社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IN2020-42 (2021-01)

誹謗中傷による被害を減らすためのツイートにおけるトゲワード検出

伊藤　圭吾†　荒澤　孔明†　服部　峻††

†,†† 室蘭工業大学 ウェブ知能時空間研究室　〒 050–8585 北海道室蘭市水元町 27–1
E-mail: †{17024019,18096001}@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

あらまし　ネット上での誹謗中傷の書き込みは以前から問題視されていたが，現在はさらに問題意識が強まったと感じている．SNS を代表する Twitter は，ユーザの投稿にリプライできるアカウントを制限する機能や，通報を受けたツイートの削除やアカウントの停止などの対策を行っている．しかしながら，このような対策は誹謗中傷を減らす根本的な対策とはならない．この問題を解決するために本稿では，投稿されているツイートが誰かを誹謗中傷しているかどうかを自動的に判定する手法を提案する．基本的な誹謗中傷の判定は，ツイートに含まれる単語と著者らがリストアップした “トゲワード”（罵詈雑言やネガティブな言葉）とのパターンマッチにより行う．さらに，ポジティブ・ネガティブ（PN）判定を用いてツイートのネガティブ度を算出したり，係り受け解析を用いてツイート中の単語（特にトゲワード）の対象を認識させたりすることで，誹謗中傷しているツイートを精度良く分類するシステムを考案する．
キーワード　誹謗中傷判定，係り受け解析，ポジティブ・ネガティブ判定，ツイート，SNS

Stinging-Word Detection in Tweets for Reducing Defamation Damage

Keigo ITOH†, Komei ARASAWA†, and Shun HATTORI††

†,†† Web Intelligence Time-Space (WITS) Laboratory, Muroran Institute of Technology
27–1 Mizumoto-cho, Muroran, Hokkaido 050–8585, Japan
E-mail: †{17024019,18096001}@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

Abstract　Posting slanderous defamation on the Internet has been regarded as a social problem, and now it seems that the social awareness of the problem has become even stronger. Twitter, which represents SNS, has a function to limit the accounts that can reply to a user's posts, and takes measures such as deleting the tweets and suspending the accounts that are reported as NG. However, such measures are not radical ones to reduce slanderous defamation and reduce its damage. To solve this problem, this paper proposes a method to automatically judge whether or not a posted tweet is defaming someone or someone's doing/something. The basic judgment of defamation is made by pattern-matching words in a posted tweet with “Stinging Words” (such as abuse words and negative words) listed by the authors. Furthermore, this paper aims to develop a system that classifies defamation tweets as precisely and exhaustively as possible, by calculating the negative degree of tweets using Positive/Negative (PN) judgment and also by recognizing the target of a word (especially, stinging word) in a tweet using dependency analysis.
Key words　Defamation Judgment, Dependency Analysis, Positive/Negative Judgment, Tweets, SNS

1. はじめに

SNS が普及し誰もが自身の考えや出来事を発言できるようになり，人とのコミュニケーションが容易に取れるようになった．Twitter や Facebook，Instagram は利用者数の観点から SNS を代表するサービスといっても良いであろう．これらのサービスが始まった頃，日本では，趣味で始める一般人が利用者の大多数を占めていたと思われる．現在では SNS が社会に浸透したことで，企業や有名人までもが実名で参入しており，ネット上で行われるコミュニケーションの領域はもはや現実と変わらないように思われる．以前なら企業と顧客，有名人とファンの間でのコミュニケーションの取り方はメールや手紙などで行われていた．メッセージの中には応援する内容であったり批判的な意見もあるであろうが，一方でメッセージを受け取った側を傷つける悪質な内容のものも存在していたが，それらは事務所の管理によって本人には見えないように隠されていた．しかしながら，それらの悪質的なメッセージは現在，SNS を通じ本人が見られる場所に存在し，また，誰もが見られる情報となり公

開され，罵詈雑言や誹謗中傷と位置づけられ社会的に問題視されるようになった．ネット上でも同様にユーザの通報により誹謗中傷している文章の削除であったり，投稿しているアカウントの凍結などの対策が人間の手によって行われているが，人間の手作業による限界や既に受け取り手に見られている可能性があり，これらの対策では不十分である．この問題を解決するためには罵詈雑言や誹謗中傷を自動的に判別し，未然に誹謗中傷している文章の投稿を防いだり，受け取る側の設定で未然に非表示にしたりする機能が必要であると考えている．

本研究では，罵詈雑言や誹謗中傷している文章を “棘のある言葉” という意味で “トゲあり文章” と定義しており，その文章に含まれる誹謗中傷を形成する単語を “トゲワード” と定義している．本稿では Twitter に投稿されたツイートの中に含まれるトゲありツイートを，トゲワードリストとのパターンマッチによって検出する手法を提案する．この単純な手法にさらにポジティブ・ネガティブ（PN）判定や係り受け解析を用いることで，ツイートのネガティブ度やツイート中の各単語の対象を機械に認識させることによって，誹謗中傷を読んだ時の人間の感覚に近づけていき提案手法の精度向上を目指す．

2. 関 連 研 究

Web 上に投稿されている文章から，誹謗中傷を検出する研究はいくつか存在する．松葉ら [1] は，学校の非公式サイトに投稿されている文章における有害情報を，SVM による二値分類で検出している．有害情報には，個人名や個人情報の他に「うざい」や「しね」などの誹謗中傷も含まれている．石坂ら [2] は，誹謗中傷を悪口として扱い，皮肉のような文脈や他の情報を必要とする中傷は対象としないと定義している．悪口文の分類には，松葉らと同様に SVM を用いて行っている．また，係り受け解析器（CaboCha）を用いて，悪口単語に否定語である「ない」が用いられていれば悪口単語から除外する手法も取り入れている．池田ら [3] は，キーワードによるパターンマッチでの違法・有害サイトの検出は，文章中でのキーワードの使われ方を考慮していないため高精度にはならないと主張し，係り受け解析を用いて精度向上を目指した．キーワードは人手によってラベル付けされた有害・無害文書を学習させ，双方で抽出される単語の出現率から，有害と無害な単語を自動抽出している．

本研究との違いは，検出しなければならない誹謗中傷の定義が難しいところである．同じような内容の誹謗中傷でも，傷つく人と傷つかない人に分かれる場合があり，受け取り手によってダメージが異なるためである．また，研究で扱うデータの特徴についても違いがある．特に，池田らの研究ではウェブサイトの文章全体を扱っており，字数制限のある SNS のショートメッセージを扱う本研究においては，文章の長さの違いや，誤字脱字，造語による解析の失敗が多くなることが予想される．そこで著者らは，係り受け関係のある単語の解析に，キーワード（本研究ではトゲワード）の対象語が抜けている場合を考慮し，ツイート中の主題語を求めることで精度を向上させる．また，SNS というショートメッセージにおいては，「楽しい」や「酷い」などの率直な感想や意見を手短に書いているという仮説を立て，ポジティブな単語とネガティブな単語が浮き彫りになると予想し，ポジティブ・ネガティブ（PN）判定を用いた分類精度の向上についても検討している．

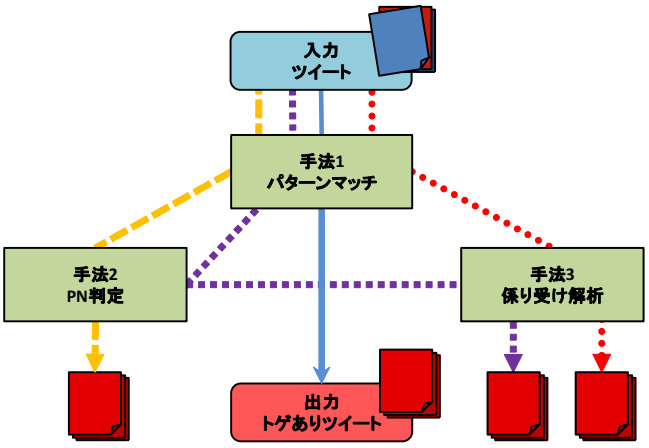

図 1　提案手法の概観

3. 提 案 手 法

著者らが作成した手製のトゲワードリストに基づくパターンマッチによって，Twitter に投稿されているツイートを誹謗中傷している「トゲありツイート」と誹謗中傷していない「トゲなしツイート」に分類する手法を提案する．提案手法の概観を図 1 に示す．また，大別して，以下の 3 つの手法からなる．

手法 1　トゲワードが見つかった場合，そのツイートを “トゲありツイート”（誹謗中傷）と判定する．

手法 2　トゲワードが見つかった場合，さらに，そのツイートの PN 判定を行い，ネガティブとなったツイートをトゲありツイートと判定する．

手法 3　トゲワードが見つかった場合，さらに，そのツイートに対して係り受け解析を行い，トゲワードの係り受け先が「固有名詞」または「一人称を除いた人称代名詞」であればトゲありツイートと判定する．

3.1　手製トゲワードリストについて

トゲワードリストに追加する単語は，誹謗中傷しているツイートに含まれる確率が高い単語である．以下に例を示す．

バカ，馬鹿，きっしょ，阿保，タヒね，死〇...etc.

実験で扱うトゲワードの数は全 138 単語である．未知語による対策は手動で追加を行っている．また「頭が悪い」，「豚みたいな顔」などの誹謗中傷となる表現については対応していない．これらはトゲワードリストに追加しても良いと思われるが，単純なパターンマッチでは，上記のトゲワードの例よりも表現のパターンが多すぎるため本提案手法では扱わないものとする．

3.2　3 種類の手法の詳細

3 種類の手法の詳細を表 1 の例文を用いて説明する．こちらの 3 件のツイートに共通するトゲワードは “馬鹿” であるが，人

表 1　トゲワードを含むツイートの例

I	お前 “馬鹿” だろ。考え方がひどすぎるｗ
II	お前ら “馬鹿” 騒ぎしすぎ ｗ めっちゃ楽しかったけど！
III	A さんが言ってたことは難しいな。俺が “馬鹿” なだけか。

間がこれらのツイートを見た時 “馬鹿” の使われ方を意識して誹謗中傷であるかを判断する．I の場合は “馬鹿” は「お前」という相手の人間に対して使われているので誹謗中傷である．II の場合は “馬鹿” は「騒ぐ」に使われていて人間ではないから誹謗中傷であるとは思わない．また「楽しかった」などのポジティブなイメージを連想させる単語を用いているため，ツイート全体に棘を感じさせない．III の場合は自分に対しての発言なので誹謗中傷ではないと判断できる．少なくとも著者の感覚では表 1 の例においては I のみが誹謗中傷であると判断している．このような人間の感覚と同様の，誹謗中傷の判定処理を提案手法に行わせる．以下に 3 種類の手法がどのようにツイートを解析するかを詳述していく．

3.2.1　手法 1 ：パターンマッチ

手法 1 では，トゲワードとツイート中の単語とのパターンマッチによって判定が行われる．表 1 の例では，トゲワードリストに含まれる “馬鹿” とのパターンマッチによる判定を行っているため，I，II，III すべてのツイートをトゲありツイートと判定する．この手法には人間の感覚を疑似化する作業は行わせていないので，このように単純な結果となる．

3.2.2　手法 2 : PN 判定

手法 2 の PN 判定では東北大学の乾・鈴木研究室の「日本語評価極性辞書」[4] を用いている．この辞書にはポジティブな単語とネガティブな単語が登録されている．こちらを利用してツイート中で使われている単語をスコア化するプログラムを作成し，ツイートのスコア合計をネガティブ度として，ある基準値（0 未満）を満たしていれば，トゲありツイートであると判定する仕組みである．単語のスコアリング方法は，ポジティブな単語であれば +1，ネガティブな単語であれば −1 という単純なものである．手法 2 に表 1 の II のツイートを解析させると，「馬鹿」が −1，「楽しい」が 1 となり，全体の文章のスコアは 0 となる．手法 2 では 0 未満のマイナスのスコアを出したツイートをトゲありツイートと判定している．よって手法 2 を用いることで，手法 1 により判定されたトゲありツイートの候補から II のツイートを除外することができる．

3.2.3　手法 3 : 係り受け解析

トゲワードの対象が何かに依って，あるトゲワードを含むツイートにおいて，そのツイートのトゲのあり・なしが変わる場合がある．そこで手法 3 では，トゲワード（例えば表 1 における “馬鹿”）の係り受け関係にある単語を，係り受け解析器の CaboCha を用いて解析する．この手法では，あるツイート中のトゲワードと係り受け関係にある単語が，固有名詞，または一人称を除いた人称代名詞（自虐を除くため）であれば，トゲありツイートであると判定する．

手法 3 でトゲありツイートと判定される例文を，CaboCha に係り受け解析させた結果を図 2 に示す．この結果では 1 つの

例文:Kさん真面目に馬鹿だなって思った

```
Kさん-D
真面目に-D
 馬鹿だ-D
  なって-D
   思った-D
```

図 2　CaboCha によるトゲありツイートの解析例

例文:Kさんの動画見てないけどなんだこれ。きっしょいわ。

```
Kさんの-D
   動画-D
見てないけど-D
 なんだこれ。---D
     きっしょ-D
      いわ。
```

図 3　トゲワードと固有名詞・代名詞に係り受けの関係がない例

文章が 5 つの文節に分けられ，文節ごとの係り受け関係が出力されている．分かれた文節の末尾に「-D」という文字列があるが，これが係り受け関係のある文節の末尾の真上に来ている．つまり図 2 の例では「K さん」は「真面目に」に係っていることを意味している．

具体的な処理としては，まず単語レベルでトゲワードリストとのパターンマッチを行いトゲワードが含まれる文節を発見すると，次に係り受け関係のある文節の単語を解析しに行く．係り受け関係のある文節の中に，固有名詞または一人称を除いた人称代名詞を見つけると，トゲワードが「人間」に対して使われていると判断し，トゲありツイートと判定を下す．図 2 の例では，3 つ目の文節でトゲワードの “馬鹿” を見つけ，係り受け関係のある文節を 2 ホップ辿り「K さん」という固有名詞を発見し，トゲありツイートと判定している．ここで，何 k ホップまで辿るのが最良か議論する必要があるので，5 章の考察で検証する．表 1 の例に戻ると，手法 3 を用いれば，I では “馬鹿” と係り受け関係にあるのは「お前」となっており，一人称を除いた人称代名詞であるためトゲありツイートと判定する．一方で III では “馬鹿” と係り受け関係にあるのは，「俺」という 1 人称であるため，トゲありツイートの候補から除外される．

上述の手法 3 では，トゲワードが含まれる文節から係り受け関係（リンク）にある文節を辿って，固有名詞または一人称を除く人称代名詞が見つかるまで k ホップ辿るといった方式を採っているが，一方で，図 3 のようなツイートの例では発見できない．トゲワードである “きっしょ” に係り受け関係があるのは 6 つ目の文節である「いわ。」だけであり，これ以上辿ることは不可能である．従って，手法 3 では，トゲワードである “きっしょ” の対象である「K さん」という固有名詞が見つからない．この例では文章が「。」によって区切られているため，“きっしょ” の含まれている文章には「K さん」という主語が抜けている．ツイートのように，文字数制限のあるショートメッセージでは，主語や対象が省略されやすい．しかし本来であれ

ば「K さん」を “きっしょ” と言っているツイートであるため見過ごすことはできない．そこで，手法 3 を改良した以下の手法 3b についても提案する．また，上記の手法 3 を以降，手法 3a と呼ぶことにする．

手法 3b では，手法 3a を試みた結果，トゲワードからの直接的な係り受け関係（リンク）を辿っても，真の対象語に辿り着けなかった場合には，トゲワードと真の対象語との何らかの係り受け関係の描写が省略されており，トゲワードよりも前の文章中における「主題語（句）」を省略されている真の対象語と仮定する．そして，手法 3a と同様に，その「主題語（句）」に固有名詞または一人称を除く人称代名詞が含まれている場合にも，トゲワードが「人間」に対して使われていると判断し，トゲありツイートと判定を下す．但し，「主題語（句）」は，トゲワードよりも前の文章中の名詞を含む文節の中で，最も係り受け関係（リンク）数が多いものを核に，1 ホップまで名詞を含む文節まで複合させたものである．図 3 の例では，“きっしょ” よりも前の文章中には名詞を含む文節として「K さんの」と「動画」があり，それぞれ係り受け関係（リンク）数を計算すると，「K さんの」は 1，「動画」は 2 となるため，「動画」が主題語（句）の核となる．さらに，1 ホップまで辿って名詞を含む文節である「K さんの」までを複合し，この例では「K さんの動画」が主題語（句）となり，「K さん」という固有名詞が含まれているため，トゲありツイートと判定している．

4. 評 価 実 験

提案手法を用いて誹謗中傷しているツイートを検出する実験を行い，トゲありツイートか否かを判定するための手法 1，2，3a（パラメータとしてホップ数 k を持つ），3b の計 4 種類の組み合わせで比較実験を行う．

4.1 実 験 手 順

実験は以下の手順により行った．

Step 1. Twitter に投稿されているツイートを 500 件抽出した．

Step 2. 抽出したツイートをもとに, トゲありツイートとトゲなしツイートにラベル付けを行った．

Step 3. 提案手法にツイートを入力し, トゲありツイートとトゲなしツイートに分けて自動でラベル付けを行った．

Step 4. 提案手法が誹謗中傷しているツイートを抽出する適合率と再現率，F 値を算出する．

実験手順 1 で抽出された実験サンプルとなるツイートは，あるネットタレントに関して投稿されたツイートである．ツイートの例を表 2 に示す．また，実験手順 2 で行われた人手によるラベル付けの結果は，トゲありツイートが 157 件，トゲなしツイートが 343 件となった．

4.2 誹謗中傷しているツイートの選別について

誹謗中傷を定義することは難しい．誹謗中傷にはいくつかの種類があると考えている．例えば，「お前のこときらいだわ」のように相手に対して直接的に誹謗中傷を浴びせるか，それとも

表 2 実験サンプルツイート

・この返しは笑える、K さん頭悪いのかな？w
・K さんそんな事言うから嫌われてるんだよ

「周りから嫌われてるじゃん」のような周りの人の言葉を間接的に相手に浴びせるかの違いや，誹謗中傷の対象が相手自身であるか，相手の所有物や所属している組織であるかつ，あるいは社会であるかの違いなどがある．これらは受け取り手によって受けるダメージが異なるため，主観で誹謗中傷を定めるのは容易ではないのである．実験手順 2 の誹謗中傷ツイートを手作業で分類する基準は以下のようなルールに基づいて行われた．

1. トゲワードの対象が人やその人の所有物及び，所属している組織であること
2. 罵詈雑言やネガティブな単語を用いて対象を直接的あるいは間接的に陥れたり貶しているもの
3. 受け取り手の行動に対する批判であると思われるツイートは誹謗中傷としない

4.3 未知語による解析精度の低下についての対応

手法 2 の PN 判定，手法 3a，3b の係り受け解析には，MeCab の解析による文章内の単語の分割や，品詞情報を用いている．MeCab の辞書に登録されていない単語は未知語として出力される．評価実験で扱う 500 件のツイートの中にも，YouTuber の活動名や「w」，名詞目的で使われていない「草」などの未知語と判定される単語がいくつか存在するが，1 つの単語として出力されなかったり，名詞や固有名詞として出力されてしまい，正しく解析されない場合がある．その結果，解析精度が低下，あるいは偶然に正解の出力を得てしまう可能性がある．この問題を解決するために，著者らが，提案手法による 500 件のツイートの解析結果を確認し，未知語に対応したユーザ辞書を作成し，MeCab の解析に用いられる辞書を拡張する．

4.4 実 験 結 果

提案手法の実験結果を表 3 に示す．手法 1 では，ツイート内のトゲワードの使われ方を考慮していないため，再現率は高くなるが，一方で適合率は低くなることが予想される．そこで，手法 2，3a，3b では，手法 1 でトゲありツイートと誤判定されたトゲなしツイートを，ツイートのネガティブ度や，トゲワードの使われ方を考慮することによって，トゲあり判定されたツイートから除外する．つまり，これらの手法は，手法 1 の適合率を向上させつつ，再現率の低下をできるだけ抑えることを目標としている．しかし，表 3 の手法 1+2 の結果を見ると，再現率と適合率を共に下げており，手法 2 が適合率を上げるというコンセプトとして機能していないことが考えられる．手法 1+3a では，適合率が上がってはいるが，共に再現率も下がっている．適合率の上がり幅は乏しいものの，予想通りの働きをしていることがわかる．

手法 3b は，ツイートのような文字数制限のあるショートメッセージの解析において，3.2.3 節の説明を率直に言い換えると，手法 3a による再現率の低下を軽減するものである．ここで手

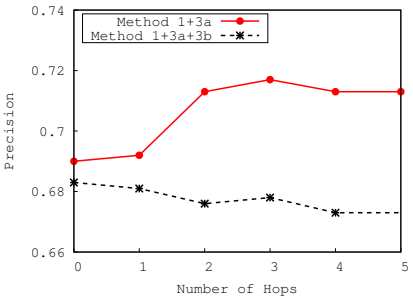

図 4　手法 1+3 のホップ数と適合率

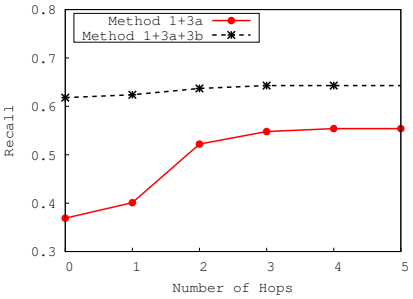

図 5　手法 1+3 のホップ数と再現率

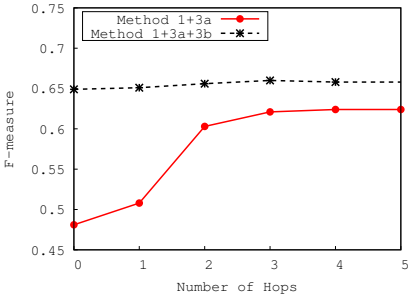

図 6　手法 1+3 のホップ数と F 値

法 3b がどれだけ精度向上に寄与しているかを見るために，手法 1+3a と手法 1+3a+3b について，ホップ数が 0〜5 の時の適合率，再現率，F 値を比較したグラフを図 4〜6 にそれぞれ示す．

図 4 では，手法 1+3a の方が適合率が高くなっていることがわかる．ホップ数の推移を見ると，手法 1+3a の方は，トゲワードと主語や対象の係り受け関係（リンク）を 3 ホップ目まで辿ることが一番良い精度が出ていることが確認できる．一方で手法 1+3a に 3b を追加した手法では，ホップ数による適合率の差はほぼ見られない．

図 5 では，手法 3b を追加した手法の方が，再現率が高くなっていることがわかる．どちらの手法もホップ数を増やすほど，トゲワードと係り受け関係（リンク）を持つ文節において固有名詞や人称代名詞を見つける確率が高くなるため，再現率が向上していることが確認できる．また，手法 1+3a の再現率は，リンクを 1 ホップ目まで辿った時に比べて，2 ホップ目まで辿った時の方が 0.1 程度高くなっていることがわかる．

図 6 では，全てのホップ数 0〜5 において，手法 3b を追加した手法の方が F 値が高くなっていることがわかる．以上より，手法 3b を追加した手法の方が，精度を向上させる働きがあるということがわかる．

表 3　提案手法の実験結果

手法					評価		
1	2	3a	#hop	3b	P	R	F
✓			–		**0.657**	**0.758**	**0.704**
✓	✓		–		0.648	0.503	0.566
✓		✓	0		0.690	0.369	0.481
✓		✓	1		0.692	0.401	0.508
✓		✓	2		0.713	0.522	0.603
✓		✓	3		**0.717**	0.548	0.621
✓		✓	4		0.713	0.554	0.624
✓		✓	5		0.713	0.554	0.624
✓		✓	0	✓	0.683	0.618	0.649
✓		✓	1	✓	0.681	0.624	0.651
✓		✓	2	✓	0.676	0.637	0.656
✓		✓	3	✓	0.678	**0.643**	**0.660**
✓		✓	4	✓	0.673	0.643	0.658
✓		✓	5	✓	0.673	0.643	0.658
✓	✓	✓	0		0.661	0.248	0.361
✓	✓	✓	1		0.656	0.268	0.380
✓	✓	✓	2		0.671	0.350	0.460
✓	✓	✓	3		0.674	0.369	0.477
✓	✓	✓	4		0.678	0.376	0.484
✓	✓	✓	5		0.678	0.376	0.484
✓	✓	✓	0	✓	0.667	0.433	0.525
✓	✓	✓	1	✓	0.663	0.439	0.529
✓	✓	✓	2	✓	0.651	0.439	0.525
✓	✓	✓	3	✓	0.651	0.439	0.525
✓	✓	✓	4	✓	0.651	0.439	0.525
✓	✓	✓	5	✓	0.651	0.439	0.525

5. 考　　察

本章では，各手法がどのようなトゲありツイートを検出したのか確認し，手法の改善策を考える．表 3 から全体的な結果を確認すると，手法 1 の精度向上のために手法 2，3a，3b を追加したにもかかわらず，手法 1 だけの解析精度の方が高くなっていることがわかる．各手法ごとの精度から，手法 1+2 の解析結果は期待していた精度とはならず，手法 1 だけの手法に比べて，適合率と再現率の両方を下げてしまっている．一方で，手法 3a と 3b は，F 値を上げられなかったものの，適合率は手法 1 だけの解析結果を超えることができた．手法 1 だけの精度を除くと，手法 1+3a+3b の，3 ホップ目まで辿った時の F 値が 0.66 となっており，組み合わせた手法の中では一番高くなることがわかった．

5.1　手法 1 の精度について

手法 1 の解析結果では，ツイート 500 件のうち 181 件をトゲあり判定している．このうち，正しくトゲあり判定できたトゲありツイートは 119 件である．実験データの中には，157 件の誹謗中傷が含まれているため，38 件の誹謗中傷は検出できなかったことがわかる．この 38 件に含まれるトゲワードは「人権ない」，「小判みたいな顔」のように，2 つの単語が組み合わさって初めてトゲワードとなるため，本実験で扱うトゲワードリストには含めていない．このようなトゲワードを含むツイー

表 4　手法 2 で検出できなかったトゲありツイート

I	K さん/さすがに（+1）/キモいわ
II	K さん/クソ/ウケる（+1）/w久々に/不快 (-1) だったわ

トをトゲあり判定するためには，単純なパターンマッチでは何通りもの表現の登録が必要となってくるため，単語の組み合わせにより生じるトゲワードを判定するのに特化した，“トゲワードペアリスト” の作成を検討する．

5.2　手法 2 の精度について

手法 2 が期待通りの精度を出せなかったことについて考察する．手法 1 がトゲあり検出した 181 件のうち，手法 2 が正確に排除したトゲなしツイートは 19 件である．一方，誤って排除したトゲありツイートは 40 件である．誤って排除したツイートの例を表 4 に示す．I の例では，日本語評価極性辞書に “キモい” が登録されていなかったためである．II の例では，「ウケる」と「不快」で PN スコアが 0 になってしまい，トゲあり検出されなかったことが原因である．「ウケる」という単語は本来は面白い時などに使われることが多いため，ポジティブになるのはわかるが，ツイート内では，皮肉のような使い方をしており，人間からするとポジティブな内容には感じられない．実験結果をすべて確認すると，ツイートに含まれる単語数の割に，スコアの登録されている単語が少なかったり，MeCab の解析ミスにより，日本語評価極性辞書に含まれる単語でも正しくスコアが付与されないケースが多く存在した．

5.3　手法 3a のホップ数について

ホップ数とは，トゲワードが検出された文節から，係り受け関係（リンク）のある文節をいくつまで辿るかを定める文節の数である．著者らは，ホップ数 k を増やすほど再現率は上がるが，一方で適合率が下がると予想していた．トゲワードが人や所有物，組織に対して使われている場合は，トゲワードを含む文節と対象を含む文節が近くにあると考えられ，文節が遠くにある程，そのトゲワードは対象に向けられたものではない可能性が高まるからである．実際の実験結果から，この予想が正しいのかを確かめる．

図 5 より，手法 1+3a において，再現率はホップ数（辿る文節）が多いほど上がっている．ホップ数毎に，トゲありツイートと判定された例を表 5 に示す．実際にトゲあり判定された，3 ホップ以上の例文は表に収まらないので割愛する．まずは，0 ホップを加えた理由を説明する．正しい挙動ならば，「むり」をトゲワード検出した後に，リンクを 2 つ辿って「K さん」という固有名詞を検出し，トゲありツイートと判定される．しかし，文章が連続して平仮名で綴られていたり，句読点が抜けていたりすると，CaboCha による係り受け解析が失敗し文節が正しく分けられない場合がある．ツイートにはこのような例が度々見受けられるため，このような例にも対応できるように，0 ホップの場合を設けた．1 ホップと 2 ホップの例文は，CaboCha による解析と提案手法が正しく動いた時の例である．

一方で，図 4 より，手法 1+3a において，適合率は 3 ホップの時がピークとなっている．これは，4 ホップの時に偽陽性（FP）が増えていることが原因であると考えられる．しかし，本実験における偽陽性のほとんどは，「K さんの虚言癖がやばい」や「K さんの喧嘩が怖い」などの，“行動” に対しての批判であるため，本稿では文節の離れ具合による適合率の低下は確認できなかった．

表 5　ホップ数毎のトゲありツイート

#hop	ツイート
0	K さんやっぱむりかも
1	てか/ K さん / 怖えよ
2	A さんと / K さんの / コラボ動画 /マジキモすぎ。

5.4　手法 3b の主題語の選択精度について

図 5 より，手法 3b を用いることで，手法 3a で検出できなかったトゲありツイートを検出し，再現率を向上させることに成功したことがわかる．実験では，手法 1+3a の 3 ホップ目まで辿る時の精度と比べて，3b を加えた手法では，トゲありツイートを 15 件多く検出できた．一方で，トゲなしツイートを誤って検出した例が 14 件増加した．ここで，計 29 件のデータの解析結果を元に主題語の選択が正しいかどうか調べた．主題語の解析結果を確認したところ，誤った選択を行っていたツイートは 9 件で，残りの 20 件はそのツイートの主題語と思える単語を正しく選択していた．主題語が誤って選択された 9 件のツイートの特徴は，固有名詞が複数存在するかつ助詞などが抜けているツイートであり，誹謗中傷の対象ではない主題語を選択してしまっている場合が多かった．よって主題語の選択精度は良好であったと思われる．

6.　まとめと今後の研究課題

本稿では，Twitter に投稿されている誹謗中傷しているツイートを，正しく検出するための手法を考案した．単純なパターンマッチによる手法では，500 件という少ないデータでも，「比喩」や「皮肉」による本来悪口とはならない単語で構成されているトゲワードが多く存在した．今後，パターンマッチに用いるトゲワードリストには，トゲワードを自動追加する機能であったり，単語の組み合わせによるトゲワードを検出するための仕組みを検討しなければならない．PN 判定は，スコアの付け方や，辞書の運用方法を再検討し，より正確に分類できるよう改善していく．係り受け解析は，精度的には悪くなかったものの，CaboCha による解析において，未知語が含まれていたり，助詞などが抜けていることによって，文節の分け方に間違いが生じている例があった．未知語を登録して解析に扱う単語を増やしたり，足りない助詞などを補完する仕組みを検討することで，トゲワード検出の精度向上を目指す．

文　献

[1] 松葉達郎, 桝井文人, 河合敦夫, 井須尚紀, “学校非公式サイトのおける有害情報検出,” 情報処理学会研究報告, NL192-15 (2009).

[2] 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, no.E1-6, pp.131–134 (2011).

[3] 池田和史, 柳原正, 松本一則, 滝嶋康弘, “係り受け関係に基づく違法・有害情報の高精度検出方式の提案,” 第 2 回データ工学と情報マネジメントに関するフォーラム, C9-5 (2010).

[4] 日本語評価極性辞書-東北大学 乾・鈴木研究室, http://www.cl.ecei.tohoku.ac.jp/index.php? (2020).