言語モデルの作成及び性能評価. 言語処理学会 第 28 回年次大会 発表論文集, pp. 285–289, 2022.

[5] 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. データ工学と情報マネジメントに関するフォーラム, pp. I14–2, 2021.

[6] 荒井ひろみ, 和泉悠, 朱喜哲, 仲宗根勝仁, 谷中瞳. 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. 言語処理学会 第 27 回年次大会 発表論文集, pp. 466–470, 2021.

[7] Vladimir N. Vapnik. **The Nature of Statistical Learning Theory**. Springer, 1995.

[8] Quoc V Le Kevin Clark, Minh-Thang Luong and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. **arXiv preprint arXiv:2003.10555**, 2020.

[9] Kenton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint-arXiv:1810.04805**, 2018.

A　付録

A.1　学習時の主な設定情報

表 5　学習時に使用した主な設定情報

per_device_train_batch_size	32
learning_rate	5e-05
lr_scheduler_type	constant
evaluation_strategy	epoch
metric_for_best_model	roc-auc
num_train_epochs	50
early_stopping_patience	10
optimizer	AdamW
loss function	CrossEntropyLoss

A.2　学習モデルの予測の振る舞い例

提案手法により構築したモデルのうち，日本語ヘイトスピーチコーパスと 5 ちゃんねる偏見コーパスの評価スコアが最も良かった **5tweet_notopical** モデルについて，不適切群ユーザ群の予測確率の振る舞いを例文とともに紹介する（表 6）．

誹謗中傷や偏見と考えられる不適切な文章は，不適切群ユーザである予測確率は高めに算出される傾向にある．また不適切なコピー文と思われる文章についても日常会話などの一般的な文章と比べると不適切群ユーザである予測確率は大きくなる傾向にあることがわかる．

表 6　モデルの予測例

#	カテゴリ	例文	予測確率
1	誹謗中傷	お前はもう永遠に無視する．二度と関わりたくない．	87.3%
2	誹謗中傷	全員地獄へ落ちろ	81.1%
3	誹謗中傷	あいつマジウザいわ，早く死んでくれ	72.5%
4	偏見	安い豚肉は腐ってるし，食べると絶対腹壊す	66.0%
5	偏見	あの地区で生まれた人は全員ケチ	57.9%
6	不適切なコピー文	これさえ飲んでいれば病気なんか罹りません	52.3%
7	不適切なコピー文	このアプリを使えば毎月 20 万円必ず稼げます．	32.8%
8	不適切なコピー文	絶対に痩せます．今なら送料無料	27.3%
9	日常会話	締め切り明日なのに，まだレポート全然書いてなかったww	8.2%
10	愚痴	マジうぜー，WiFi の接続すぐ切れるじゃん	2.5%
11	日常会話	明日みんなでカラオケ行こうぜー	2.3%
12	日常会話	サッカーめっちゃ盛り上がったよね！本戦マジで惜しかったなー	1.2%
13	愚痴	雨かぁー，最悪．洗濯物干しっぱなしにしちゃったよ．	0.1%