D-003

アスリートに対する誹謗中傷ツイートの分析と検出法の初期検討
Analysis of Tweets Slandering Athletes and an Initial Consideration of a Detection Method

西谷 千乃与† 安藤 一秋‡
Senchiyo Nishitani Kazuaki Ando

1. はじめに

近年，パソコンやスマートフォンの普及に伴い，SNS が幅広く用いられ，アスリートに対する応援がより簡単になった．しかし，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，多くのアスリートたち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，Twitter を利用するアスリートを対象に，本人のツイートに対するリプライツイートに含まれる誹謗中傷を自動検出する技術の実現を目指す．本稿では，誹謗中傷を含むツイート候補を収集し，ツイートを分析するとともに検出手法について初期検討する．

2. 関連研究

石坂らの研究[1]では，2 チャンネル上で誹謗中傷を含む文を検出する手法を提案している．石坂らの手法では，誹謗中傷に使用される単語であるか否かを判断する悪口度を，SO-PMI（Semantic Orientation Using Pointwise Mutual Information）を用いて単語ごとに算出する．そして，各単語の悪口度を用いて，対象とする文が持つ誹謗中傷性を Support Vector Machin（SVM）で二値分類する．2 ちゃんねるから収集した悪口文 1,400 文と非悪口文 1,400 文を使用した実験により，F 値で 89.97 が得られたと述べている．今後の課題として，2 ちゃんねる以外の文に対しても，同様な結果が得られるかの確認を挙げている．

大友らの研究[2]では，Twitter 上のテキストを対象に，いじめ表現辞書を構築し，この辞書と 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．

本研究では，大友らの研究を参考に，悪口辞書を構築して，誹謗中傷候補の収集に利用し，得られた候補を機械学習により分類する手法について検討する．まずは，悪口辞書を用いた誹謗中傷候補ツイートの収集法について述べる．

† 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

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 アスリートの選定とリプライツイートの収集

アスリートの選定法について述べる．まず，有名人 Twitter ランキング[3]のスポーツ選手・格闘技カテゴリから 433 選手分の Twitter ユーザ ID を任意に取得した．その後，リプライツイートが多い選手であるほど，誹謗中傷を受けやすいと仮定し，433 選手に対する 1 週間分のリプライツイートを収集した．そして，35 件以上リプライツイートが収集できた 115 選手を対象とする．

次に，選手本人のツイートに対するリプライツイートを収集する．2022/6/19~2022/8/30 の期間に収集した結果，92,873 件のリプライツイートが収集できた．

3.2 リプライツイートの収集と辞書拡張による抽出

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[2]といえる．そこで，誹謗中傷に関連する単語辞書（以降，悪口単語辞書）を利用することで，まずは誹謗中傷候補ツイートを抽出する．本稿では，西原ら[4]の研究で用いられていた悪口単語 55 単語を，悪口単語辞書のベースに利用する．

92,873 件のリプライツイートに悪口単語辞書を適用して誹謗中傷候補ツイートを抽出した結果，1,185 件が得られた．しかし，抽出できていない誹謗中傷ツイートが存在し，抽出漏れを確認したため，辞書拡張を検討する．

まず，単語感情極性値対応表[5]を用いて，悪口単語と共起する単語の内，極性値が 0 より小さいネガティブな単語を悪口単語の候補とする．次に，悪口単語を含むツイートと含まないツイートでそれぞれツイートごとの単語の tf-idf 値を求め，単語ごとに合計する．悪口単語を含む文の中には，日常的に使用される単語も多く含まれるため，悪口単語を含む文から，悪口単語を含まない文に対する tf-idf 値の合計値が上位の単語を削除する．その後，単語の共起回数の上位，tf-idf 値の合計の上位をそれぞれ目視し，妥当な単語を辞書に追加する．結果として，図 1 に示す 19 単語を追加し，74 単語からなる悪口単語拡張辞書を構築した．

悪口単語拡張辞書を用いて 92,873 件のリプライツイートから誹謗中傷候補ツイートを抽出した結果，2,216 ツイートが得られたため，追加した単語の有効性を確認できた．

図 1 悪口単語辞書に追加された単語

悪い，障害，タヒ，やめる，なめる，生意気，下手，むかつく，ブタ，ガキ，迷惑，イキる，ゴミ，ごみ，負け犬，くそ，死ぬ，消える，殺す

4. 誹謗中傷候補ツイート分析

本稿では，115 選手の内，スポーツカテゴリによる偏りをなくすため，14 個のスポーツカテゴリ，計 70 選手を対象に設定する．そして，2022/8/9 から 2022/12/15 の期間に収集したリプライツイート 400,090 件を分析対象とする．

4.1 目視による分析

全てのカテゴリの内，目視で確認可能な 3,635 件の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），誹謗中傷ツイートと断言できない（不明）に分類した．その結果，正例が 453 件，負例が 3,132 件，不明が 50 件となり，悪口単語を含んでいる文のうち，12.0%が誹謗中傷ツイートであった．また，不明ツイートも検出対象に加えた方がよいと考えられるため，全体の 16.0%が抽出対象といえる．

4.2 カテゴリ別の分析

カテゴリ別に分析した結果のリプライ数の上位を表 1 に示す．リプライツイート数が最も多いスポーツカテゴリは「格闘家」カテゴリで 130,684 件，誹謗中傷候補ツイート数も「格闘家」カテゴリが最も多く 4,919 件であった．抽出対象となる正例と不明ツイートの合計値の割合も 24.4%と他のスポーツカテゴリよりも高く，リプライツイート数と誹謗中傷数は純粋に比例することが確認できた．

表 1 カテゴリ別の分析結果の一部

カテゴリ名	格闘家	サッカー	プロレス
リプライ数	130,684	115,417	46,958
誹謗中傷候補ツイート（%）	4,919/130,684 (3.76)	2,337/115,417 (2.02)	873/46,958 (1.85)
抽出対象（%）	305/1,250 (24.4)	127/800 (15.8)	18/600 (3.0)

4.3 カテゴリ間の類似性分析

SNS での誹謗中傷はアスリート以外に対しても問題にある．そこで，スポーツカテゴリ間およびその他のカテゴリ（「お笑い」と「政治家」）における誹謗中傷ツイート候補の類似性を確認することで，誹謗中傷のドメイン依存性の有無を確認する．

各カテゴリには特有の単語が存在しているため，悪口単語のみでベクトルを作成し，cos 類似度で比較する．各 5 ユーザに対するリプライツイートを収集した後，悪口単語拡張辞書を用いて誹謗中傷候補ツイートを無作為に各 300 ツイート抽出する．そして，各カテゴリの悪口単語ベクトルの cos 類似度により分析する．

分析結果の一部を表 2 に示す．スポーツカテゴリ間およびスポーツと「お笑い」カテゴリについては，悪口単語の差は確認できなかった．しかし，「政治家」カテゴリとは差があることを確認した．よって，誹謗中傷のドメイン依存性の有無については検証を継続する必要がある．

表 2 各カテゴリの悪口単語ベクトルの類似度

	格闘家	サッカー	プロレス	お笑い	政治家
格闘家	1.000	0.870	0.830	0.893	0.702
サッカー	0.870	1.000	0.933	0.953	0.641
プロレス	0.830	0.933	1.000	0.930	0.545
お笑い	0.893	0.953	0.930	1.000	0.740
政治家	0.702	0.641	0.545	0.740	1.000

5. 検出手法の初期検討

ラベル付けした 3,635 件のリプライツイートを用いて，誹謗中傷ツイートの検出手法について初期検討する．本稿では，関連研究[2]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．特徴量については，bag-of-words，Tf-idf，Word2Vec（W2V）をそれぞれ使用する．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，学習器と素性をそれぞれ組み合わせて，precision，recall，F1-score で評価する．誹謗中傷ツイートの検出タスクでは，曖昧ツイートも幅広く検出したいため，「不明」ツイートを「正例」に加え，503 件の正例データで実験する．なお，負例のデータ数は正例数に揃える．

学習器と素性それぞれ組み合わせた手法の性能を評価した．表 3 には，最良値を得た手法を表 3 に示す．全ての素性を使用した場合，F1-score は 0.804 で最良値になることを確認した．また，敬語表現の数を素性とした実験では，各評価において高い値となり，誹謗中傷の検出に役立つことを確認した．

表 3 各評価指標における最良手法の性能

	precision	recall	F1-score
W2V_mlp 品詞+敬語	**0.798**	0.722	0.758
Bow_svm 文字数+敬語	0.629	**0.929**	0.690
W2V_svm 全ての素性	0.740	0.881	**0.804**

検出できなかった誹謗中傷ツイートのエラー分析の結果，「話にならんのはこちらのセリフやで、先発やめて中継ぎか敗戦処理やれば？」のように，文脈を考慮しなければ判定が難しい事例を確認した．今後は，大規模事前学習済みモデルを利用することで，文脈を考慮できる手法について検討する．また，辞書ベースで候補ツイートを収集する手法では，検出漏れが生じてしまうため，辞書を用いない検出手法についても検討する．

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，誹謗中傷候補ツイートを分析し，検出手法を初期検討した．カテゴリ間の類似性分析から，スポーツとその他のカテゴリでは，誹謗中傷のドメイン差がある可能性を確認した．また，検出手法の初期検討により，Word2Vec とすべての提案素性を用いた SVM の F1-score が 0.804 で最良値を得ることを確認した．今後は，誹謗中傷のドメイン依存性の有無の検証を継続する．また，文脈を考慮する大規模事前学習済みモデルを検討し，検出性能を向上するとともに，辞書を用いない検出手法を検討する．

参考文献

[1] 石坂他, “Web 上の誹謗中傷を表す文の自動検出”, NLP2017 年次大会発表論文集, pp.131-134 (2011).
[2] 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.
[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] 高村他, “スピンモデルによる単語の感情極性抽出”, 情報処理学会論文誌, Vol47, No2, pp.627-637, 2006.