多様な表現を含む攻撃的テキストの自動分類

山崎 慶朋　　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{s2210175,kshirai}@jaist.ac.jp

概要

テキストの攻撃性の強さを推定するモデルの学習用データセットを構築する手法を提案する．多くの人から非難を浴びている炎上ツイートに着目し，それに対する反応を攻撃的テキストとして収集する．さらに，収集されたテキストから攻撃的でないものを自動的に除外することでデータセットの品質を高める．提案手法は攻撃的キーワードを手がかりとしないため，多様な表現を含む攻撃的テキストを収集できる．実験の結果，提案手法によって構築されたデータセットから学習された攻撃性判定モデルがベースラインを上回ることを確認した．

1　はじめに

昨今，ソーシャルメディアは世代を問わず多くの人が利用している．しかし，悪意のある攻撃的な表現が人を不快にさせることが社会的な問題になっている．そのため，ソーシャルメディアにおける攻撃的投稿を自動検出する技術の需要が高まっている．

テキストが攻撃的か否を分類するモデルを教師あり学習するためには，攻撃的なテキストとそうでないテキストを収集し，正解ラベルを付与したデータセットが必要である．従来研究の多くは，予め攻撃的な単語のリストや表現を用意し，それを含むテキストを攻撃的テキストとして収集する．しかし，攻撃的なテキストの中には攻撃的な表現を明示的に使わず暗に他者を攻撃するものも含まれる．攻撃的な単語を手がかりに収集されたデータセットから学習されたモデルは，そのような暗黙的な攻撃的テキストを正しく分類できない可能性がある．また，多くのテキストに対して人手で攻撃的か否かのラベルを付与する作業のコストが大きいという問題もある．

本研究は，攻撃的な単語を必ずしも含まない多様な表現の攻撃的テキストを分類するために，攻撃的か否のラベルを付与したデータセットを自動的に構築する手法を提案する．非道徳的な発言により多くの他者から非難されている炎上ツイートに着目し，これに対するリプライに対して「攻撃的」のラベルを付与し，攻撃的テキストとして収集する．さらに，自動付与されたラベルを修正する手法も提案する．

2　関連研究

先行研究の多くは，キーワードを手がかりに攻撃的テキストや有害テキストを収集し，テキストの攻撃性や有害性を判定するモデルを学習している [1, 2, 3, 4, 5]．牧元と徳永は，BiLSTM(Bidirectional Long Short-Term Memory) や ALBERT[6] を用いてツイート全体が攻撃的か否かを分類し，BiLSTM-CRF(Conditional Random Fields) や BERT-CRF を用いてツイートにおける攻撃的表現の位置を特定している [7]．人手でラベル付けされたデータセットを訓練データとして用いるが，攻撃的テキストは攻撃的キーワードを含む投稿を複数回行ったユーザーの投稿から収集されている．Zampieri らは，テキストが攻撃的か，その攻撃に対象が存在するか，その対象は何か，を判定する新しいタスクを提案し，これらのタスクのためのデータセットとして Offensive Language Identification Dataset(OLID) を構築している [8]．その際，攻撃的なテキストは，政治的な話題など攻撃的な反応が生じやすい話題のキーワード検索によって収集されている．

本研究は，攻撃的単語を含むという制約なしに攻撃的なテキストを収集することで，多様な表現を含む攻撃的テキストを自動分類することを目指す．

3　提案手法

3.1　データ収集

攻撃的テキストを収集するに当たり，本研究では炎上ツイートとそれに対するリプライに着目する．

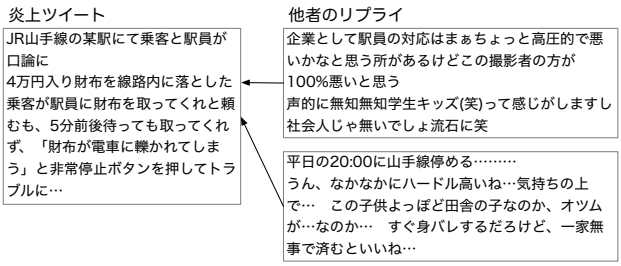

図 1 炎上ツイートとそれに対するリプライの例

1 節で述べたように，ここでの炎上ツイートとは，モラルに欠けた行動や非道徳的な振舞いを報告し，他者から非難を集めるツイートと定義する．図 1 にその例を示す．炎上ツイートに対するリプライは攻撃的なテキストが多いと考えられる．

ソーシャルメディア Twitter(現 X) においてフォロワー数が多く，様々な話題を取り上げて投稿しているユーザを選び，そのユーザの投稿から炎上ツイートを人手で選別する．次に，それに対するリプライと引用リツイート (以下，まとめて「反応ツイート」と呼ぶ) を攻撃的テキストとして収集する．同様に，動物や新製品の発売など，他者の非難を集めにくい話題のツイートを非炎上ツイートとして選別し，それに対する反応ツイートを非攻撃的テキストとして収集する．データ収集は 2023 年 1 月 16 日から 2 月 11 日にかけて実施した．収集したツイートの投稿日時の範囲は 2020 年 8 月 18 日から 2023 年 1 月 23 日であった．以下，収集したツイート集合を「炎上・非炎上ツイートデータ」と呼ぶ．同データの統計を表 1 に示す．

表 1 炎上・非炎上ツイートデータの統計

	ユーザ数	元ツイート数	反応ツイート数
炎上	3	12	20,396
非炎上	2	69	20,045

3.2 収集データの予備評価

炎上ツイートに対する反応ツイートが実際に攻撃的テキストであるかを予備的に評価した．攻撃的テキストとして収集したデータの中から 68 件をランダムに選び，それが攻撃的か否かを人手で判定した．その結果，攻撃的と判定されたツイートは全体の 29.4%であり，炎上ツイートに対する反応の多くは実際には攻撃的ではないことがわかった．一方，非炎上ツイートに対する反応ツイートを概観すると，そのほとんどが非攻撃的であった．

3.3 攻撃性判定モデルの学習

3.3.1 概要

本研究では，与えられたテキストの攻撃性のスコアを予測する．攻撃性のスコアとは，0 から 1 までの値を取り，大きいほどそのテキストが攻撃的であることを意味する．以下，テキストの攻撃性スコアを推定する回帰モデルを攻撃性判定モデルと呼ぶ．

攻撃性判定モデルとして Bidirectional Encoder Representations from Transformers(BERT)[9] を用いる．3.1 項で収集した炎上・非炎上ツイートデータにおいて，攻撃的テキスト (炎上ツイートに対する反応) のスコアは 1，非攻撃的テキスト (非炎上ツイートに対する反応) のスコアは 0 として，BERT をファインチューニングする．ただし，3.2 項で述べたようにこのデータセットには誤りが多く含まれる．そのため，テータセットの誤りの訂正とモデルの学習を交互に繰り返す．

3.3.2 初期データの作成

初期データは最初の攻撃性判定モデルの学習に用いる．初期データの作成手法として以下の 3 つを提案する．

手法 i (intact) 炎上ツイートに対する反応を攻撃的，非炎上ツイートに対する反応を非攻撃的とラベル付けする手法．攻撃的テキストには多くの誤りが含まれる．

手法 ii (PtoN) 正例 (攻撃的テキスト) のうち負例 (非攻撃的テキスト) と類似したものがあれば，そのラベルを「攻撃的」から「非攻撃的」に修正する手法．正例のテキスト p に対し，その負例に対する類似度 $NS(p)$ を式 (1) で算出する．

$$NS(p) = \mathrm{ave}_{n_i \in TOP_5(p)}\ sim(p, n_i) \tag{1}$$

ここで，$sim(p, n_i)$ は正例 p と負例 n_i の類似度，$\mathrm{TOP}_5(p)$ は p との類似度が大きい上位 5 件の負例の集合を表し，$NS(p)$ はその上位 5 件の類似度の平均値と定義する．正例と負例の類似度 $sim(p, n_i)$ は，両者を日本語用 Sentence-BERT モデル [10] を用いて埋め込み表現に変換し，そのコサイン類似度で算出する．$NS(p)$ の値が 0.7 を超えるとき，その正例のラベルを「非攻撃的」に修正する．

手法 iii (scoring) 初期のデータセットではテキストに対して攻撃的 (1) か非攻撃的 (0) の二値のスコアしか付与されていないが，ここでは $[0, 1]$ の範囲の攻撃性のスコアを推測して付与する．その手続きは以下の通りである．

1. 炎上ツイート，非炎上ツイートに対する反応ツイートの集合をそれぞれ P，N とする．
2. データセット $P \cup N$ に出現する全ての単語 bi-gram(bg_i と記す) に対し，P, N における出現回数 $\mathrm{F}_P(bg_i)$，$\mathrm{F}_N(bg_i)$ をカウントする．
3. 各 bg_i に対し，式 (2) に示す $\mathrm{R}(bg_i)$ を求める．$\mathrm{R}(bg_i)$ は，bg_i が P に偏って出現するときは大きく，N に偏って出現するときは小さくなる．
$$\mathrm{R}(bg_i) = \begin{cases} \frac{\mathrm{F}_P(bg_i)}{\mathrm{F}_N(bg_i)} & \text{if } \mathrm{F}_P(bg_i) \geq \mathrm{F}_N(bg_i) \\ -\frac{\mathrm{F}_N(bg_i)}{\mathrm{F}_P(bg_i)} & \text{if } \mathrm{F}_P(bg_i) < \mathrm{F}_N(bg_i) \end{cases} \quad (2)$$
4. 以下のいずれかの条件を満たす単語 bi-gram を除外し，残りの単語 bi-gram の集合を B とする．
   (a) $\mathrm{F}_P(bg_i) < 3$ または $\mathrm{F}_N(bg_i) < 3$
   (b) $|\mathrm{R}(bg_i)| > 100$
5. ツイート t の攻撃性スコアを式 (3) のように定義する．
$$OS(t) = \begin{cases} \mathrm{Nor}\left(\frac{\sum_{bg_i \in T \cap B} \mathrm{R}(bg_i)}{|T \cap B|}\right) & \text{if } T \cap B \neq \emptyset \\ \mathrm{median}(OS(t)) \text{ in } N & \text{if } T \cap B = \emptyset \end{cases} \quad (3)$$
   T はツイート t における単語 bi-gram の集合，Nor は攻撃性スコアを $[0, 1]$ の値にするための正規化関数である．[1] ステップ 3. の条件を満たす単語 bi-gram がツイート中にひとつも出現しないとき ($T \cap B = \emptyset$ のとき) は，N に属するツイートの攻撃性スコアの中央値とする．

直観的には，炎上ツイートの反応ツイート群 P(または N) に頻出する単語 bi-gram を多く含むツイートに高い (または低い) 攻撃性スコアを与えている．ここで，収集したデータは少数の炎上ツイート・非炎上ツイートに対する反応を集めたものであり，その話題に偏りがあることに留意する．すなわち，元のツイートの話題に関連のある話題語が P と N のどちらか一方に偏って出現する可能性がある．そのような話題語は攻撃性の強さを表すものではないため，ステップ 4.(b) の条件によって P・N のどちらかに偏って出現する単語 bi-gram を削除している．

3.3.3 モデルの学習方法

攻撃性判定モデルを学習する 3 つの手法を提案する．

手法 A (vanilla) 初期の訓練データを用いて BERT を一度だけファインチューニングする手法．

手法 B (bootstrap) ブートストラップの手法を用いてデータのラベル付けとモデルの学習を交互に繰り返す手法．まず，初期の訓練データを用いて BERT をファインチューニングし，攻撃性判定モデル M_1 を学習する．次に，M_1 を用いて訓練データの攻撃性スコアを予測し，その上位 500 件 (攻撃的テキスト) と下位 500 件 (非攻撃的テキスト) に対し，攻撃的または非攻撃的のラベルを付与する．ラベルを付与できたデータを用いて BERT をファインチューニングし，攻撃性判定モデル M_2 を得る．以下，モデル M_i によるラベルの付与と，ラベルを付与されたデータによってモデル M_{i+1} を再学習することを繰り返す．このとき，モデルの学習に用いるデータは (初期データを除いて) 漸進的に増加する．

手法 C (relabeling) 手法 B では，ブートストラップ学習の初期の段階では少量の訓練データしか使えないため，それから学習されたモデルの性能が低いことが懸念される．そこで，データセットにおける全てのサンプルを常に使いつつ，その攻撃性スコアだけを更新する手法を提案する．この手法では，モデル M_i によって攻撃性スコアを推測して付与し，スコアが更新されたデータから新しいモデル M_{i+1} を学習することを繰り返す．データセットの全てのサンプルに対するモデル M_i と M_{i+1} による予測スコアの分布を記録し，その分布のユークリッド距離が 7 以下になったら反復学習を停止する．

1) データセットに出現する全ての bg_i について攻撃的スコアを計算した後，それらの最小値を引くことで，スコアを 0 より大きい値に変換する．次に，スコアの最大値で割ることで $[0, 1]$ の範囲の値に変換する．

4 評価実験

4.1 実験データ

攻撃的か否かを人手でラベル付けしたテストデータを作成した．3.1 項で述べた炎上・非炎上ツイートデータから，炎上ツイート 2 件，非炎上ツイート 2 件を選び，それらに対する反応ツイート 280 件を抽出した．これらに対し，3 名の被験者が「攻撃的」「非攻撃的」「判定不能」「非文」のいずれかのラベルを付与した．被験者間の一致度を示す Fleiss's κ は 0.511 であった．1 名以上によって「判定不能」「非文」と判定されたツイートを除き，最終的なラベルは多数決で決定した．

訓練データは炎上・非炎上ツイートデータから前述のテストデータを除いたものである．実験データの統計を表 2 に示す．前処理として，ユーザ名，URL，改行を削除し，日本語が含まれないツイートは除外した．また，手法 i と ii については，初期

データ作成後，「草」「笑った」といった短文のツイートで，攻撃的ツイート群と非攻撃的ツイート群の両方に出現するものは削除した．

表 2　実験データ

テストデータ		訓練データ	
攻撃的	70	炎上ツイートへの反応	19,671
非攻撃的	203	非炎上ツイートへの反応	18,955

4.2　実験設定

ベースラインとして，攻撃的単語を手がかりとしたモデルを用意する．具体的には，先行研究 [1, 2, 3] で挙げられていた 40 の攻撃的単語 (「死ね」「蛆虫」など) を含むツイートを抽出し，これを正例 (攻撃的テキスト) とした．また，正例と同数の負例を非炎上ツイートの反応ツイートからランダムに選択した．正例と負例の件数はそれぞれ 671 件であった．

事前学習済み BERT モデルとして東北大学が公開している BERT base[11] と BERT large[12] を用いた．

提案手法のモデルは攻撃性スコアを予測するが，閾値を設定し，予測したスコアがその閾値以上もしくは以下であるかによってテキストを「攻撃的」「非攻撃的」「不明」のいずれかに分類する分類問題の性能を評価する．評価指標として ROC-AUC と PR-AUC を用いる．

4.3　結果と考察

BERT base を用いて学習した提案手法ならびにベースラインの評価結果を表 3 に示す．

初期データの作成手法 i, ii, iii を比較すると，ROC-AUC では ii(PtoN) が最も良いが，PR-AUC では手法 ii, iii のいずれも手法 i と比べて改善は見られなかった．モデルの学習手法 A, B, C を比較すると，手法 C(relabeling) は手法 A(vanilla) を上回った．データセットのエラーを修正する手法 C は有効であると言える．一方，手法 B(bootstrap) の指標は全体的に悪い．これは，ブートストラップの初期の段階では訓練データの数が少ないため，有効な攻撃性判定モデルが学習できなかったことが原因と考えられる．一番良い手法の組み合わせは，ROC-AUC では ii×C，PR-AUC では i×C であった．

表 3　実験結果 (BERT-base)

ROC-AUC	A(vanilla)	B(bootstrap)	C(relabeling)
i (intact)	0.792	0.567	0.804
ii (PtoN)	0.811	0.755	**0.817**
iii (scoring)	0.781	0.686	0.776
baseline		0.790	

PR-AUC	A(vanilla)	B(bootstrap)	C(relabeling)
i (intact)	0.563	0.303	**0.634**
ii (PtoN)	0.551	0.451	0.567
iii (scoring)	0.550	0.356	0.522
baseline		0.555	

表 4　実験結果 (BERT-large)

ROC-AUC	A	C
i (intact)	**0.838**	0.828
ii (PtoN)	0.803	0.810
iii (scoring)	0.775	0.780

PR-AUC	A	C
i (intact)	**0.633**	0.617
ii (PtoN)	0.525	0.526
iii (scoring)	0.530	0.484

ベースラインと比較すると，提案手法は ROC-AUC，PR-AUC のいずれも上回った．ただし，ベースラインの訓練データは提案手法よりもサイズが小さい．そこで，ベースラインの訓練データと同じ件数の炎上・非炎上ツイートに対する反応ツイートをランダムに選択し，手法 i×A でモデルを学習したところ，ROC-AUC(0.810) はベースラインを上回り，PR-AUC(0.549) はベースラインと同等であった．この結果から，炎上ツイートの反応を攻撃的テキストとして収集する提案手法は，攻撃的単語を手がかりとする手法に比べて，多様な表現を含む大規模なデータセットを構築し，より優れた攻撃性判定モデルを学習することができると言える．

BERT large を用いて学習した提案手法の評価結果を表 4 に示す．手法 B(bootstrap) は表 3 での結果が悪かったため，ここでは比較の対象としていない．BERT-base を用いたモデルと比べて，全体的に ROC-AUC や PR-AUC が改善されている．一番良い手法の組み合わせは i×A であり，BERT large を用いたときは，手法 C(relabeling) や ii(PtoN)，iii(scoring) の有効性は確認できなかった．

5　おわりに

本論文では，炎上・非炎上ツイートの反応を収集することで攻撃性のラベルが付与されたデータセットを自動構築し，その誤りを自動的に訂正した上で，テキストの攻撃性の強さを判定するモデルを学習する手法を提案した．評価実験により，提案手法が攻撃的単語を手がかりにデータセットを構築する手法よりも優れていることを確認した．今後の課題として，炎上ツイートの反応ツイートの中から非攻撃的なテキストを除外するより良い方法を探究することや，収集するツイート数を増やしてより大規模なデータセットを構築することなどが挙げられる．

参考文献

[1] 石坂達也, 山本和英. 2 ちゃんねるを対象とした悪口表現の抽出. 言語処理学会第 16 回年次大会発表論文集, pp. 178–181, 2010.

[2] 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 (第 27 回), 2013.

[3] 畠山鈴生, 桝井文人, プタシンスキ・ミハウ, 山本和英. 有害表現抽出に対する種単語の影響に関する一考察. 人工知能学会全国大会論文集 (第 30 回), 2016.

[4] 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. 情報処理学会東北支部研究報告, Vol. 2018, No. 9, 2019. B1-1.

[5] 尾崎航成, 向井宏明, 松井くにお. SNS における不適切投稿の検知. 情報処理学会第 82 回全国大会, pp. 621–622, 2020.

[6] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite bert for self-supervised learning of language representations, 2020.

[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 回年次大会発表論文集, pp. 1961–1965, 2022. (B8-4).

[8] 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, 2019.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4171–4186, 2019.

[10] 日本語用 Sentence-BERT モデル, (2024 年 1 月閲覧). https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens.

[11] BERT base Japanese – Hugging Face, (2024-1 閲覧). https://huggingface.co/cl-tohoku/bert-base-japanese-v3.

[12] BERT large Japanese – Hugging Face, (2024-1 閲覧). https://huggingface.co/cl-tohoku/bert-large-japanese-v2.

[13] Perspective API, (2024-1 閲覧). https://perspectiveapi.com/.

[14] Perspective | Developers , (2024-1 閲覧). https://developers.perspectiveapi.com/s/about-the-api-training-data.

A　Perspective API との比較

Perspective API[13] はテキストの有害性を判定する著名なシステムである．Perspective API は大量の人手でラベル付けされたデータセットから学習されており [14]，データセットの自動構築を目指す本研究とは異なるが，本論文の実験データを用いて比較する．Perspective API と提案手法 (手法 i×A, BERT-large を使用) の ROC 曲線と PR 曲線を図 2 に示す．Perspective API の ROC-AUC は 0.862，PR-AUC は 0.737 であり，提案手法を上回る．しかし，PR 曲線で再現率が高い付近では，提案手法の方が精度が高い．モデルが様々な攻撃的表現を認識できるほど再現率は高くなるが，そのような状況で提案手法が Perspective API よりも高い性能を示していることは，炎上ツイートに対する反応を収集することで多様な攻撃的表現を獲得し，攻撃性判定モデルも多様な表現を認識できていることを示唆する．

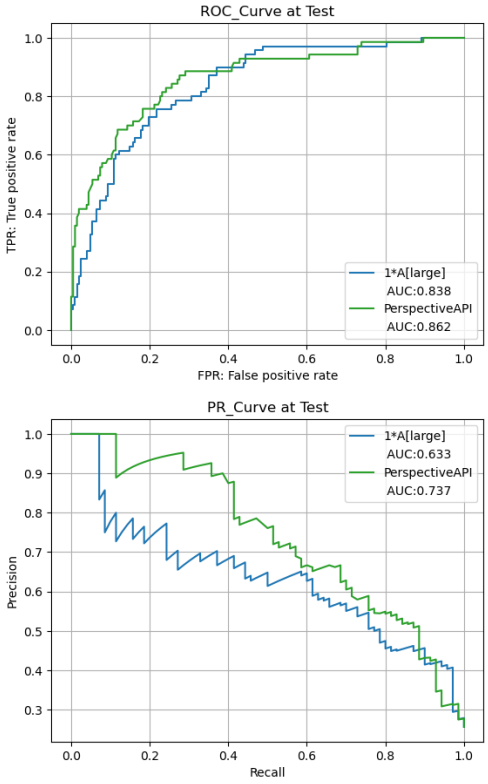

図 2　Perspective API との比較 (上:ROC 曲線/下:PR 曲線)