D-008

アスリートに対する SNS 上の誹謗中傷検出におけるリプライ元ツイートの有効性の検証
Verification of Effectiveness of Source Tweets in Detecting Defamatory Posts Targeting Athletes on Social Media

西谷 千乃与† 安藤 一秋‡
Senchiyo Nishitani Kazuaki Ando

1. はじめに

SNS を利用することで，アスリートの応援が容易になった．その一方で，SNS における誹謗中傷が問題になっている．2021 年の東京オリンピックでは，SNS における誹謗中傷により，多くのアスリートが苦しんだことは大きな話題となった．誹謗中傷を受けたアスリートは，試合で納得できるパフォーマンスを発揮できない可能性もあるため，今後を左右する大きな問題である．したがって，アスリートに対する誹謗中傷を早期に自動検出する技術が必要になる．

本研究では，SNS 上で発信されるアスリートに対する誹謗中傷を自動検出する手法の実現を目指す．本稿では，リプライツイートとリプライ元ツイートの 2 つを用いた実験により，誹謗中傷検出におけるリプライ元ツイートの有効性について検証する．

2. 関連研究

大友らの研究[1]では，Twitter 上のテキストを対象に，いじめ表現辞書と n-gram や Word2vec，Doc2vec といった複数の特徴量を用いたモデルにより，ネットいじめを自動検出する手法を提案している．6 種の分類モデルと特徴量を組み合わせた実験により，すべての特徴量を用いたロジスティック回帰の性能が最良となり，F 値で 0.921 が得られたと述べている．また，実験を通じて，いじめ表現辞書が正しい検出に貢献したと述べている．今後の課題として，目的にふさわしい基本単語を厳選し，いじめ表現辞書を改善する方法の検討を挙げている．

松本らの研究[2]では，煽り投稿を対象とし，BERT を用いて煽りを自動検出する手法を提案している．特に，リプライツイートとリプライ元ツイートを合わせたデータを用いて有効性を比較している．結果として，リプライツイートのみを学習・分類した場合の F 値が 0.715，両ツイートを学習・分類した場合の F 値が 0.711 となり，リプライ元ツイートを用いる効果が確認できなかったと述べている．今後の課題としては，検出性能の向上を図るために，人物同士の関係性などを特徴として追加することを挙げている．

本研究では，大友らの研究を参考に，悪口辞書を構築して誹謗中傷候補を収集し，得られた候補を 2 値分類する手法の実現を目指す．松本らの研究ではリプライ元ツイートを用いる効果が確認できなかったが，本稿では，アスリートに対する誹謗中傷検出におけるリプライ元ツイートの有効性について検証する．

3. 誹謗中傷候補ツイートの収集

本研究では，誹謗中傷を「他人を罵倒すること」や「他人の人格を否定すること」と定義する．以下，対象選手の選定法と選手本人に対するメンションツイート，選手のツイートに対するリプライツイート（誹謗中傷候補ツイート）を収集する方法について述べる．

3.1 悪口単語辞書

誹謗中傷を含むツイートには，誹謗中傷に関連する単語が含まれていることが多い[1]．そこで，誹謗中傷に関連する単語辞書（以降，悪口単語辞書）を利用して，誹謗中傷候補ツイートを抽出する．本研究では，西原ら[4]が用いた悪口単語 55 語を悪口単語辞書のベースに利用する．そして，アスリートを対象とする悪口単語 19 語[5]を追加し，74 語からなる悪口単語辞書を用いる．

3.2 誹謗中傷候補ツイートの抽出

誹謗中傷を含むツイートの分析とデータセット構築のために，誹謗中傷候補ツイートを抽出する．対象は，Twitter ランキング[3]のスポーツカテゴリ 14 種から 5 人ずつ選定したアスリート 70 選手である．選手本人のツイートに対するメンション・リプライツイートを，2022/6/19~2022/12/15 の期間に収集した結果，400,090 件が収集できた．その後，悪口単語辞書を用いて，9,887 件を抽出した．

4. 誹謗中傷検出手法の検討

14 種のスポーツカテゴリに対し，目視で確認可能な誹謗中傷候補ツイートを，誹謗中傷ツイート（正例），誹謗中傷ツイートではないツイート（負例）にラベル付けし，正例と負例のそれぞれ 503 件からなるデータセットを構築した．以降，このデータセットを用いて評価する．

4.1 実験設定

関連研究[1]と同様，線形サポートベクトルマシン（SVM），ロジスティック回帰（LR），ランダムフォレスト（RF），多層パーセプトロン（MLP）を分類器として使用する．素性については，bag-of-words，TF-IDF，Word2Vec（W2V）を使用する．また，追加素性として，1 ツイートにおける文字数，形容詞と副詞の数，敬語表現の数を使用し，分類器と素性をそれぞれ組み合わせて，precision，recall，F1-score で評価する．

また，事前学習済みモデルである BERT（東北大学 2023 版），RoBERTa（京都大学 2022 版，早稲田大学 2021 版，rinnna 社 2021 版），DeBERTa（東京大学 2023 版，京都大学 2022 版），ELECTRA（東京大学 2023 版），bigbird（早稲田大学 2021 版），LUKE（Studio Ousia2020 版）をファインチューニングして二値分類で誹謗中傷ツイートを検出する手法も検討する．訓練バッチサイズは，4，8，16，32 の 4 つで実験する．訓練バッチサイズと言語モデルをそれぞれ組み合わせて性能を評価する．

† 香川大学大学院創発科学研究科 Graduate School of Science for Creative Emergence, Kagawa University
‡ 香川大学創造工学部 Faculty of Engineering and Design, Kagawa University

4.2 誹謗中傷検出手法の実験結果

素性に基づいた検出手法と事前学習済みモデルに基づく検出手法の評価結果のうち，F1-score で最良値を得た手法をそれぞれ表 1 に示す．全ての手法において，最も高い F1-score はバッチサイズ 32 の東北大学が提供する BERT-v3 で 0.838 を得た．

表 1 カテゴリ別の分析結果の一部

	precision	recall	F1-score
東北大 BERT-v3 (size32)	**0.856**	0.822	**0.838**
W2V_SVM+全ての素性	0.740	**0.881**	0.804

5. リプライ元ツイートの有効性検証

誹謗中傷検出におけるリプライ元ツイートの有効性について検証する．

5.1 実験設定

正例・負例ツイート 503 件に対して，リプライ元ツイートを収集した結果，それぞれ 76 件と 129 件のリプライ元ツイートが得られた．そこで，正例の件数を負例の件数に揃えるため，正例ツイートを目視で新たに 53 件取得した．計 258 件のリプライ元ツイートを含むデータを使用して，リプライ元ツイートの有効性について検証する．なお，検出手法は 4.1 節と同様で，リプライツイートとリプライ元ツイートの両方を用いたデータとリプライツイートのみのデータでそれぞれ実験し，結果を比較する．

5.2 素性に基づく手法における検証

分類器と素性をそれぞれ組み合わせた手法の評価結果のうち，リプライ元ツイートを加えた実験とリプライツイートのみの実験において，F1-score が上位 3 件の結果をそれぞれ表 2 と表 3 に示す．

表 2 と 3 より，リプライ元ツイートを加えたデータで敬語表現の数を用いたロジスティック回帰の F1-score は 0.810 で全手法において最良値を得た．また，組み合わせた手法において F1-score はリプライ元ツイートを加えた方が高い値となり，誹謗中傷の検出に有効であることを確認した．

表 2：F1-score における上位 3 件の結果（リプライ元＋リプライ）

	precision	Recall	F1-score
Bow_LR+敬語	0.742	**0.885**	**0.810**
Bow_MLP+全て	**0.778**	0.808	0.793
Bow_MLP+品詞，敬語	0.697	**0.885**	0.780

表 3：F1-score における上位 3 件の結果（リプライのみ）

	precision	Recall	F1-score
Bow_LR+品詞，敬語	0.760	**0.731**	**0.745**
Tf-idf_SVM+文字数，敬語	0.760	**0.731**	**0.745**
Bow_LR+全て	**0.783**	0.692	0.735

5.3 事前学習済みモデルに基づく手法における検証

事前学習済みモデルは文脈を考慮するため，リプライ元ツイートをリプライツイートの前に加えたデータ（正順）とリプライ元ツイートをリプライツイートの後ろに加えたデータ（逆順）の 2 つのデータを用いるとともに，リプライツイートのみのデータとも比較する．

バッチサイズ別における各言語モデルに基づく検出手法の評価結果のうち，正順，逆順，リプライツイートのみのデータで F1-score が最良値を得た上位 3 件の手法を表 4，表 5，表 6 に示す．

表 4，5，6 より，正順のデータでバッチサイズ 8 の京大 RoBERTa に基づく手法が全手法における F1-score で最良値 0.889 を得た．また，リプライツイートのみのデータと比較すると，F1-score はリプライ元ツイートを加えた正順，逆順の方が高い値となり，リプライ元ツイートの有効性を確認した．表 4 と 5 より，F1-score は逆順より正順の方が高い値を得ており，正順データの有効性を確認した．

表 4：F1-score における上位 3 件の結果（正順）

	precision	Recall	F1-score
京大 RoBERTa (size8)	0.857	**0.923**	**0.889**
早稲田 RoBERTa (size4)	0.913	0.808	0.857
京大 RoBERTa (size8)	**0.952**	0.769	0.851

表 5：F1-score における上位 3 件の結果（逆順）

	precision	Recall	F1-score
rinna RoBERTa (size4)	0.793	**0.885**	**0.836**
rinna RoBERTa (size8)	**0.909**	0.769	0.833
rinna RoBERTa (size16)	0.815	0.846	0.830

表 6：F1-score における上位 3 件の結果（リプライのみ）

	precision	Recall	F1-score
京大 RoBERTa (size4)	0.875	**0.808**	**0.840**
東大 ELECTRA (size32)	**0.910**	0.769	0.833
早稲田 bigbird (size4)	0.909	0.769	0.833

5.3 考察

F1-score で最良値を得た素性に基づく手法と事前学習済みモデルに基づく手法の出力についてエラー分析した．その結果，両者の誤り傾向には重複が見られなかった．誤抽出されたリプライツイートには，リプライ元ツイートを投稿したアスリート本人への誹謗中傷ではなく，対戦相手や他の投稿者への誹謗中傷が多く見られた．今後は，リプライ元ツイートをさらに分析することで，どのような投稿が誹謗中傷を受けやすいのかを検証する．

6. おわりに

本稿では，アスリートに対する誹謗中傷ツイートの検出を目的に，リプライ元ツイートの有効性について検証した．検証の結果，誹謗中傷検出におけるリプライ元ツイートの有効性を確認した．

今後は，どのようなリプライ元ツイートが誹謗中傷を受けやすいのかを検証するとともに，検出性能を向上させる手法を検討する．また，辞書を用いない検出手法についても検討する．また，辞書を用いない検出手法についても検討する．

参考文献

[1] 大友他, “いじめ表現辞書を用いた Twitter 上のネットいじめの自動検出”, DEIM2020 論文集, 7 pages, 2020.
[2] 松本他, “BERT を利用した煽りツイート検出の一手法”, DEIM2021 論文集, 2021.
[3] 有名人 Twitter ランキング, https://www.talentteit.com
[4] 西原他, “電子掲示板からの文脈を考慮した誹謗中傷コメントの抽出”, JSAI2014, 4 pages, 2014.
[5] 西谷他, “アスリートに対する誹謗中傷の分析と検出法の初期検討”, FIT2023 講演論文集，第 2 分冊，pp.217-218，2023.