修士論文

多様な表現を含む攻撃的テキストの自動検出

山崎慶朋

主指導教員 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
（情報科学）

令和６年３月

Abstract

Malicious and offensive expressions in social media have become a serious social problem in recent years. To protect people from the damage caused by such offensive expressions, technology to automatically detect offensive posts in social media is in a huge demand. Supervised learning of a model to classify whether a text is offensive or not requires a labeled dataset consisting of offensive and non-offensive texts. Most previous studies constructed a dataset by collecting offensive texts using a pre-defined list of offensive words or expressions and manually annotating them with gold labels. Since it is rather difficult to create a comprehensive list of offensive words and expressions, however, texts including implicit offensive expressions or expressions containing unknown offensive words may not be included in the dataset, and a model trained from it may fail to classify such texts correctly. In addition, manual annotation of a large number of offensive texts requires much human labor.

This study proposes a method to automatically construct a dataset annotated with offensive/non-offensive labels that includes a wider variety of offensive expressions, and to train a model to predict intensity of offensiveness of a text using this dataset. Specifically, since bashing in social media, where users are accused of their immoral or unethical behavior by others, is likely to receive offensive responses, replies to a bashed post are collected as offensive texts. This method might be able to collect a wide variety of offensive expressions, since collected offensive texts are not restricted to ones that contain a limited number of offensive words. In addition, the quality of the constructed dataset is improved by automatically finding wrong offensive labels and fixing them.

First, users who have a large number of followers and often post about various topics related to bashing on Twitter (currently X) are selected. From their posts, tweets that receive a particularly large number of responses (called "bashed tweet") are manually selected, and tweets replying to those bashed tweets are collected as offensive texts. Similarly, non-offensive texts are obtained by collecting tweets replying to "non-bashed tweet", which is considered unlikely to attract criticism. Then, the validity of the constructed dataset was manually verified. It was found that only about 30 percent of the responses to bashed tweets were actually offensive texts. Therefore, we propose methods to obtain a more accurate model for prediction of the offensive intensity by iteratively correcting errors in the dataset and training the model. Our method consists of two kinds of modules: "initial data construction" and "model training".

The methods of "initial data construction" are methods to construct initial training data that is used at the beginning of the iterative training of the model. We propose the following three methods. Method i(intact) uses the aforementioned

dataset consisting of responses to bashed tweets and non-bashed tweets without any modification. Method ii(PtoN) calculates the similarity between tweets by Sentence BERT and modifies the labels of responses to bashed tweets that are similar to responses to non-bashed tweets from positive (offensive) to negative (non-offensive). Method iii(scoring) assigns an offensiveness score to each text, which is calculated by considering the bias of the frequency of the word bi-grams in the sets of responses to bashed tweets and non-bashed tweets.

As for "model training", three methods are proposed. In all these methods, BERT is used as the model for prediction of offensive intensity. Method A(vanilla) is a method to fine-tune a BERT model only once using an initial data. Method B(bootstrap) incrementally increases labeled samples in a bootstrapping manner. Method C(labeling) repeats re-estimation of offensive scores of all texts in the dataset and training the model using the updated dataset alternatively until the predicted offensive scores are converged.

In the experiments, we compare nine proposed methods obtained by combining three methods of initial data construction and three methods of model training, as well as a baseline model. The baseline model is a fine-tuned BERT model using a dataset constructed as follows. A list of 38 offensive words suggested by the previous studies was created, then the tweets including one of those words were excerpted from our dataset as offensive texts, while the same number of tweets not including any pre-defined offensive words are excerpted as non-offensive texts. As test data, we use 273 tweets annotated by three males in their 20s. We use "BERT base" and "BERT large" as pre-trained BERT models. The task in this experiment is a classification problem to determine whether a text is offensive or not, and the evaluation criteria are ROC-AUC and PR-AUC.

In the evaluation of the models using BERT base, the best method was the combination of Method ii(PtoN) and Method C(labeling) with respect to ROC-AUC and the combination of Method i(intact) and Method C(labeling) with respect to PR-AUC. Comparing methods of initial data construction i, ii, and iii, Method i and Method ii performed better than Method iii. Comparing the methods of model training A, B, and C, Method B(bootstrap) entirely poorly preformed; their ROC-AUC and PR-AUC were worse than the baseline. It indicated that Method B was ineffective. Seeing the PR curve, the proposed methods achieved higher precision than the baseline when the recall was low, lower when the recall was medium, and became higher again when the recall was high. The fact that the proposed methods outperformed the baseline when the recall was high indicated that the proposed methods were capable of detecting various offensive expressions. However, it was unclear why the proposed methods performed better even when the recall was low.

In the evaluation of the models using BERT large, the combination of Method i(intact) and Method A(vanilla) achieved the best ROC-AUC and PR-AUC. In a comparison of methods of initial data construction i, ii, and iii, on the one side, Method i performed the best in both ROC-AUC and PR-AUC, on the other hand, Method iii performed worse than the other methods. These results were consistent with the experiment using BERT base. Therefore, we can conclude that the methods of initial data construction can be ordered as i>ii>iii in terms of their effectiveness. The performance of Method A and Method C for model training were almost the same, and Method B poorly performed in the experiment using BERT base. Therefore, the methods of model training can be ordered as A$\simeq$C>B. Comparing the proposed methods to the baseline, both ROC-AUC and PR-AUC of our methods i$\times$A and method i$\times$C were higher than the baseline.

概要

昨今，ソーシャルメディアにおいて悪意のある攻撃的な表現が人を不快にさせてしまうことが社会的な問題になっている．このような攻撃的な表現による被害から利用者を保護するため，近年では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判定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となる．従来研究の多くは，予め用意した攻撃的な単語のリストや表現に基づいて攻撃的なテキストを収集し，また，これに対して人手によるアノテーションを実施することによってデータセットを構築している．しかし，様々な攻撃的な単語や表現を網羅した包括的なリストを作成することは困難であることから，暗黙的な攻撃的表現や未知の攻撃的単語を含む表現がデータセットに含まれず，それから学習したモデルはそのようなテキストの分類を誤る可能性がある．また，大量の攻撃的テキストを人手でラベル付けすることは，作業者の負担が大きいという問題もある．

本研究では，より多様な攻撃的表現を含む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し，これを基にテキストの攻撃性の強さを推定するモデルを学習する方法を提案する．具体的には，ソーシャルメディアにおける非道徳的な発言，非常識的な振る舞いによって多くの他者から非難を浴びる現象，いわゆる炎上現象が攻撃的な反応を受けやすいという性質に着目し，炎上現象の原因となった投稿に対するリプライを攻撃的テキストとして自動的に収集する．この手法では攻撃的単語を含むという条件なしに攻撃的テキストを収集するため，多様な攻撃的表現を収集することが期待できる．さらに，上記の方法で「攻撃的」のラベルが誤って付与されたテキストを自動的に検出し，それを修正することで，ラベル付きデータセットの品質を向上させる．

まず，Twitter(現 X) において，フォロワー数が多く，炎上現象に関する様々な話題を取りあげて投稿しているユーザを選び，そのユーザの投稿から，特に反応の多いツイート（炎上ツイート）を人手で選別し，その投稿に対する反応ツイートを攻撃的テキストとして収集する．同様に，非難が集まりにくいとみられるツイート (非炎上ツイート) 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．収集したテキストを検証したところ，炎上ツイートに対する反応のうち，実際に攻撃的なテキストは約 30%であった．そのため，テータセットのラベル誤りの訂正とモデルの学習を交互に繰り返すことで，より精度の高いモデルを構築する方法を提案する．本手法は「初期データの作成手法」と「モデルの学習手法」から構成される．

「初期データの作成手法」は，モデルの反復学習の最初に用いる訓練データを構築する手法である．以下の 3 つを提案する．手法 i(intact) は，前述の炎上ツイート・非炎上ツイートの反応からなるデータセットをそのまま用いる手法である．手法 ii(PtoN) は，Sentence BERT によってツイート間の類似度を計算し，炎上ツイートに対する反応のうち非炎上ツイートに対する反応に類似したもののラベルを正

例から負例に修正する手法である．手法 iii(scoring) は，単語 bi-gram の炎上・非炎上ツイート反応群における出現頻度の偏りから算出した攻撃性スコアをテキストに付与する手法である．

「モデルの学習手法」として以下の 3 つを提案する．なお，本研究では攻撃性のスコアを予測するモデルとして BERT を用いる．手法 A(vanilla) は，BERT モデルを初期データを用いて 1 度だけファインチューニングする手法である．手法 B(bootstrap) は，Bootstrap によって漸進的にデータセットを増やす手法である．手法 C(relabeling) は，データセットにおける全てのテキストに対するモデルによる攻撃性スコアの再推定と，更新されたデータセットによるモデルの学習を収束するまで繰り返す手法である．

評価実験では，3 種類の初期データ作成手法と 3 種類のモデルの学習手法を組み合わせた 9 つの提案手法，ならびにベースラインモデルを比較する．ベースラインモデルは，先行研究を参考に攻撃的単語を 38 個用意し，これを含むテキストを攻撃的テキスト，含まないものを非攻撃テキストとして収集し，このデータセットを用いて BERT をファインチューニングしたモデルである．テストデータとして，20 代男性 3 名によってアノテーションされた 273 件のツイートを用いる．事前学習済み BERT モデルとして BERT base と BERT large を用いる．評価タスクはテキストが攻撃的か否かを判定する分類問題であり，評価基準は ROC-AUC と PR-AUC とする．

BERT base を用いた実験では，最も良い手法は，ROC-AUC では手法 ii(PtoN) と手法 C(relabeling) の組み合わせ，PR-AUC では手法 i(intact) と手法 C(relabeling) の組み合わせであった．初期データの作成手法 i,ii,iii を比較すると，手法 i,ii は手法 iii と比較して良い成績が得られた．モデルの学習手法 A,B,C を比較すると，手法 B(bootstrap) を使用したモデルの成績は全体的に低く，ROC-AUC，PR-AUC ともにベースラインを下回り，あまり有効ではないことがわかった．PR 曲線を観察すると，Recall が低いときには提案手法の方がベースラインよりも Precision が高く，中程度になるとベースラインの方が高くなり，Recall が高いときには再び提案手法の方が高くなる傾向が見られた．Recall が高いときに提案手法がベースラインを上回ることから，提案手法が多様な攻撃的表現を検出する能力が高いことがわかった．ただし，Recall が低い範囲でも提案手法の成績が良い原因は不明であった．

BERT large を用いた実験では，最も ROC-AUC と PR-AUC が高かった手法は，手法 i(intact) と A(vanilla) の組み合わせであった．初期データの作成手法 i，ii，iii の比較では，両 AUC で手法 i が最も成績が良かった．手法 iii が他の手法と比較して評価値が低いことは BERT base による実験と同様であった．これらを踏まえると，初期データの作成手法は，i>ii>iii の順に有効であると言えた．モデルの学習手法については，A と C の成績はほぼ同等であった．BERT base での実験結果も踏まえると，モデルの学習手法の優劣は A≃C>B となることがわかった．提案手法とベースラインと比較すると，手法 i×A と手法 i×C については，ROC-AUC と

PR-AUC の両方で提案手法がベースラインを上回った.

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

昨今，スマートフォンの普及に伴い，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てしまうことが社会的な問題になっており，ソーシャルメディア上での攻撃により精神的な不調をきたすといった被害を受けた事例もよく発生している．攻撃的な表現による被害から利用者を保護し，より安全なサービスを提供することを目的に，近年では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判定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となる．従来研究の多くは，予め用意した攻撃的な単語のリストや表現に基づいて攻撃的なテキストを収集し，また，これに対して人手によるアノテーションを実施することによってデータセットを構築している．しかし，攻撃的なテキストの中には攻撃的な単語や表現を明示的に使わず暗に他者を攻撃するものも存在する．また，あらかじめ作成された攻撃的単語のリストに含まれていない攻撃的な単語を含むものも存在する．攻撃的な単語は時間とともに変化し，増えていくと考えられるため，攻撃的な単語を網羅的に収集したリストを作成することは困難である．したがって，攻撃的な単語を手がかりに収集されたデータセットには，暗黙的な攻撃的表現や未知の攻撃的単語を含む表現が含まれておらず，それから学習したモデルはそのようなテキストの分類を誤る可能性がある．また人手によるラベル付きデータの構築について考察すると，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はそのコストが高く，また心無いテキストを読む機会も多いために作業者の心理上の負担も大きいといった問題がある．

以上から，ソーシャルメディア上の攻撃的投稿を自動的に検出するためには，攻撃的な単語を含まないような多様な攻撃的表現を含むデータセットを構築すること，人手によるアノテーションを必要とせずにラベル付きデータセットを構築することが，重要な課題として挙げられる．

1.2 目的

本研究は，攻撃的な単語を必ずしも含まない，あるいは事前に用意した攻撃的単語以外の単語を含むような，多様な表現の攻撃的テキストを検出することを目的とする．これを実現するために，攻撃的か否のラベルを付与した多様な表現を含むデータセットを自動的に構築し，これを基にテキストの攻撃性の強さを推定するモデルを学習する方法を提案する．

まず，ソーシャルメディアにおける非道徳的な発言，非常識的な振る舞いによって多くの他者から非難を浴びる現象，いわゆる炎上現象が攻撃的な反応を受けやすいという性質に着目し，炎上現象の原因となった投稿に対するリプライを攻撃的テキストとして自動的に収集する．この手法では攻撃的単語を含むという条件なしに攻撃的テキストを収集するため，多様な攻撃的表現を収集することが期待できる．さらに，上記の方法で「攻撃的」のラベルが誤って付与されたテキストを自動的に検出し，それを修正することで，ラベル付きデータセットの品質を向上させる．自動構築したラベル付きデータセットを用いて，テキストの攻撃性の強さを推定するモデルを学習し，その有効性を実験的に検証する．

1.3 本論文の構成

本論文の構成は以下のとおりである．2 章では，本研究の主要な関連研究を紹介し，これらの研究と本研究の違いについて論じる．3 章では，炎上ツイートを手掛かりとして攻撃的ラベルが付与されたデータセットを構築する手法について述べ，その妥当性を考察する．4 章では，自動構築したラベル付きデータの誤りを訂正する手法や，これを用いて攻撃性を判定するモデルを学習する手法について述べる．5 章では，提案手法の評価実験について，実験の手続き，実験結果，それに対する考察を述べる．最後に 6 章では，本論文のまとめや今後の課題について述べる．

第2章　関連研究

本章では，本研究の関連研究について述べる．SNS 上に投稿されたテキストの攻撃性を判定する研究は数多く行われている．2.1 節では，攻撃的キーワードとの共起度によりテキストの攻撃性を判定する先行研究について述べる．2.2 節では，攻撃的キーワードを手がかりにラベル付きデータセットを構築し，これを基に教師あり機械学習によって攻撃性を判定する先行研究について述べる．2.3 節では，攻撃的キーワード以外の情報を用いて教師あり学習のためのデータセットを構築する手法について述べる．最後に，2.4 項では，先行研究と本研究の違いについて論じる．

2.2 節と 2.3 節で紹介する研究は教師あり機械学習に基づくものであり，攻撃性のラベルが付与されたデータセットを必要とする．ラベル付けの対象となるテキストはウェブや SNS から収集されることが多いが，一般に，攻撃的なテキストはそうでないテキストと比較すると圧倒的に少ない．このため，ランダムにテキストを収集すると，負例 (非攻撃的なテキスト) の数が正例 (攻撃的なテキスト) の数を大きく上回り，攻撃性を判定するモデルがうまく学習できないことが考えられる．2.2 節と 2.3 節では，主に攻撃的テキストをどのように獲得するかという観点から，機械学習に基づく先行研究を概観する．

2.1　攻撃的キーワードとの共起度を利用した研究

石坂と山本は，巨大電子掲示板サイト「２ちゃんねる」に書き込まれた文が悪口か否かを分類する手法を提案した [11]．彼らの手法では Support Vector Machine(SVM) によって分類器を学習するが，その際に素性選択を行う．具体的には，SO-PMI(Semantic Orientation Using Pointwise Mutual Information) によって単語が悪口を示唆する強さを算出し，それが高い単語のみを SVM の学習素性とした．SO-PMI は 2 つの極性の基本単語を基に，対象の単語がそのどちらと文書内共起しやすいかを表す指標である．この研究では「死ね」「消えろ」など，人手で用意した悪口単語を悪口極性を持つ基本単語とし，そのような悪口単語と共起せず，出現頻度の高い単語を非悪口極性の基本単語とした．SO-PMI の計算式を式 (2.1)-(2.3) に示す．ここで，w_p と w_n はそれぞれ悪口，非悪口極性の単語であり，$hit(w)$ は w をクエリとしたときのウェブ検索ヒット件数である．式 (2.3) は $C(w)$ と $f(a)$ の和であるが，a は関数 $f(a)$ の重みを設定する定数であり，この研究では

$a = 0.9$ としていた．

$$C(w) = log\frac{hit(w, w_p) * hit(w_n)}{hit(w, w_n) * hit(w_p)} \tag{2.1}$$

$$f(a) = a * log\frac{hit(w_p)}{hit(w_n)} \tag{2.2}$$

$$SO\text{-}PMI(w) = C(w) + f(a) \tag{2.3}$$

全ての単語 w について $SO\text{-}PMI(w)$ を計算し，それが閾値を超える単語のみを学習素性として，SVM を学習する．実験の結果，悪口基本単語を「消えろ」，非悪口基本単語を「振替」とし，SO-PMI が -0.2 を超える単語のみを学習の素性としたとき，最高の F 値 89.97 が得られた．

新田らは，自己相互情報量（Pointwise Mutual Information; PMI）を用いたスコアリングによって学校非公式サイトにおける有害書き込みを検出する手法を提案した [14]．ここでの PMI は，フレーズと有害極性単語との共起の強さを表す指標として用いた．有害極性単語はあらかじめ人手で選別した．文部科学省による有害語の定義に基づいて学校非公式サイトの書き込みを調査し，合計 255 語の有害語を選別した．これらのうち「卑猥語」「暴力誘発語」「誹謗中傷語」の 3 つのカテゴリに該当する有害語の中から出現頻度の上位 3 語の有害語をそれぞれ抽出し，合計 9 語を有害極性単語と定義した．そして，テキストの有害度スコアを式 (2.5) によって計算した．式 (2.4) に示す PMI-IR は，PMI の考えに基づき，ウェブ検索エンジンのヒット件数で 2 つの単語の共起の強さを測る指標である．p_i は有害度スコアを求める対象となる文に出現するフレーズを，w_j は有害極性単語を表す．また，$hits(p_i)$ は p_i をクエリとしたときのウェブ検索ヒット件数であり，$hits(p_i\&w_j)$ は p_i,w_j をクエリとしたときウェブ検索ヒット件数である．テキストの有害度スコアは，テキスト内のフレーズ p_i と，ある有害カテゴリにおける 3 つの有害極性単語 w_j の組み合わせについて，$PMI\text{-}IR(p_i, w_j)$ を計算し，その最大値と定義されている．

$$PMI\text{-}IR(p_i, w_j) = log_2\frac{hit(p_i\&w_j)}{hit(p_i) * hit(w_j)} \tag{2.4}$$

$$score = max(max(PMI\text{-}IR(p_i, w_j))) \tag{2.5}$$

畠山らは，新田らの手法の性能を向上させるために，有害極性単語の拡張や組み合わせを検討した [10]．石坂と山本 [11] が用意した悪口極性の基本単語によって有害極性単語を拡張し，またこれらを組み合わせて用いる手法を提案した．その結果，有害極性単語の適切な選別や組み合わせによって有害判定の精度が向上することを確認した．また，ウェブアンケートを実施し，人手で有害と判断された単語群を有害極性単語として用いたときの精度を比較した．その結果，人手で有害と判定された種単語を使用した場合，多くの有害語は高い極性値を持つことがわかり，また種単語の選択によって有害判定の精度が大きく変わることを示した．

2.2　攻撃的キーワードを含むテキストを利用した研究

あらかじめ用意した攻撃的キーワードをクエリとしてテキストを検索し，攻撃的テキストの候補として収集し，これを基にテキストの攻撃性を判定するモデルを機械学習する試みがいくつか行われている．

尾崎らは，Twitter(現 X) 上の投稿が不適切であるか否かを分類する手法を提案した [17].「死ね」「消えろ」「嫌い」「罵り」「馬鹿」「クズ」「ゴミ」「ボケ」「アホ」「いじめ」の 10 個の攻撃的キーワードを用意し，各キーワードを含むテキストを 1,000 個ずつ，計 10,000 個のテキストを収集した．そのテキスト群を人手でアノテーションすることによってデータセットを構築し，SVM によってテキストが攻撃的か否かを分類するモデルを学習した．交差検証による実験の結果，正解率の平均は 0.9701 であったと報告している．

大友と張は，Twitter(現 X) 上におけるネットいじめの自動検出を目的とし，その検出に貢献度の高い単語を選定し，選定した単語に基づいてラベル付きデータセットを構築し，SVM や決定木などの様々な機械学習手法によってツイートがネットいじめに該当するか否かを分類した [15]．いじめの検出に貢献する単語を選定する際は，まず文献 [11, 14, 10] の研究で用いられていた 36 個の「いじめ単語」を使用し，これを含む 2,349,052 個のツイートを収集した．次に，いじめ単語の数に応じたスコアリングによって，この中から 3,450 個のツイートを抽出した．これらのツイートに対して，Yahoo!クラウドソーシングを利用してアノテーションを行った．ここでのアノテーションでは，1 ツイートに 3 人のワーカーがいじめ文か非いじめ文かを判定した．

荒井らは，日本語の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の種類をアノテーションするためのガイドラインを設計した [6]．さらに，在日外国人に関連する差別語をキーワードにテキストを検索・収集し，このガイドラインに従い，攻撃の対象者と攻撃の種類を付与したデータセットを構築した．

Founta らは，Twitter における様々な種類の攻撃行為を総合的に調査し，攻撃行為を包括的に分類できるアノテーションスキームを提案した [9]．まず最初に，収集した 3,200 万件のツイートから，アノテーションに使用するツイートを選別する必要があったが，非攻撃的なツイートと比較して攻撃的なツイートは圧倒的に少ないという問題があった．そこで，ランダムサンプリングのほかに，攻撃的である可能性の高いテキストをブーストして抽出する手法を取り入れた．具体的には，感情分析の結果強い負の極性（極性のスコアが −0.7 未満）を持ち，少なくとも 1 つの攻撃的な単語を含むツイートを攻撃的である可能性のあるテキストとして抽出した．また，このスキームによってアノテーションされた 8 万件のラベル付きデータセットを公開した．

2.3 攻撃的キーワードを手がかりとしないデータセット構築手法

攻撃的キーワードを直接的に利用してデータセットを構築するアプローチには，網羅的な攻撃的キーワードを用意することが難しいこと，限られた攻撃的表現を含むテキストのみから構成されたデータセットが構築されやすいことなどの問題がある．これに対し，攻撃的キーワード以外の情報を手がかりに攻撃性判定のためのデータセットを構築する試みも行われている．

ユーザに着目した研究　牧元と徳永は，BiLSTM(Bidirectional Long Short-Term Memory) や ALBERT[12] 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 [13]．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単語を含む投稿を複数回行ったユーザの投稿から収集されている．このときに用いた攻撃的単語は，畠山らが実施した人手のアンケートにおいて有害とされた種単語群から選ばれた 17 単語であった．

話題に着目した研究　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) を構築した [19]．データセットを構築する際，政治的な話題には攻撃的な反応を得やすいとした知見から,「she is」や「to:BreitBartNews」などの政治的な話題のキーワードをクエリとした検索によってデータを収集した．

2.4 本研究の特徴

本研究では，攻撃的な単語を含むテキストだけではなく，そのような単語を用いずに暗黙的に他者を攻撃するテキストも含めて，多様な攻撃的表現を正確に検出することを目的とする．そこで，攻撃的表現の多様性という観点から，関連研究を考察する．

2.1 節で述べた攻撃的キーワードとの共起度を用いた手法では，極めて少ない攻撃的極性を持つ単語との共起度を測ることでテキストの攻撃性を判定していた．しかし，攻撃的な単語と文書内で共起しないが攻撃的である単語，フレーズ，表現が存在する可能性があり，これらを含む攻撃的テキストを正確に分類できない可能性がある．

2.2 節で述べた研究では，あらかじめ用意した攻撃的な単語を含むテキストを収集する方法が用いられている．しかし，特定の攻撃的な表現に偏ったデータセッ

トが構築されやすく，それから学習されたモデルは既知の攻撃的単語が使われない攻撃的表現を検出することは難しいと考えられる．

2.3 節では，攻撃的単語を含むテキストを複数回投稿しているユーザの投稿を収集した研究と，攻撃的テキストが多く寄せられることが見込まれる政治的な話題語に基づいてデータを収集した研究を紹介した．これらの収集方法では，2.2 節で述べた研究と比較して多様な攻撃的表現を含むテキストを獲得できると考えられる．しかし，掲示板やソーシャルメディアサービスにおいては，限られたユーザによって構成され，一般的な人との交流がほとんどない閉じたコミュニティが存在することも多く，その中では汎用的ではない特殊な表現や言い回しが使われることがある．したがって，これらの研究では，あらかじめ用意した攻撃的単語をよく使用する集団や政治について批判的な集団など，比較的狭いコミュニティに特有の攻撃的表現しか獲得できない可能性がある．Founta らの研究 [9] においては，攻撃的単語を含むテキストだけでなく，ランダムサンプリングによって選ばれたテキストもデータセットに含まれており，最終的に構築されたデータセットにおける攻撃的テキストの中には攻撃的単語を含まないものもあると考えられる．しかし，ランダムサンプリングによって得られた攻撃的テキストは約 4%程度と少なく，多様な攻撃的表現を得るのに十分とは言えない．

また，攻撃性のラベルが付与されたデータセットを構築する際には，上記のようなデータ収集に関する問題の他に，アノテーションのコストが高いという問題も存在する．多くの先行研究では，作業者が攻撃的か否かをアノテーションすることによってデータセットを構築していた [17, 16, 15, 6, 9, 13, 19]．しかし，これには作業者の負担が大きく，また数多くの不快な攻撃的な表現を読まなくてはならないという心理的な負担も大きいというデメリットがある．

以上をまとめると，攻撃性のラベルが付与されたデータセットを構築する既存研究には，(1) 攻撃的な表現の多様性に欠ける，(2) アノテーターに対する負担が大きい，という 2 つの課題がある．本研究では，Zampieri らの研究 [19] のように，攻撃的なテキストを集めやすい話題に着目する．ただし，政治の話題に限るのではなく，多くの話題に関する攻撃的な表現を収集する．このため，いわゆる炎上という現象に着目し，炎上に対する他者の返信や反応を収集することで，より多様な攻撃的な表現の獲得を目指す．また，提案手法は，人手で炎上ツイートを選別する必要があるが，それに対するリプライやリツートは自動的に収集できる．そのため，大量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必要はない．

第3章　データセット構築

3.1　炎上ツイートと非炎上ツイート

1.2 節で述べたように，攻撃性のラベルが付与されたデータセットを構築するにあたり，炎上現象について着目する．本研究では,「炎上ツイート」を，モラルに欠ける，または非常識的な言動に関する投稿，もしくはそれをまとめて報告する投稿であるとする．炎上ツイートに対しては，攻撃的な表現を用いて非難する反応が多く寄せられることが予想される．一方，そのような反応が見込まれない，動物の話題，非政治的な話題，新商品話題等のツイートを非炎上ツイートとする．図 3.1 に炎上ツイート・非炎上ツイートとそれぞれに対する反応の例を示す．炎上ツイートに対する反応には攻撃的表現が含まれているが，非炎上ツイートに対する反応には見られない．

3.2　データ収集とラベル付け

Twitter において，フォロワー数が多く，炎上現象に関する様々な話題を取りあげて投稿しているユーザを選び，そのユーザの投稿から，特に反応の多いツイートを人手で選別する．これを炎上ツイートとし，その投稿 ID を記録する．次に，その炎上ツイート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」とする）を攻撃的テキストとして収集する．同様に，動物の話題，非政治的な話題，新商品発売等の普通のニュースなど，他者から非難が集まりにくいことが予想されるツイートを非炎上ツイートとして投稿 ID を記録し，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．

Twitter では，あるツイートに対する他のユーザの反応には，直接のリプライと引用リツイートが存在するが，本研究ではどちらも収集する．引用リツイートとは，リツイート（あるツイート投稿をタイムラインに共有すること）にコメントを付与することのできる機能である．収集には Twitter API を利用した．リプライの収集には `https://api.twitter.com/2/tweets/search/recent?query=conversation_id:TWEETID`，引用リツイートの収集には `https://api.twitter.com/2/tweets/TWEETID/quote_tweets` のエンドポイントをそれぞれ使用した．ただし，リプライを収集する際には，Twitter API の仕様により，直接のリプライのみを取り出すのではなく，リプライに対するリプライなど，基点となる炎上ツイー

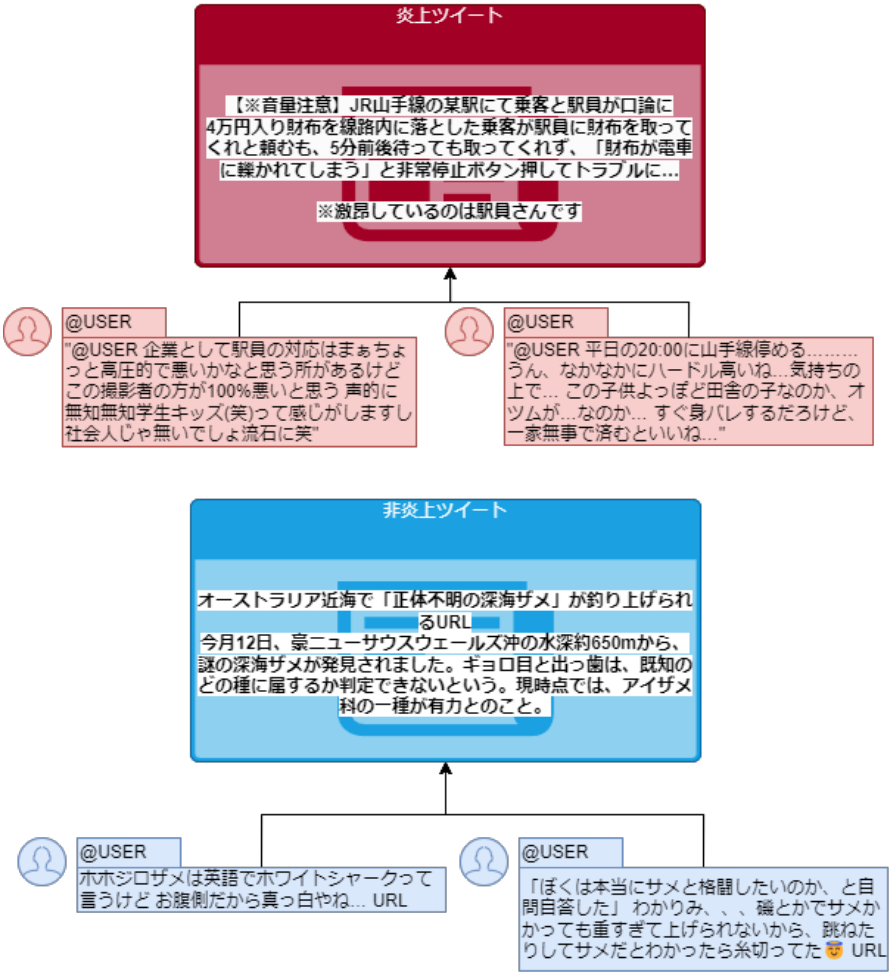


図 3.1: 炎上ツイート・非炎上ツイートとそれに対する反応の例

ト・非炎上ツイートから展開されるすべてのリプライを収集対象とする．図 3.2 は上記のデータ収集の手続きを図解したものである．

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．収集したデータの統計を表 3.1 に示す．炎上ツイートとして 12 件，非炎上ツイートとして 69 件のツイートを選び，それらに対する反応をおよそ 20,000 件ずつ収集した．

3.3 考察

炎上ツイートに対する反応が実際に攻撃的テキストであるかを予備的に評価する．具体的には，炎上ツイートもしくは非炎上ツイートに対する反応が攻撃的テ

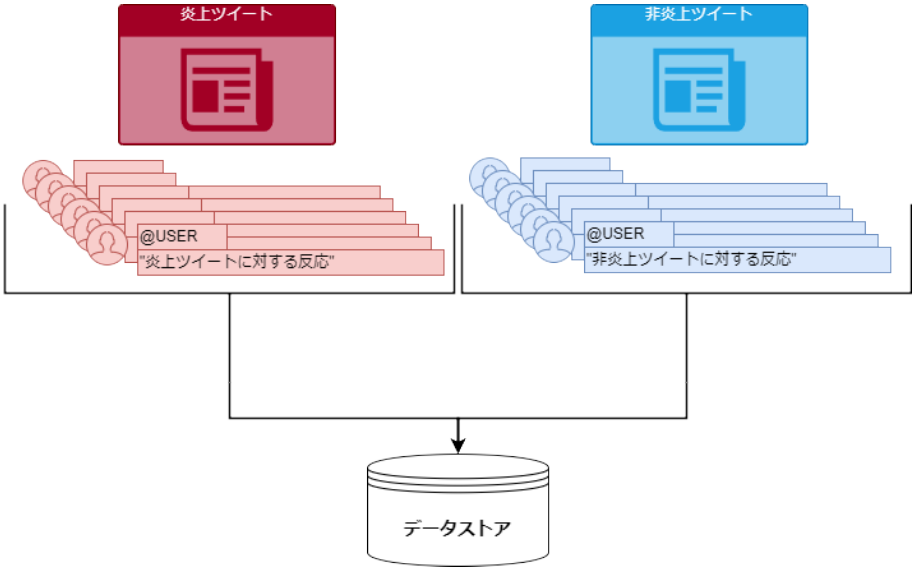


図 3.2: 炎上・非炎上ツイートに対する反応の収集

表 3.1: 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

キストかどうかを人手で判定する．収集した全てのデータを人手で評価するのは困難であるため，ランダムサンプリングによる標本調査を行う．検査するサンプル数は式 (3.1) に従って決める．N は調査対象とするサンプルの母数，z は信頼度，e は許容誤差，p は回答比率を示す．予備調査では，信頼度，許容誤差はそれぞれ 90%，10%とした．回答比率 p は不明のため，最も必要サンプリング数 n が大きくなる 50%とした．n 件のサンプルをランダムに選択し，それが攻撃的か否かを人手で判定した．

$$n = \frac{Nz^2p(1-p)}{e^2(N-1) + z^2p(1-p)} \tag{3.1}$$

サンプルの母数 (N)，検査数 (n), 攻撃的テキストと判定されたテキストの割合を表 3.2 に示す．攻撃的と判定されたツイートは全体の 30.9%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は，そのほとんどが非攻撃的であった．

次に，炎上ツイートまたは非炎上ツイートに対する反応のうち攻撃的キーワードを含むものについて，それが攻撃的テキストであるかを検証する．文献 [11, 14, 10] を参考に，表 3.3 に示す 38 個の攻撃的キーワードのリストを作成した．次に，炎上・非炎上ツイートへの反応でかつこれらの攻撃的キーワードのいずれかを含む

表 3.2: データセットの予備調査の結果 (AR は攻撃的テキストの割合)

	母数 (N)	検査数 (n)	AR
炎上ツイートへの反応	20,396	68	30.9%
非炎上ツイートへの反応	20,045	68	1.5%
攻撃的キーワードを含む反応	660	62	69.4%

ツイートの中から，62 件のツイートをランダムに選択し，それが攻撃的かを人手で評価した．その結果を表 3.2 の最後の行に示す．攻撃的テキストの割合は 69.4% であり，炎上ツイート全体の 30.9%と比べてかなり高かった．したがって，炎上ツイートの反応の多くは攻撃的ではないが，攻撃的なキーワードが含まれているものは攻撃的である可能性が高いことがわかった．

表 3.3: 攻撃的キーワードの一覧

うざい	きもい	イボヲタ	ウザい	カス
キチガイ	キモい	クズ	クズマスゴミ	ゴキヲタ
脱糞	ダセー	チョン	バカサヨ	ビッチ
蛆虫	ブサイク	ブス	ホモ	殴る
マジキモ	ヤリマン	不細工	厨	殺す
嫌い	孤独	愚民	死ね	死ねよ
殺せ	池沼	消えろ	無能	目糞
童貞	糞尿	糞虫		

第4章　攻撃的テキストの判定

4.1　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) [7] を用いる．BERTは，左右双方向の文脈を学習できるように設計されており，モデルの構造としてはTransformer[18] のEncoderを多層に重ねたものとなっている．Masked Language Model(MLM)，Next Sentence Prediction (NSP) によって事前学習を行い，その後適用したい自然言語処理タスクのデータセットを用いてモデルのパラメタをファインチューニングする．

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を攻撃性スコアを算出するモデルとしてファインチューニングする．ただし，3.3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．また，攻撃性のスコアを予測するモデルではなく，炎上・非炎上ツイートデータにおいて炎上している話題と炎上していない話題を識別するモデルが学習される可能性もある．そのため，テータセットのラベル誤りの訂正とモデルの学習を交互に繰り返すことでより精度の高いモデルの構築を目指す．ラベル誤りを訂正するために，訓練データの構築方法を改善する「初期データの作成方法」と，モデルを漸進的に学習する「モデルの学習手法」の2つのアプローチを検討する．

4.2　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．

4.2.1 手法 i(intact)

3.1 項で述べたように，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そのままラベル付けする手法．ただし，3.3 項で述べたように，特に攻撃的テキストには多くの誤りが含まれている．

4.2.2 手法 ii(PtoN)

炎上ツイートに対する反応のうち，攻撃的でないツイートを自動的に識別し，そのラベルを攻撃的から非攻撃的に修正する手法を提案する．PtoN とは，正例 (Positive Sample; この場合は「攻撃的」) から負例 (Negative Sample; この場合は「非攻撃的」) への修正を意味する．この手法の概要を図 4.1 に示す．4.2.1 項で作成したラベル付きデータについて，攻撃的ラベルを持つ正例テキスト群（赤色）のテキストのうち，非攻撃的ラベルを持つ負例テキスト群（青色）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．同図の「正例修正」の操作の詳細を以下に説明する．

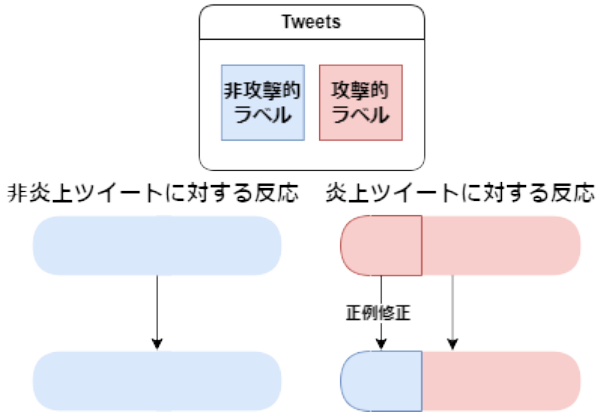


図 4.1: 手法 ii(PtoN) の概要

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 (4.1) で算出する．

$$NS(p) = \mathrm{ave}_{n_i \in TOP_5(p)}\ sim(p, n_i) \tag{4.1}$$

ここで，$sim(p, n_i)$ は正例 p と負例 n_i のコサイン類似度，$\mathrm{TOP}_5(p)$ 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．なお，$sim(p, n_i)$ におけるコサイン類似度の計算は式 (4.2) の通りである．

$$sim(p, n_i) = \frac{\boldsymbol{p} \cdot \boldsymbol{n_i}}{\|\boldsymbol{p}\|\|\boldsymbol{n_i}\|} \tag{4.2}$$

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 [5] 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．

この手法は，データセットの正例の多くは実際には攻撃的ではないため，ある正例に対し意味や表現がよく似ている負例が見つかった場合には，その正例を負例に修正するべきという考えに基づく．

4.2.3 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) か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．この手法の概要を図 4.2 に示す．この図では，これまでは 1(赤色のデータ) または 0(青色のデータ) のスコアのみが付与されていたツイートに対し，「単語 bi-gram に基づくスコアリング」によって [0,1] のスコアを付与する様子を表している．

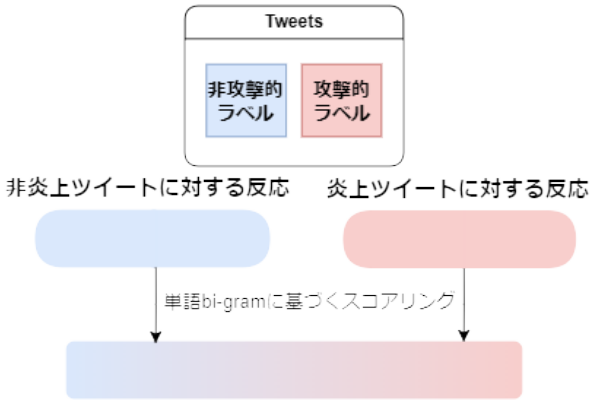


図 4.2: 手法 iii(scoring) の概要

以下，攻撃性のスコアを付与する方法の詳細を述べる．

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 (4.3) 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．

$$\mathrm{R}(bg_i) = \begin{cases} \dfrac{\mathrm{F}_P(bg_i)}{\mathrm{F}_N(bg_i)} & \text{if } \mathrm{F}_P(bg_i) \geq \mathrm{F}_N(bg_i) \\ -\dfrac{\mathrm{F}_N(bg_i)}{\mathrm{F}_P(bg_i)} & \text{if } \mathrm{F}_P(bg_i) < \mathrm{F}_N(bg_i) \end{cases} \tag{4.3}$$

4. 以下のいずれかの条件を満たす単語 bi-gram を除外し，残された単語 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 (4.4) のように定義する．

$$OS(t) = \begin{cases} \mathrm{Nor}\left(\dfrac{\sum_{bg_i \in T \cap B} \mathrm{R}(bg_i)}{|T \cap B|}\right) & \text{if } T \cap B \neq \emptyset \\ \mathrm{median}(OS(t)) \text{ in } N & \text{if } T \cap B = \emptyset \end{cases} \tag{4.4}$$

   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．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．上記の関数 Nor による正規化の手順は以下の通りである．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．

直観的には，炎上ツイートの反応ツイート群 P に頻出する単語 bi-gram を多く含むツイートに高い攻撃性スコアを，非炎上ツイートの反応ツイート群 N に頻出する単語 bi-gram を多く含むツイートに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 のどちらかに極端に偏って出現する単語 bi-gram を削除している．

ツイート t に出現する単語 bi-gram のうち，ステップ 4. で作成した単語 bi-gram の集合 B に属するものがひとつもない場合には，すなわち式 (4.4) において $T \cap B = \emptyset$ の場合には，その攻撃性スコアを計算できない．このとき，3.3 項で述べたように，収集したデータセットの大部分は非攻撃的テキストであることから，B に属する単語 bi-gram をひとつも含まないツイートも非攻撃的テキストであると仮定する．また，その攻撃的スコアは非攻撃的テキストの攻撃的スコアの中央値を仮に与える．

4.3 モデルの学習手法

この節では，攻撃性判定モデルを学習する 3 つの手法を提案する．なお，初期の訓練データは 4.2 項で述べた手法のいずれかを用いて作成する．

4.3.1 手法 A(vanilla)

初期データを用いて BERT を一度だけファインチューニングする手法である．最も基本的な手法であるが，初期データのラベル誤りがモデルの学習に悪影響を与える可能性がある．

4.3.2 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である．その概要を図 4.3 に示す．まず，初期訓練データ D_{origin}(赤: 攻撃的ラベルが付いたデータ, 青: 非攻撃的ラベルが付いたデータ) 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_1 を得る．このモデルは 4.3.1 項で述べた手法で得られるモデルと同じである．

次に，ラベルなし訓練データ $D_{unlabel}$ を用意する．図 4.3 では右上の黒色のデータがこれに相当する．$D_{unlabel}$ は，初期訓練データ D_{origin} から攻撃性のラベルを除去したデータであり，ツイート群自体は初期訓練データと同一である．判定モデル M_1 を用いてラベルなし訓練データ $D_{unlabel}$ の攻撃性スコアを予測する．攻撃性スコアの上位 500 件 (攻撃的) と下位 500 件 (非攻撃的) のテキストに対し，1(攻撃的) または 0(非攻撃的) のスコアを付与し，更新訓練データ $D_{updated(1)}$ とする．また，残されたテキストを $D_{unlabel}$ とする．ここでは，攻撃性スコアの極性が高いほどラベルの信頼度が高いという考えに基づき，信頼度が高いデータのみから新しい訓練データを構築している．

その後，$D_{updated(1)}$ 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 を得て，M_2 で攻撃性スコアを予測した $D_{unlabel}$ のうち，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，1(攻撃的) または 0(非攻撃的) のスコアを付与し，これを $D_{unlabel}$ から除き，$D_{updated(1)}$ と合わせて $D_{updated(2)}$ とする．以下，モデル M_i による $D_{unlabel}$ に対する攻撃性スコア予測と，攻撃性スコア上位下位 500 件と $D_{updated(i)}$ を合わせて得た $D_{updated(i+1)}$ によるモデル M_{i+1} の再学習を繰り返す．この手法では，モデルの学習に用いるデータは (初期訓練データを除いて) 漸進的に増加する．最終的に $D_{unlabel} = \emptyset$ となった時点で訓練データが完成したとみなし，この訓練データを用いて最終的な判定モデル M を学習する．

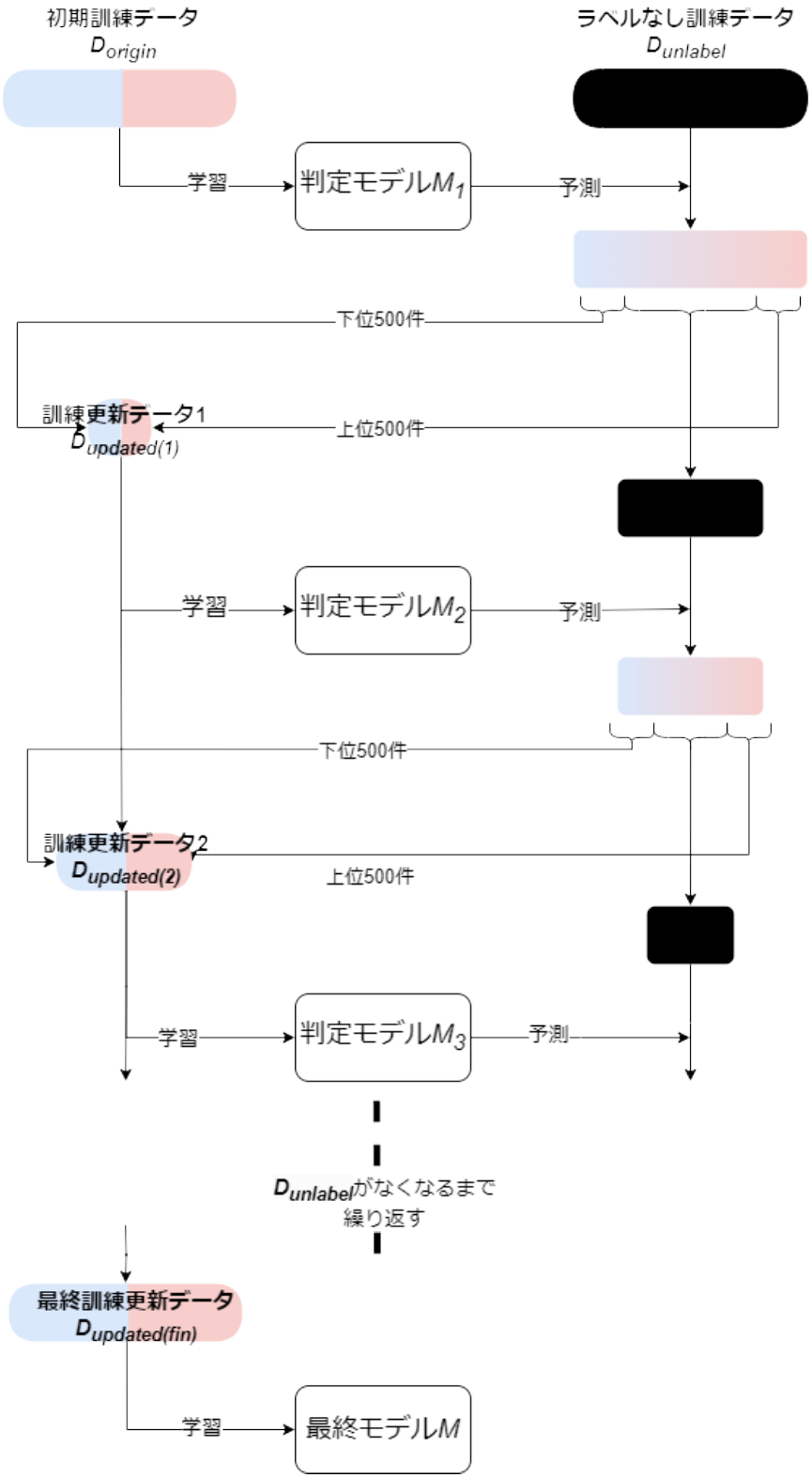


図 4.3: 手法 B(bootstrap) によるモデル学習の流れ

4.3.3 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．この手法の処理の流れを図 4.4 に示す．まず，初期訓練データ D_{origin} から最初の攻撃性判定モデル M_1 を学習する．次に，モデル M_1 によって訓練データの各ツイートの攻撃性スコアを推定し，スコアが付与されたデータセット $D_{score(1)}$ を得る．以下，攻撃性判定モデル M_i によって攻撃性スコアを推測することで [0,1] のスコアが付与されたデータ $D_{score(i)}$ を得て，これを用いて新しい判定モデル M_{i+1} を学習することを繰り返す．データに付与する攻撃性のスコアが収束した時点での訓練データを最終の訓練データ (図 4.4 における「最終スコアデータ」) とし，最終的な攻撃性判定モデルを学習する．

データセットに付与する攻撃性スコアが収束したかは以下の手続きで判定する．データセットの全てのサンプルに対する判定モデル M_i と M_{i+1} による予測スコアの分布 $\boldsymbol{S_i}$ と $\boldsymbol{S_{i+1}}$ をそれぞれ記録する．そして，$\boldsymbol{S_i}$ と $\boldsymbol{S_{i+1}}$ 間のユークリッド距離が閾値 T_c 以下になったとき，反復学習を停止する．ユークリッド距離の計算は式 (4.5) の通りである．a はデータセットにおけるサンプル数，$S_{i(t)}$ ならびに $S_{i+1(t)}$ は i 回目ならびに $i+1$ 回目の反復操作で得られたツイートの攻撃性スコアを表す．

$$d(\boldsymbol{S_i}, \boldsymbol{S_{i+1}}) = \sqrt{\sum_{t=1}^{a}(S_{i+1(t)} - S_{i(t)})^2} \tag{4.5}$$

以上の手法は攻撃性スコアの計算 (ラベル付け) を繰り返すため，手法 C(relabeling) と呼ぶ．本手法は，訓練データに付与する攻撃性スコアの再計算により攻撃性判定モデルの性能が次第に向上し，初期の訓練データでは攻撃的ラベルが付与されたテキストのうち実際には非攻撃的であるもののスコアが低い値に修正されていくことを期待している．

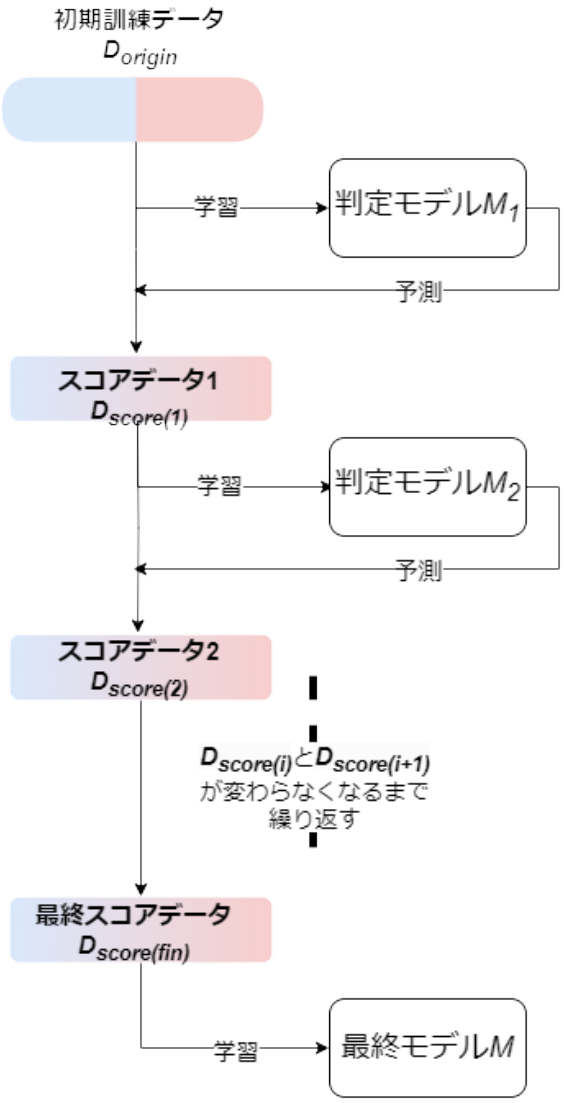


図 4.4: 手法 C(relabeling) によるモデル学習の流れ

第5章　評価

5.1　実験データ

提案手法を評価するために，ツイートテキストに対して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する．3.1 節で述べた炎上・非炎上ツイートから，炎上ツイート 2 件，非炎上ツイート 2 件を選択し，それらに対する反応ツイート計 280 件を抽出した．これらに対して，3 名の被験者 (いずれも 20 代男性) が 0(非攻撃的)，1(攻撃的)，2(その他) のいずれかのラベルをアノテーションした．それぞれのラベルの定義を表 5.1 に示す．これは被験者に提示したアノテーションの指針でもあり，文献 [4] で用いられていたアノテーションの指針に準拠している．また，非文の場合を除き,「その他」のラベルは原則として選ばず,「攻撃的」「非攻撃的」のいずれかのラベルを付与するよう依頼した．

被験者間のアノテーションの一致度を確認するため，Fleiss's κ 係数を求めた．Fleiss's κ 係数は，被験者が 3 人以上のときにアノテーションの一致度を測る指標である [8]．Fleiss's κ 係数は 0.511 となり，被験者によって付与されたラベルはある程度一致していると言える．

表 5.1: 攻撃性ラベルの定義

ラベル	ラベル名	説明
0	非攻撃的	中立的 or 礼儀正しい，好意的なコメント
1	攻撃的	憎悪的 or 攻撃的 or 無礼 or 理不尽で ユーザーに見解の共有を諦めさせる可能性のあるコメント
2	その他	非文，もしくは攻撃的かどうか確信が持てないもの

3 名の被験者によるアノテーションの結果を基に，最終的なテストデータを作成した．まず，1 名以上によって「非文もしくは判定不能」と判定されたツイート 7 件を除去した．次に，各ツイートの最終的なラベルを 3 名の被験者によって付与されたラベルの多数決によって決定した．

攻撃性判定モデルの学習に使用する訓練データは炎上・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．前処理として,「ユーザ名」「URL」「改行」を削除した．また，攻撃的ラベルが付与されたツイート群と非攻撃的ラベルが付与されたツイート群の両方に出現するツイートは訓練データから除去した．これ

に該当するツイートは,「草」「笑った」など比較的短く，定型文のようなツイートが多かった．また，この処理は初期データの作成方法として提案した手法 i(intact)，手法 ii(PtoN) とで個別に実施した．攻撃的スコアを与える手法 iii(scoring) については，ツイートを攻撃的ツイート群と非攻撃的ツイート群に分けることはできないため，この処理は行わなかった．

テストデータ，手法 i の初期訓練データ，手法 ii の初期訓練データの統計を表 5.2 に示す．手法 ii については，正例のうち負例と類似しているもののラベルを「攻撃的」から「非攻撃的」に修正した後の結果である．訓練データでは攻撃的ラベルと非攻撃的ラベルが付与されたツイートの数はほぼ同じだが，テストデータは攻撃的ラベルより非攻撃的ラベルが付与されたツイートの方が多い．手法 iii について，ツイートに付与した攻撃的スコアのヒストグラムを図 5.1 に示す．横軸は攻撃的スコア，縦軸はそのスコアが割り当てられたツイートの数を表す．予想に反し，大半のツイートには 0.5 付近のスコアが割り当てられていることがわかる．すなわち，手法 iii で構築した初期の訓練データにおいては，ツイートが攻撃的が非攻撃的かを明確に区分できていない．

表 5.2: 実験データの統計

	非攻撃的	攻撃的
手法 i(intact) の初期訓練データ	18,955	19,671
手法 ii(PtoN) の初期訓練データ	20,757	17,869
評価データ	203	70

5.2 実験設定

5.2.1 比較する手法

ベースラインモデルとして，攻撃的単語を手がかりとして構築した訓練データから学習した攻撃性判定モデルを用意する．具体的には，3.2 節で述べた方法で収集した炎上・非炎上ツイートデータの中から表 3.3 の攻撃的キーワードのいずれかを含むツイートを抽出し，これを正例（攻撃的テキスト）とする．また，正例と同数の負例を非炎上ツイートに対する反応からランダムに抽出する．その結果，正例と負例の件数はそれぞれ 653 件となった．このデータを用いて BERT モデルを一度ファインチューニングしたモデルをベースラインとし，提案手法と比較する．

提案手法は，初期データの作成手法 (i,ii,iii)，モデルの学習手法 (A,B,C) をそれぞれ組み合わせた 9 つの手法 (i×A, i×B, i×C, ii×A, ii×B, ii×C, iii×A, iii×B, iii×C) とし，これらを比較する．また，ベースラインの訓練データのサイズは提案手法のそれよりも小さいが，一般に機械学習されたモデルは訓練データが多いほど性能が高い．訓練データ量が同じという条件下でベースラインと提案手法を比較するた

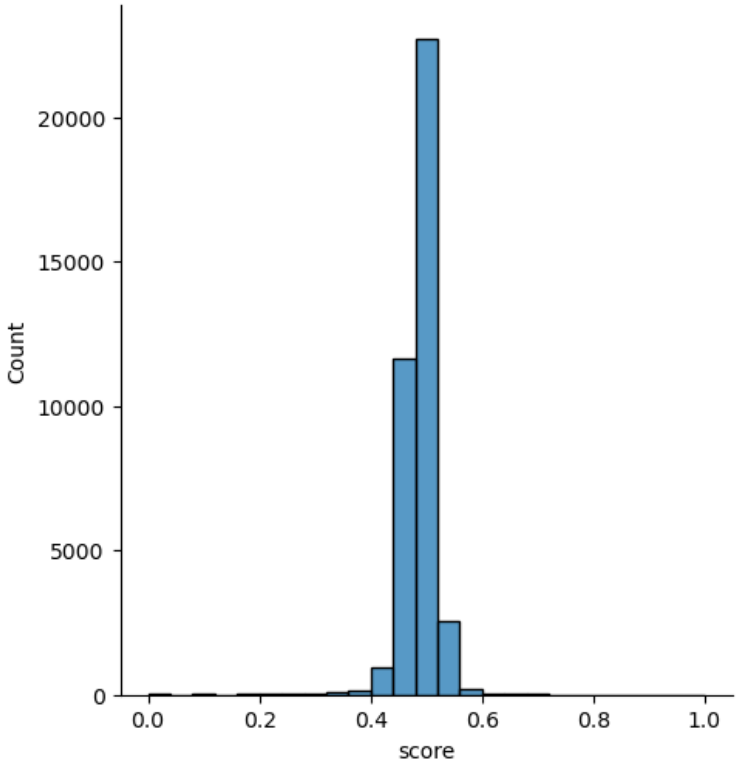


図 5.1: 手法 iii(scoring) によって与えられた攻撃的スコアの分布
(中央値:0.49，平均値:0.49，標準偏差:0.038)

め，ベースラインと同じサンプル数の訓練データから手法 i × A によって学習された攻撃性判定モデルも評価する．炎上ツイートに対する反応から正例を 653 件，非炎上ツイートに対する反応から負例を 653 件ランダムに選択し，これを訓練データとして BERT をファインチューニングする．以下，この手法を「手法 i × A(small)」と呼ぶ．

5.2.2 評価基準

本実験は，テキストが攻撃的か否かを判定する分類問題を評価タスクとする．一方，4.1 節で述べた通り，提案手法によって学習されるモデルは攻撃性スコアを予測する回帰モデルである．提案手法の攻撃性判定モデルを用いても，予測したスコアが閾値以上または以下のときに攻撃的または非攻撃的と判定することにより，分類問題を解くことができる．しかし，その性能は閾値の設定に大きく依存する．そこで，本研究ではモデルの評価指標として，Receiver Operatorating Characteristic Curve(ROC 曲線) ならびに Precision-Recall Curve(PR 曲線) の Area Under Curve(AUC) を用いる．ROC 曲線は閾値の変化による False Positive Rate(FPR) と True Positive Rate(TPR) の変動を表し，PR 曲線は閾値の変化による Precision

と Recall の変動を表す．AUC は ROC 曲線と PR 曲線の下の領域の広さを表す指標であり，様々な閾値に対するモデルの総合的な良さを評価する．また，ROC-AUC と PR-AUC は [0,1] の値をとる．

FPR,TPR,Recall,Precision は式 (5.1)，(5.2)，(5.3) のように定義される．なお，TPR と Recall の定義は同じである．

TP(True Positive)
　モデルが正例データを正しく正例と判定した数

TN(True Negative)
　モデルが負例データを正しく負例と判定した数

FP(False Positive)
　モデルが負例データを誤って正例と判定した数

FN(False Negative)
　モデルが正例データを誤って負例と判定した数

FPR
　負例データのうち，モデルが誤って正例と判定した割合

$$FPR = \frac{FP}{TN + FP} \tag{5.1}$$

TPR(Recall)
　正例データのうち，モデルが正しく正例と判定した割合

$$TPR = \frac{TP}{TP + FN} \tag{5.2}$$

Precision
　モデルが正例と判定したデータのうち，実際に正例データであった割合

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

5.2.3　モデルの学習

本研究では 4.1 節で述べた通り，攻撃性判定モデルとして BERT モデルを使用する．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[1] と BERT-large[2] を使用する．BERT をファインチューニングする際のハイパーパラメータなどを表 5.3 に示す．

4.3.3 項で述べたように，手法 C(relabeing) において，訓練データに付与した攻撃的スコアの分布 S_i と S_{i+1} のユークリッド距離が閾値 T_C 以下になったとき，反復学習を停止する．本実験では，BERT base モデルを用いたときは $T_C = 7$，BERT large モデルを用いたときは $T_C = 10$ と設定した．

表 5.3: BERT モデル学習時の設定

訓練データにおける検証データの比率	0.05
Epoch 数	3
バッチサイズ	16
学習率	$3e^{-5}$

5.3 実験結果と考察 (BERT base)

5.3.1 AUC による評価 (BERT base)

BERT base モデルを使用したときの 9 つの提案手法，ベースライン，ならびに手法 i×A(small) の ROC-AUC，PR-AUC を表 5.4 に示す．

一番良い手法は，ROC-AUC では手法 ii(PtoN) と手法 C(relabeling) の組み合わせ，PR-AUC では手法 i(intact) と手法 C(relabeling) の組み合わせであった．初期データの作成手法 i,ii,iii を比較すると，ROC-AUC においては ii(PtoN) が最も成績が良いが，PR-AUC では i(intact) の方が成績が良くなる傾向が見られた．手法 iii(scoring) は他の手法と比較して悪い成績が得られた．このことから，初期データの作成手法として，手法 i(intact) または ii(PtoN) が有望であると言える．

モデルの学習手法 A,B,C を比較すると，ROC-AUC,PR-AUC のいずれにおいても C(relabeling) において最高の評価値が得られた．このことから，手法 C(relabeling) の有効性が確認された．ただし，手法 A(vannila) との差は全体的にはそれほど大きくはなかった．反対に，手法 B(bootstrap) の AUC 値は全体的に低いことから，この手法はあまり有効ではなかったと言える．

ベースラインと比較すると，手法 i×A，i×C，ii×A，ii×C，iii×A については ROC-AUC，PR-AUC ともにベースラインを上回った．ベースラインを下回った手法のうち 3 つは手法 B との組み合わせであった．ベースラインと同じサイズの訓練データからモデルを学習する提案手法 i×A(small) とベースラインを比較すると，ROC-AUC では提案手法が上回り，PR-AUC ではベースラインが上回った．

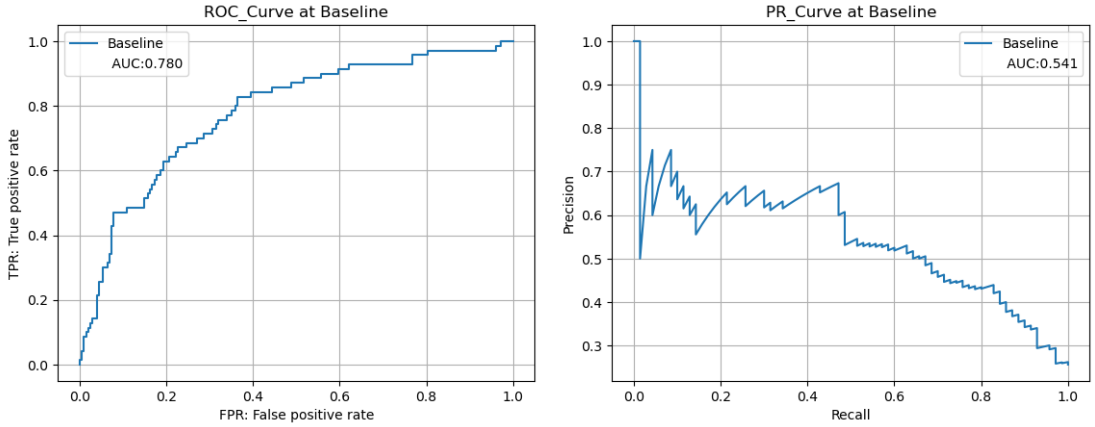
表 5.4: 攻撃性判定モデルの評価結果 (BERT base)

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
ベースライン	0.780		
手法 i×A(small)	0.804		

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
ベースライン	0.541		
手法 i×A(small)	0.496		

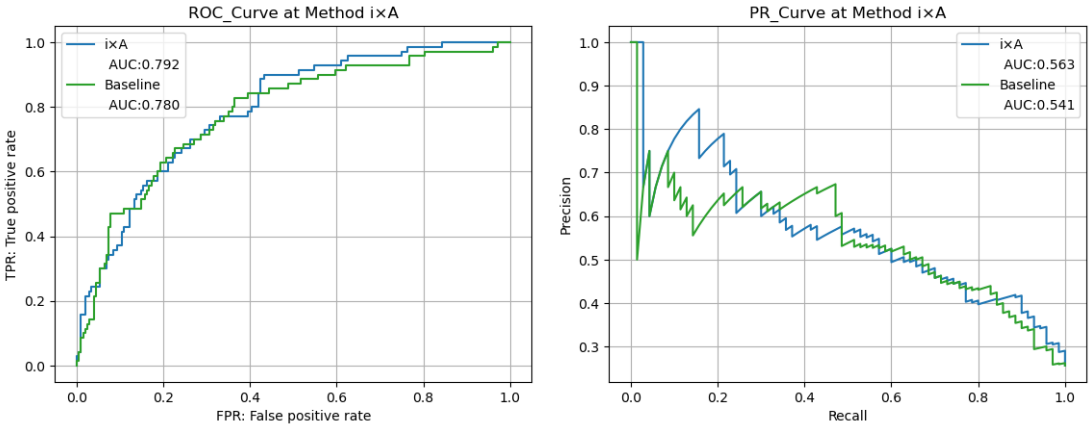
5.3.2 PR 曲線に関する考察 (BERT base)

ベースラインモデルの ROC 曲線と PR 曲線を図 5.2 に示す．同様に，9 つの提案手法について，ROC 曲線と PR 曲線を図 5.3-図 5.11 に示す．比較のため，これらの図ではベースラインの結果も掲載している．緑の線はベースライン，青の線は提案手法である．最後に，手法 i×A(small) の結果を図 5.12 に示す．



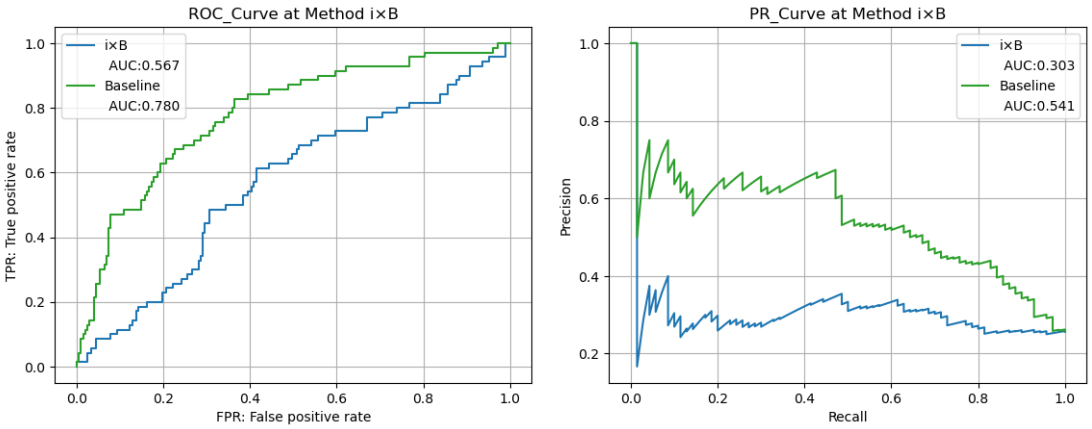
(a) ROC 曲線　　(b) PR 曲線

図 5.2: ベースライン学習の実験結果 (BERT base)



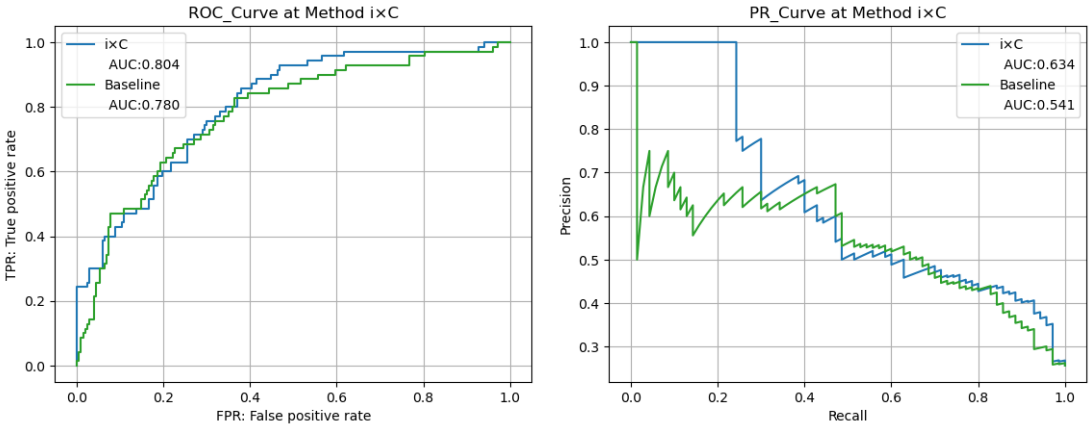
(a) ROC 曲線　　(b) PR 曲線

図 5.3: 手法 i(intact)×A(vanilla) の実験結果 (BERT base)



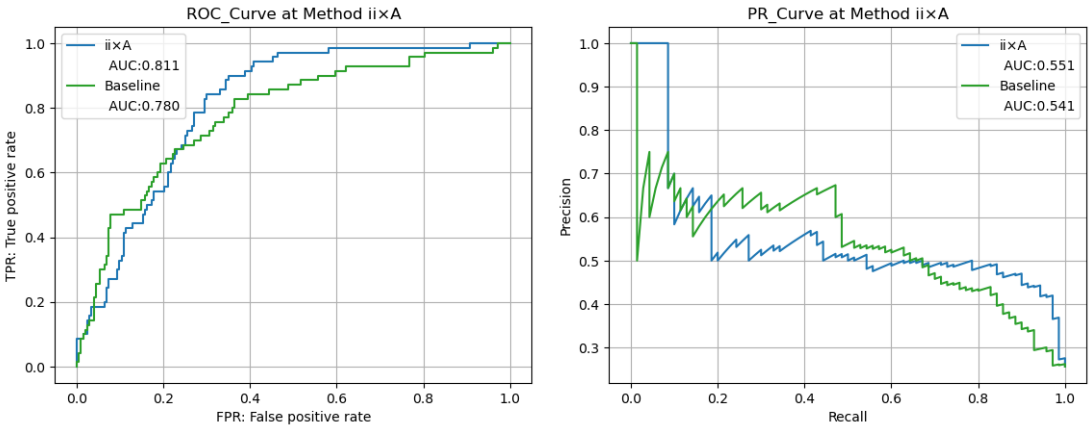
(a) ROC 曲線　　(b) PR 曲線

図 5.4: 手法 i(intact)×B(bootstrap) の実験結果 (BERT base)



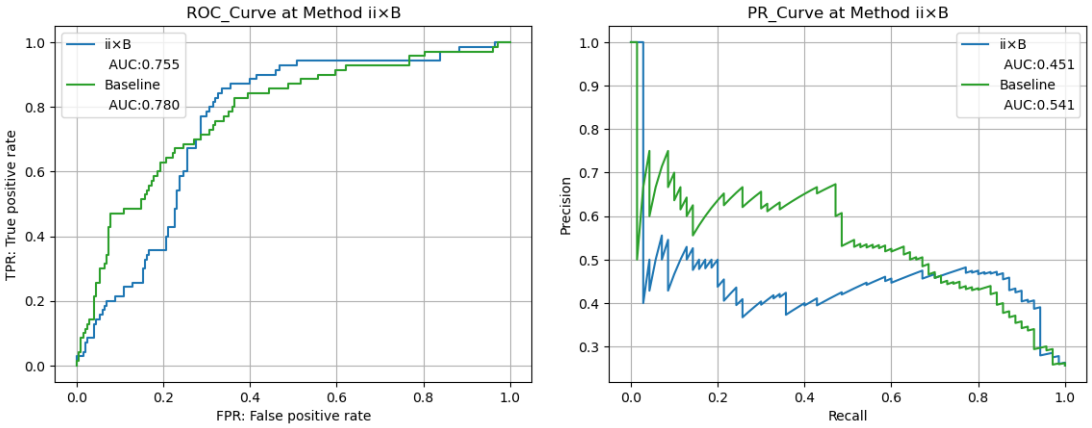
(a) ROC 曲線　　(b) PR 曲線

図 5.5: 手法 i(intact)×C(relabeling) の実験結果 (BERT base)



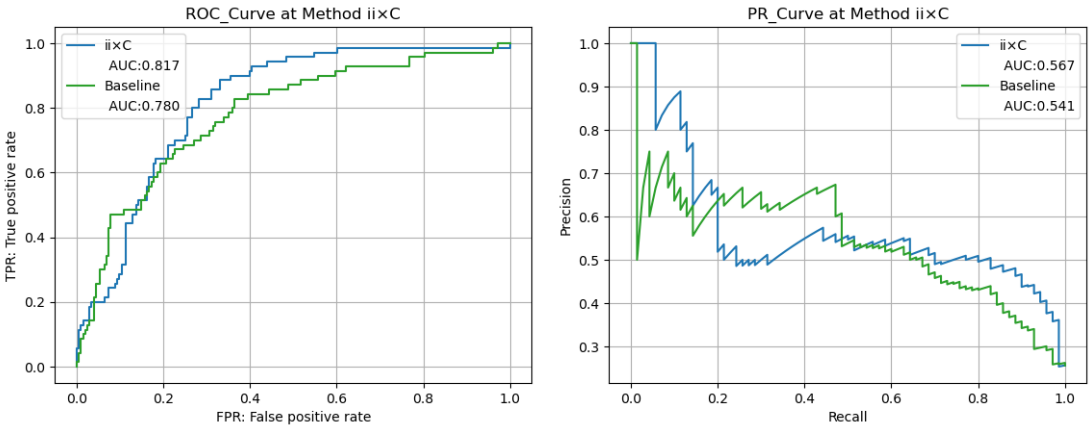
(a) ROC 曲線　　(b) PR 曲線

図 5.6: 手法 ii(PtoN)×A(vanilla) の実験結果 (BERT base)



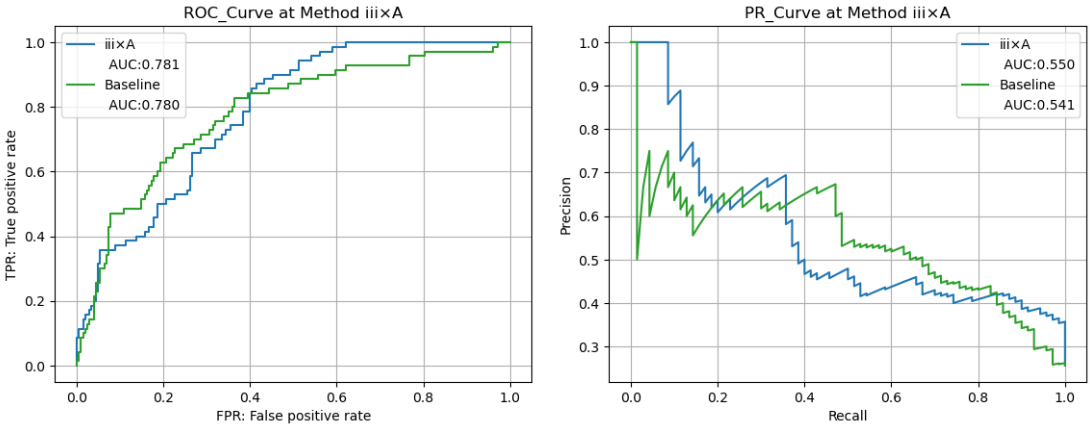
(a) ROC 曲線　　(b) PR 曲線

図 5.7: 手法 ii(PtoN)×B(bootstrap) の実験結果 (BERT base)



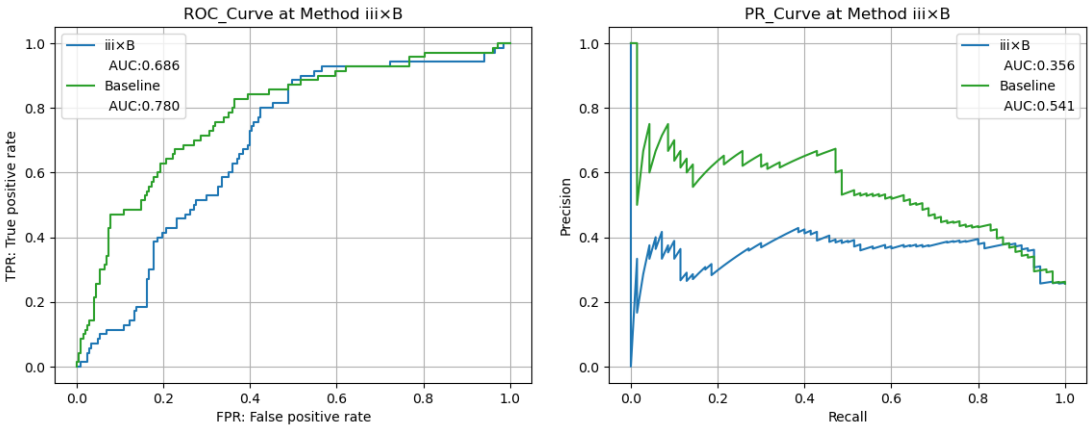
(a) ROC 曲線　　(b) PR 曲線

図 5.8: 手法 ii(PtoN)×C(relabeling) の実験結果 (BERT base)



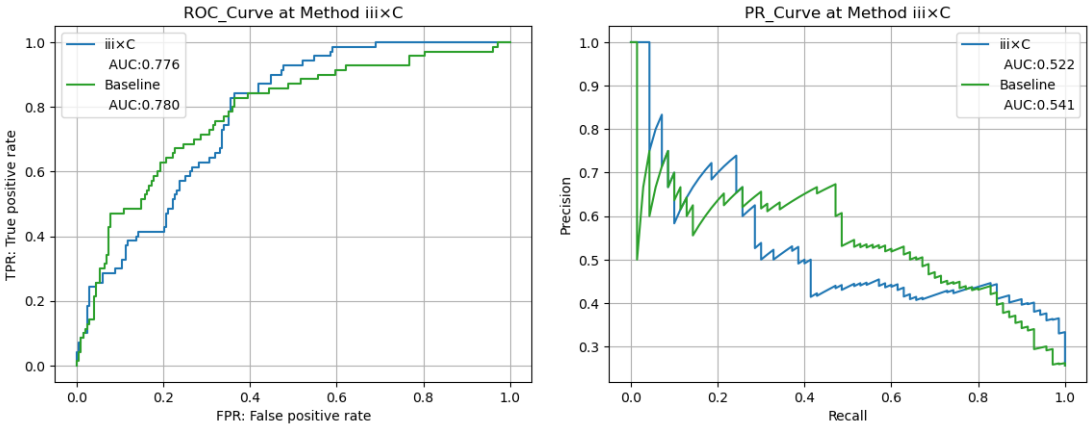
(a) ROC 曲線　　(b) PR 曲線

図 5.9: 手法 iii(scoring)×A(vanilla) の実験結果 (BERT base)



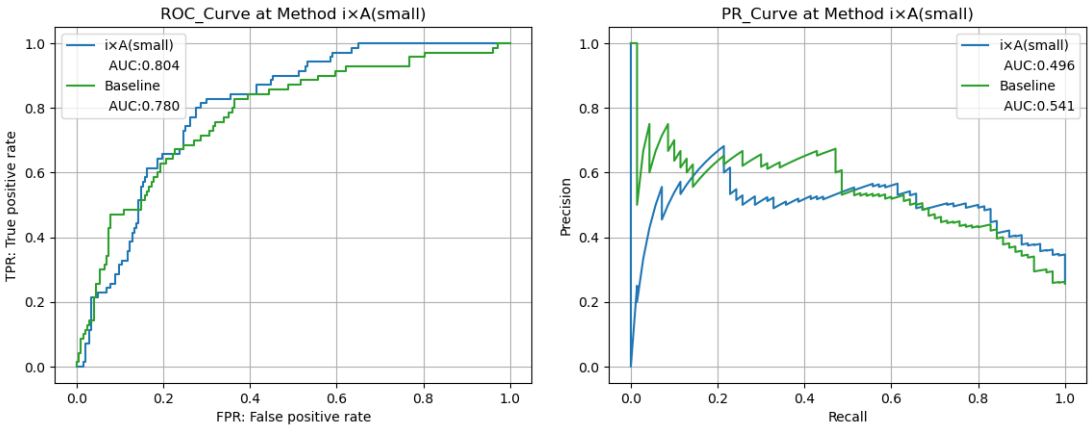
(a) ROC 曲線　　(b) PR 曲線

図 5.10: 手法 iii(scoring)×B(bootstrap) の実験結果 (BERT base)



(a) ROC 曲線　　(b) PR 曲線

図 5.11: 手法 iii(scoring)×C(relabeling) の実験結果 (BERT base)



(a) ROC 曲線　　(b) PR 曲線

図 5.12: 手法 i×A(small) の実験結果 (BERT base)

図 5.3-図 5.11 において，ベースラインと提案手法の PR 曲線を観察すると，提案手法はおおむね 3 つのパターンに分けることができる．

1. Recall が低いときには提案手法の方が Precision が高く，中程度になるとベースラインの方が高くなり，Recall が高いときには再び提案手法の方が高くなるパターン．具体的には以下の手法.

   i(intact)×A(vanilla)(図 5.3), i(intact)×C(relabeling)(図 5.5),
   ii(PtoN)×A(vanilla)(図 5.6), ii(PtoN)×C(relabeling)(図 5.8),
   iii(scoring)×A(vanilla)(図 5.9), iii(scoring)×C(relabeling)(図 5.11)

2. Recall が低いときにはベースラインの方が Precision が高いが，Recall が高くなるにつれて提案手法の方が高くなるパターン．具体的には以下の手法.

   ii(PtoN)×B(bootstrap)(図 5.7)

3. 全体的にベースラインを下回るパターン．具体的には以下の手法.

   i(intact)×B(bootstrap)(図 5.4), iii(scoring)×B(bootstrap)(図 5.10)

パターン 1 やパターン 2 に該当する手法が多いことから，全体的には，炎上ツイート・非炎上ツイートの反応からデータセットを構築する提案手法は，Recall が高いという条件下で，攻撃的キーワードを手がかりにデータセットを構築するベースラインより優れていると言える．この理由として，攻撃的ラベルの品質に差があること，データセットにおける攻撃的テキストの多様性に差があることが考えられる．3.3 節で述べた通り，攻撃的キーワードを含むテキストが実際に攻撃的である可能性は，炎上ツイートに対する反応のそれと比較して高いことから，データセットに付与されるラベルの品質（正確性）が高いと言える．Recall が低い場面，すなわち Precision が高く，FPR が低い場面では，ノイズの少ないラベル付きデータから学習したモデルの方がテキストの攻撃性を正確に予測できたと言える．しかし，攻撃的キーワードを含むテキストのみでは攻撃的表現の多様性に欠けていたため，攻撃的テキストを漏れなく検出することが難しくなり，Recall が高いことを重視する場面では，ベースラインの性能が落ちたと考えられる．一方，提案手法では，攻撃的キーワードを必ずしも含まない多様な攻撃的表現をデータセットに含み，このことが Recall が高い状況で優れた性能を発揮した原因であると考えられる．訓練データ数が同じという条件での比較でも，すなわちベースラインと手法 i×A(small) との比較でも，図 5.12 に示すように，Recall が低い範囲ではベースラインの方が成績が良いが，Recall が高い範囲では提案手法が上回っている．PR-AUC の比較では提案手法はベースラインに劣るものの，多様な攻撃的テキストを検出することが要求される場面ではベースラインよりも優れている．以上の考察から，提案手法によって多様な攻撃的表現を含むデータセットを自動構築できたと言える．

ただし，パターン 1 に該当する手法について，Recall が低い範囲では提案手法がベースラインを上回っている．このことは，これまでに考察した攻撃性ラベルの品質の差や攻撃的表現の多様性の差では説明できないものである．このような提案手法の挙動の原因を解明するためには，追加実験も含めたさらなる検討が必要である．

提案手法がベースラインよりも明らかに悪いパターン 3 には，手法 B(bootstrap) を使用した 2 つのモデルが該当した．このことは，5.3.1 項で報告した実験結果において，手法 B の成績が全体的に悪かったことと一致する，手法 B(bootstrap) はやはり有効な手法とは言えない．この原因として，反復学習の初期の段階ではモデルの訓練データの量が少なく，その時点で学習されたモデルの性能が低いこと，またそのような性能の低いモデルによって誤った攻撃性スコアが付与されたことなどが考えられる．

5.4 実験結果と考察 (BERT large)

5.4.1 AUC による評価 (BERT large)

BERT large モデルを使用したときの 6 つの提案手法，ベースライン，並びに手法 i×A(small) の ROC-AUC，PR-AUC を表 5.5 に示す．なお，この実験では，BERT base モデルを用いたときの実験結果が悪かった手法 B(bootstrap) の評価は省略している．

一番良い手法の組み合わせは，5.3 節で述べた BERT base を用いた実験とは異なり，ROC-AUC，PR-AUC のいずれでも手法 i(intact) と A(vanilla) の組み合わせであった．また，初期データの作成手法 i，ii，iii の比較においても，BERT large を用いたときは，両 AUC で手法 i が最も成績が良かった．また，手法 iii が他の手法と比較して評価値が低いことは BERT base による実験と同様であった．これらを踏まえると，初期データの作成手法の有効度は，i>ii>iii の順であると言える．また，モデルの学習手法 A と C を比較すると，5.3 節で報告した BERT base モデルの実験では手法 C が最高の成績を収めたが，BERT large モデルを用いた実験では ROC-AUC，PR-AUC のいずれにおいても A(vanilla) の方が評価値が高かった．しかし，手法 i 以外では手法 C が手法 A を上回る場合もある．これらを踏まえると，モデル学習手法の有効性の順序は A≃C>B であると言える．

BERT base モデルを用いた結果 (表 5.4) と BERT large モデルを用いた結果 (表 5.5) を比較すると，ROC-AUC と PR-AUC の両方について BERT large モデルの方が評価値が高かったのは手法 i × A のみであった．

提案手法とベースラインと比較すると，手法 i×A，i×C については，ROC-AUC と PR-AUC の両方で提案手法がベースラインを上回った．ROC-AUC でのみベースラインを上回ったのは手法 ii×A，ii×C であった．手法 iii を用いた 2 つの手法 (iii×A と iii×C) は ROC-AUC，PR-AUC ともにベースラインを下回った．これら

の結果は，初期訓練データの作成方法については手法 i,ii,iii の順で良いという上記の考察と一致するものであった．ベースラインと同じサイズの訓練データからモデルを学習する提案手法 i×A(small) とベースラインを比較すると，ROC-AUC では提案手法が上回り，PR-AUC ではベースラインが上回った．この結果は BERT base モデルを用いたときと同じであった．

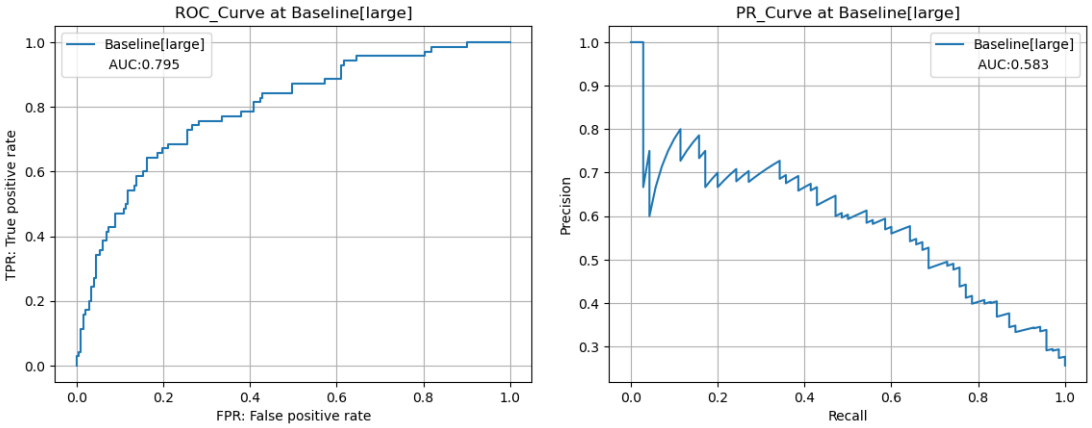
表 5.5: 攻撃性判定モデルの評価結果 (BERT large)

ROC-AUC	A(vanilla)	C(relabeling)
i (intact)	**0.838**	0.828
ii (PtoN)	0.803	0.810
iii (scoring)	0.775	0.780
ベースライン	0.795	
i×A(small)	0.812	

PR-AUC	A(vanilla)	C(relabeling)
i (intact)	**0.633**	0.617
ii (PtoN)	0.525	0.526
iii (scoring)	0.530	0.484
ベースライン	0.583	
i×A(small)	0.551	

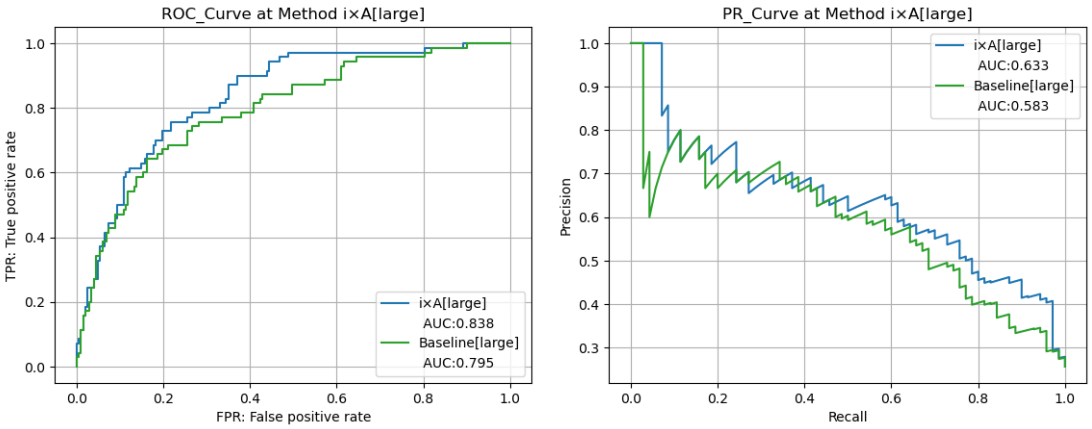
5.4.2 PR 曲線に関する考察 (BERT large)

ベースラインモデルの ROC 曲線と PR 曲線を図 5.13 に示す．同様に，6 つの提案手法について，ROC 曲線と PR 曲線を図 5.14-図 5.19 に示す．これらの図にはベースラインの結果も掲載する．最後に，手法 i×A(small) の結果を図 5.20 に示す．



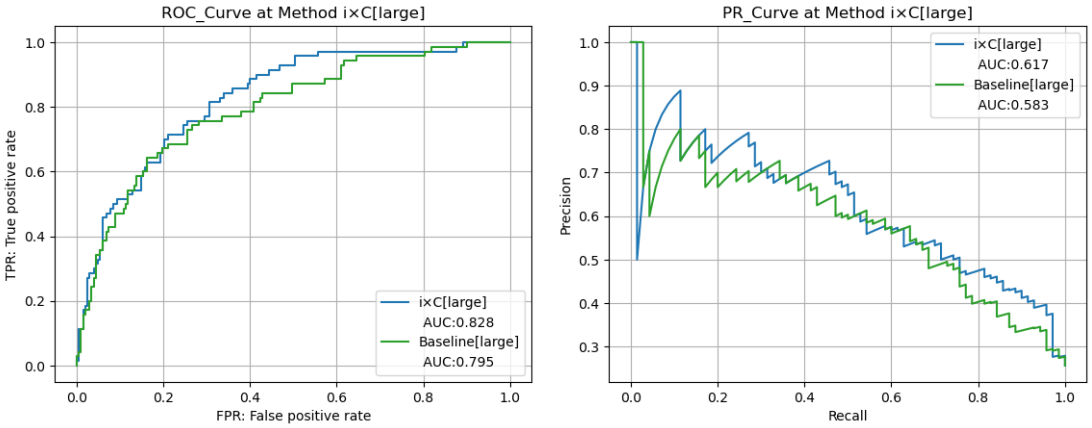
(a) ROC 曲線　　(b) PR 曲線

図 5.13: ベースラインの実験結果 (BERT large)



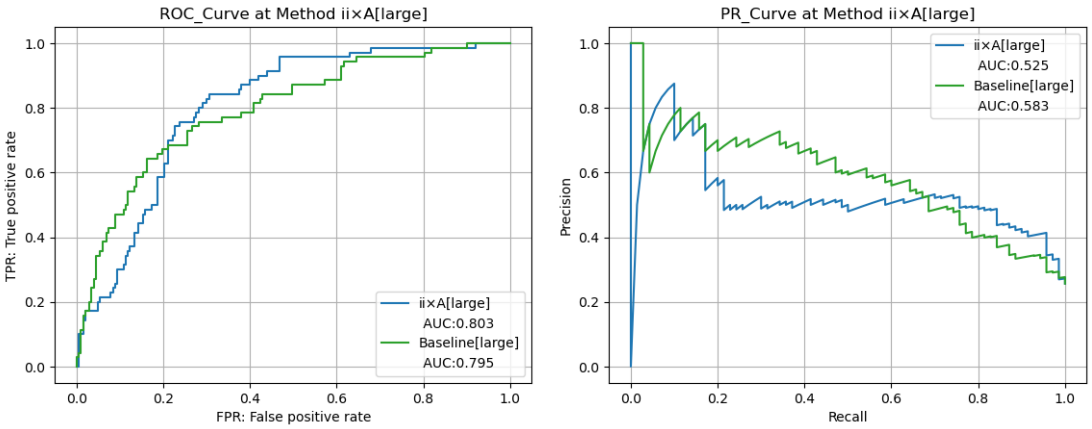
(a) ROC 曲線　　(b) PR 曲線

図 5.14: 手法 i(intact)×A(vanilla) の実験結果 (BERT large)



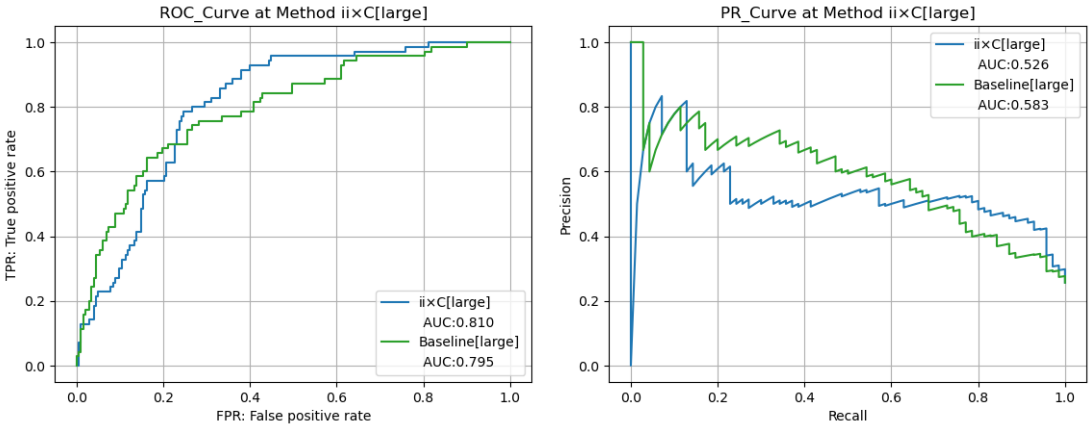
(a) ROC 曲線　　(b) PR 曲線

図 5.15: 手法 i(intact)×C(relabeling) の実験結果 (BERT large)



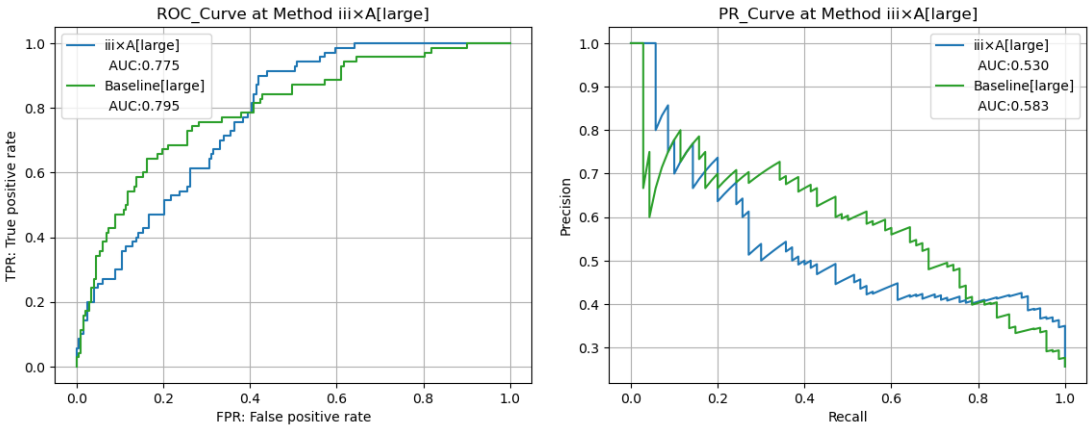
(a) ROC 曲線　　(b) PR 曲線

図 5.16: 手法 ii(PtoN)×A(vanilla) の実験結果 (BERT large)



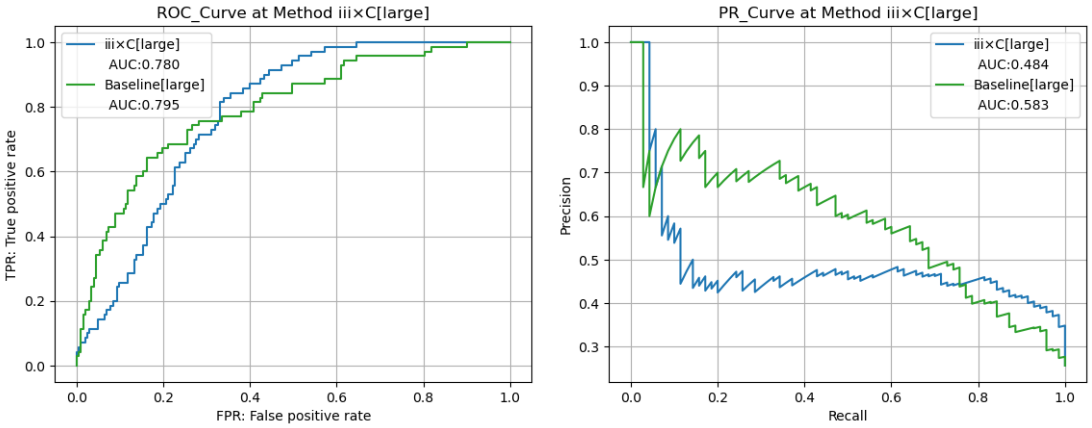
(a) ROC 曲線　　(b) PR 曲線

図 5.17: 手法 ii(PtoN)×C(relabeling) の実験結果 (BERT large)



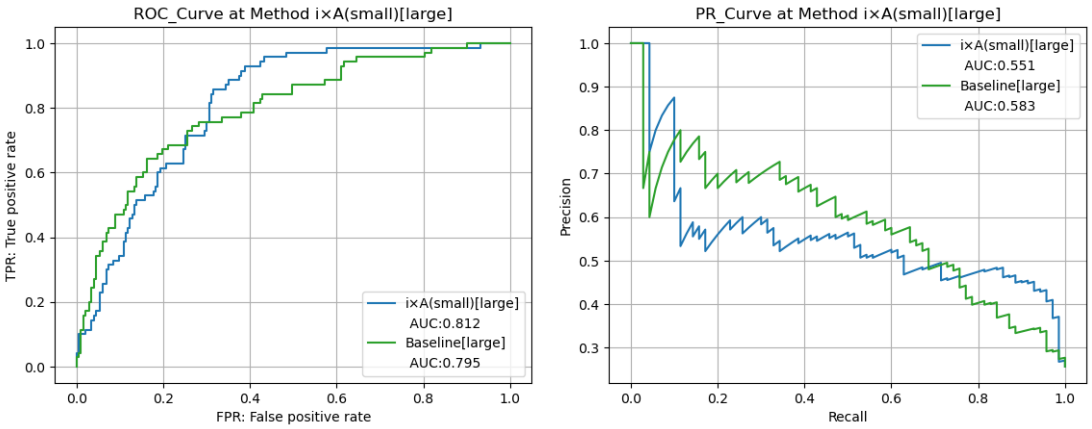
(a) ROC 曲線　　(b) PR 曲線

図 5.18: 手法 iii(scoring)×A(vanilla) の実験結果 (BERT large)



(a) ROC 曲線　　(b) PR 曲線

図 5.19: 手法 iii(scoring)×C(relabeling) の実験結果 (BERT large)



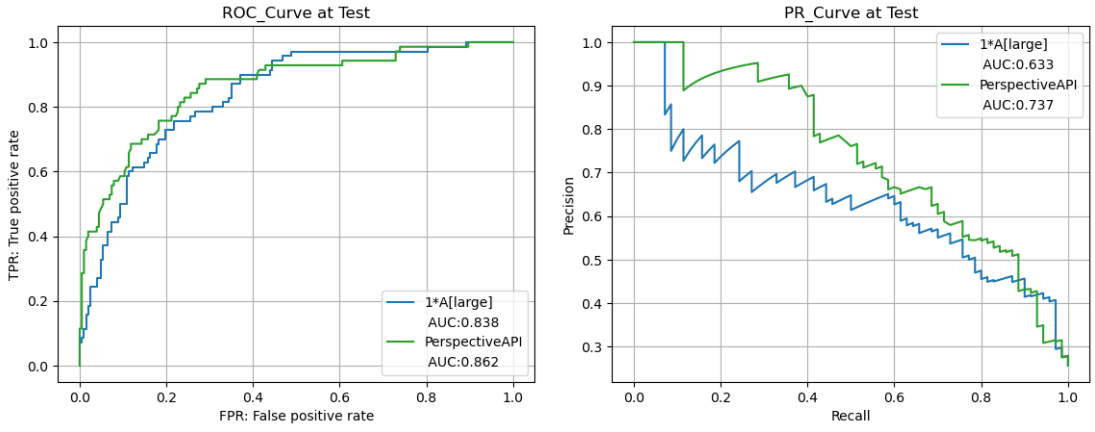
(a) ROC 曲線　　(b) PR 曲線

図 5.20: 手法 i×A(small) の実験結果 (BERT large)

PR 曲線に着目して提案手法ならびにベースラインを比較すると，提案手法は，訓練データのサイズが同じという条件での手法 i×A(small) も含めて，5.3.2 項で述べたパターン 1 もしくは 2 に当てはまる．すなわち，Recall が大きくなると提案手法の方が precision が高くなる．この結果は，BERT base モデルと BERT large モデルを用いた実験で共通している．したがって，BERT large モデルを用いた実験でも，提案手法が多様な攻撃的表現を集めたデータセットを構築し，また多様な攻撃的表現を含む攻撃的テキストの検出に優れていることが確認された．

5.5 Perspective API との比較

Perspective API[4] 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 [3]，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 と提案手法の中で最も AUC の高かった手法，すなわち BERT large モデルを利用し，手法 i(intact) と手法 A(vanilla) を組み合わせた手法の ROC 曲線と PR 曲線を図 5.21 に示す．緑の線は Perspective API，青の線は提案手法である．Perspective API の ROC-AUC は 0.862，PR-AUC は 0.737 であり，いずれも提案手法 (ROC-AUC 0.838，PR-AUC 0.633) を上回る．しかし，PR 曲線で Recall が高い付近では，提案手法の方が Precision が高い．既に述べたように，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．



(a) ROC 曲線　　(b) PR 曲線

図 5.21: 提案手法と Perspective API との比較

第6章　おわりに

6.1　本論文のまとめ

本論文では，多様な表現を含む攻撃的テキストを正確に判定するモデルを学習するために，攻撃的ラベルが付与されたデータセットを自動構築し，それを基に攻撃性の強さを推定するモデルを学習する手法を提案した．

ラベル付きデータセットの自動構築については，炎上しやすい話題を多く取り上げる Twitter ユーザの投稿から攻撃的反応が見込まれる炎上ツイートを人手で選別し，それに対する反応を収集した．また，攻撃的反応がほとんどないことが見込まれる話題のツイート（非炎上ツイート）をニュースアカウント等から人手で選別し，それに対する反応を収集した．炎上ツイートに対する反応に攻撃的，非炎上ツイートに対する反応に非攻撃的のラベルを仮に付与した．構築したデータセットの品質を予備的に調査したところ，非炎上ツイートに対する反応のほとんどは非攻撃的であったが，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．そこで，攻撃性判定モデルの学習と訓練データの誤りを訂正する処理を反復的に繰り返すことで，元のラベル付きデータの誤りの影響を軽減する手法を検討した．この際，最初のモデルを学習する初期データについて，元のデータセットの誤りを自動的に修正する手法も検討した．

初期データの作成方法については，最初に用意したデータセットをそのまま使用する手法 i(intact)，Sentence BERT を用いて非攻撃的テキストと類似した攻撃的テキストを探索し，そのラベルを攻撃的から非攻撃的に修正する手法 ii(PtoN)，テキスト中の単語 bi-gram が攻撃的テキスト群または非攻撃的テキスト群のどちらかに偏って出現する傾向を基に算出した [0,1] の攻撃性スコアを付与する手法 iii(Scoring) 手法を提案した．

攻撃性判定モデルの反復学習については，モデルを一度だけ学習する単純な手法 A(vanilla)，Bootstrap によって信頼度の高いサンプルから順にラベルを決定する手法 B(bootstrap)，学習したモデルによる攻撃性ラベルの修正と修正済み訓練データを用いたモデルの学習を収束するまで繰り返す手法 C(relabeling) を提案した．攻撃性判定モデルとして BERT を用いた．

評価実験では，手法 i,ii,iii と手法 A,B,C の組み合わせによる 9 つの提案手法ならびにベースライン手法を比較した．ベースライン手法は，本研究で構築したデータセットの中から攻撃的単語を含むテキストとそれを含まないテキストを 653 件ずつ選別し，これを用いてファインチューニングした BERT モデルとした．テス

トデータを作成するために，データセットの中から選別した 273 件のツイートに対し，20 代男性 3 名が攻撃的・非攻撃的のいずれかのラベルを付与し，その多数決によって最終的なラベルを決定した．テストデータに対して攻撃性の強さ (スコア) を予測し，スコアの大きさによってツイートが攻撃的か否かを分類し，ROC-AUC と PR-AUC を評価基準として提案手法を評価した．

評価実験の結果，BERT base を用いたとき，ROC-AUC が最大となったは手法 ii×C，PR-AUC が最大となったのは手法 i×C であった．全体としては，手法 C の成績が良く，手法 B の成績が悪い傾向が見られた．ベースラインとの比較においては，ROC-AUC,PR-AUC ともに上回ったものは手法 i×A,i×C,ii×A,ii×C,iii×A であった．学習モデルを BERT-large としたとき，ROC-AUC,PR-AUC 共に手法 i×A が最高の成績となった．以上を踏まえ，初期データの作成方法については i>ii>iii の順，モデルの学習手法については A≃C>B の順に有効であると結論付けた．

PR 曲線について提案手法とベースラインを比較すると，再現率が低い範囲と高い範囲では提案手法が優位であり，再現率が中程度の範囲ではベースラインが優位である傾向が観察された．本研究では多様な攻撃的表現を含む攻撃的テキストを正しく検出することを目的としているが，これが実現できると再現率が高くなる．再現率が高い範囲で提案手法がベースラインを上回ったことは，多様な攻撃的表現に対応するモデルが学習できていることを示唆するものである．テキストの攻撃性を判定する公開ツールである Perspective API との比較実験においても，ROC-AUC や PR-AUC については提案手法は Perspective API に劣るものの，PR 曲線で再現率が高い範囲では Perspective API を上回る結果が得られた．

6.2 今後の課題

評価実験では，多様な攻撃的表現を含むデータを構築し，それを基に再現率が高い場面で特に有効な攻撃性判定モデルを学習できたことが確認できたが，ROC-AUC や PR-AUC による評価では，提案手法の全てがベースラインを上回るわけではなかった．上記を踏まえて今後の課題を以下に挙げる．

ラベル誤りを削減する手法の改善　自動構築したデータセットのラベルの誤りを訂正する手法として，手法 ii(PtoN) や iii(Scoring)，あるいは手法 B(Bootstrap) や C(Relabeling) を提案したが，手法 iii と手法 B については，多くの場合でベースラインと比べて ROC-AUC や PR-AUC が低いことから，その効果は十分ではなかった．今後は手法 iii や手法 B を改善する方法や，ラベルの誤りを訂正する新しい手法を探究することが必要である．

提案手法の特徴の検証　既に述べたように，再現度が低い範囲において提案手法がベースラインを上回る明確な理由は不明であった．この原因を追求するために

提案モデルによる判定結果をより詳細に分析し，その長所や短所が明らかになれば，提案手法を改善する新たな手法やアプローチの発見につながる可能性がある．

大規模なデータセットの構築　本研究で構築したデータセットのサイズは 40,000 件程度であり，大規模なデータセットとは言えないものであった．理論的には炎上ツイートや非炎上ツイートに対する反応を更に収集することでデータセットの規模を拡大することができるが，本研究を進めている間に Twitter API の仕様が変更され，ツイートを容易に収集することが困難になった．一方，多様な攻撃的表現に対応するためには，より大規模なデータセットの構築は不可欠である．また，攻撃的表現は時間とともに変化することから，攻撃的テキストを定期的に収集することも求められる．したがって，継続的に攻撃的テキストを効率良く収集するスキームを考案することも重要な課題である．

参考文献

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

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

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

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

[5] 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.

[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] 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.

[8] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5, p. 378, 1971.

[9] 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. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12, , 2018.

[10] 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.

[11] 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.

[12] 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.

[13] 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).

[14] 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.

[15] 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, 第 2018 巻, 2019. B1-1.

[16] 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.

[17] 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.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 2017.

[19] 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.