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.