令 和 2 年 度　卒 業 研 究 論 文

題目　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に関する研究

指導教員　　服部 峻

提出者　　室蘭工業大学 情報電子工学系学科

氏　名　　伊藤　圭吾

学籍番号　　17024019

提出年月日　　令和 3 年 2 月 12 日

目次

図目次

表目次

第 1 章

序論

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内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ることで提案手法の精度向上を目指す．

第 2 章

関連研究及び既存のシステム

本章では，本稿で行う実験の内容や扱う技術と関連のある研究の紹介と，実際に誹謗中傷を減らすために行われている対策（システム）を紹介する．

2.1　有害情報の検出に関する研究

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．SVM で分類する時に用いる素性の特徴量には，IDF 値を用いることで分類精度が上がる可能性があるという結果を出している．

石坂ら [2] は，単語の悪口度を算出し自動で悪口単語を抽出する実験と，悪口文と非悪口文に分類する実験を行っている．単語の悪口度の算出には Wang and Araki [3] が提案した SO-PMI (Semantic Orientation Using Pointwise Mutual Information) を使用しており，その結果は詳細には記載されていないが，悪口度の高い上位 5 件の単語は的確に悪口を抽出している．悪口文の分類には抽出した悪口単語を素性とした SVM を用いている．また否定語に対応するため CaboCha を用いた係り受け解析を行い分類精度の向上を目指した．評価実験の結果は F 値が最大で 0.9 という高い精度を出している．しかし造語の対応が難しく，形態素解析を行った時に悪口単語が分けられてしまい分類に失敗してしまう可能性があることを指摘している．

池田ら [4] 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．係り受け解析による分類精度を上げるために，有害情報を持つ単語と係り受け関係のある単語に概念辞書を用い，使われている単語の認識に柔軟性を持たせた．結果，分類精度は F 値 0.75 という結果を出している．

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，受け取り手に依っ

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，ツイート内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．

2.2　誹謗中傷を減らすためのシステム

SNS や掲示板における，インターネット上の誹謗中傷をなくすために行われている対策を 2 つ紹介する．1 つ目は，ユーザの通報による投稿の削除である．SNS は自由に発言できるという特徴があるため，卑猥な動画や画像のアップロード，誹謗中傷などの，あるユーザにとっては危険性のある内容が投稿される可能性がある．それらの投稿を削除するために，ユーザが通報できる窓口を設け，プロバイダは通報を受けた投稿を確認し削除や閲覧制限をかけるなどの対策を行っている．2 つ目は，返信できるアカウントを制限することで嫌がらせ行為を行う人との連絡を機械的に絶つ方法である．この対策は Twitter で行われており，実際に自分のアカウントでは返信ができないというツイートを見かけたことがある．一方で，この機能を利用していないユーザの返信欄では，他のユーザによる誹謗中傷を見かけることもある．

これら 2 つの対策は，実際に誹謗中傷している文章を減らす働きをしているが，全ての誹謗中傷の削除を手動で行うことは困難であったり，返信欄以外では自由に投稿できたり，投稿された後の削除であれば既に誹謗中傷による被害を受けている可能性も考えられ，本質的には不十分な対策であると認識している．

本稿では，誹謗中傷している文章をテキスト内の単語から得られる情報を基に，誹謗中傷しているツイートを自動で検出する実験を行う．検出する精度が向上し正しく誹謗中傷を検出できる状態になれば，誹謗中傷を含む投稿を未然に防いだり，閲覧者側から制限できる機能として誹謗中傷を含む投稿のみを非表示にするといった対策が可能になると考えている．

第 3 章

誹謗中傷の抽出におけるツイートの有用性

本章では，誹謗中傷を抽出する実験のデータとして，数ある SNS や掲示板から Twitter を選んだ理由を詳述する．

3.1　TwitterAPI を用いたツイートの抽出について

ツイートの抽出には TwitterAPI を用いた自動抽出を行った．本実験で利用している API には制限があり，自動で抽出できるツイートにも 3 つの条件がある．1 つ目は，キーワードを入力することでヒットするツイートであること．2 つ目は，アカウント名を入力することで得られる，そのアカウントが投稿したツイートであること．3 つ目は，7 日前までのツイートであること．この 3 つの条件の基で得られるツイートを本実験では扱うこととなる．しかしながら，誹謗中傷を含むツイートを手作業で多く用意するのは困難であるため，このような自動抽出が可能な API が存在し，時事問題や個人に対する意見が自由に飛び交う側面を持つ Twitter を利用するのは有効であると考えている．

3.2　ツイートの特徴と解析精度について

SNS における誹謗中傷を検出するために，ツイートの特徴と得られる情報について考える．まず押さえるべき特徴として，ツイートには字数制限がある．現在一つの投稿における文字数は最大 140 字とされている．ツイートのようなショートメッセージでは，主語や助詞などが抜けていたり，流行している造語が用いられたりすることがある．本稿では，これらの特徴を持つ文章に対し形態素解析や係り受け解析を用いるため，誤字や脱字，未知語などによる解析精度の低下が予想される．

一方で，本稿では活用していないが，これらの解析精度を下げてしまう特徴に対し，解析精度の低下を抑える働きをする特徴もいくつか存在する．例えば，ツイート内の情報から誰に対しての返信であるかどうかが分かるようになっている．この情報からは，ツイート内に主語が

含まれていなくても罵詈雑言が用いられている場合に，返信先の相手に対して誹謗中傷しているということが機械的に判定できるといった活用方法が挙げられる．さらに，あるツイートが誹謗中傷であるかどうかを機械的に判断する時に扱う情報は，そのツイートのテキストだけでなく「そのツイートへのリプライ」の内容も活用できると検討している．また「いいね」や「リツイート」といった数値的にそのツイートがどのくらい注目を浴びているかが見て分かる情報も利用できそうな特徴であるが，こちらは「BAD ボタン」がないため，「いいね」が押されているからといって良い内容の投稿であるとは断言できない．本稿ではこれらの解析精度を抑える働きについては触れていないため，今後の課題として検討することとする．

第 4 章

提案手法

著者が作成した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 4.1 に示す．また，大別して，以下の 3 つの手法からなる．

手法 1 トゲワードを検出した場合，そのツイートを誹謗中傷と判定する．
手法 2 トゲワードを検出した場合，さらに，そのツイートのポジティブ・ネガティブ（PN）判定を行い，ネガティブとなったツイートを誹謗中傷と判定する．
手法 3 トゲワードを検出した場合，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」であれば誹謗中傷と判定する．

4.1 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．以下にデータから一部抜粋したトゲワードを示す．詳細は付録 A に示す．

実験で扱う手製のトゲワードの単語数は全 138 単語である．「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，以下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．

バカ，馬鹿，きもい，嫌い，嫌われる，死ね，あたおか，無理，不快，怖い，臭い

4.2 WordNet による拡張トゲワードリストについて

トゲワードリストに登録されている単語が少なければ，それだけ誹謗中傷を検出することができない可能性があり，再現率を向上させるため，ある程度単語を増やすべきであると考えて

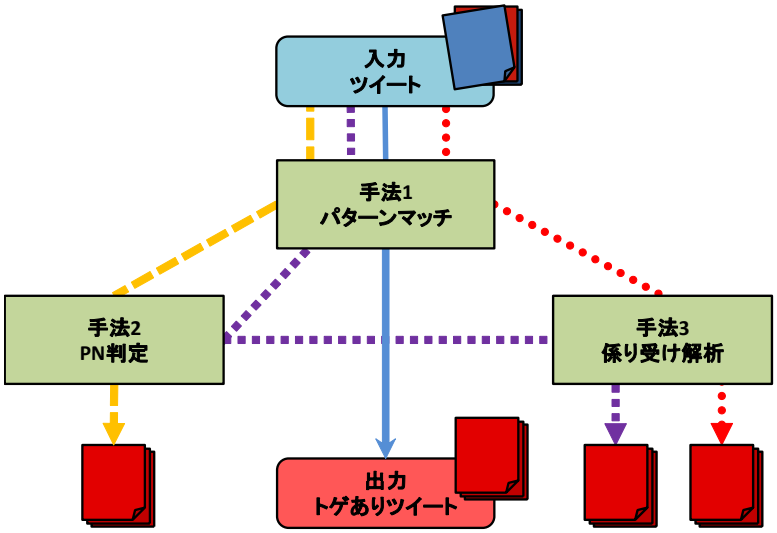

図 4.1　提案手法の概観

表 4.1　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 最高に楽しかったけど！
III	A さんが言ってたことは難しいな。俺が “馬鹿” なだけか。

いる．手製のトゲワードリストに登録されている単語は 138 語であるが，意味で分類すると 40 種類となる．手製トゲワードリストは，トゲワードの基本形（一部例外あり）となる 40 単語に対し，著者が思いつく範囲で誹謗中傷となりそうな類義語を全 98 単語追加した全 138 単語で構成される辞書である．しかし手作業で類義語を追加するのは困難であるため，類義語を自動で出力する WordNet[5] を用いてトゲワードの自動追加を試みる．出力された単語は 5.4.2 節で実験結果として載せる．本実験では，手製トゲワードリストと拡張トゲワードリストをそれぞれ用いた実験を行う．

4.3　3 種類の手法の詳細

3 種類の手法の詳細を表 4.1 の例文を用いて説明する．全てのツイートにトゲワードの “馬鹿” が含まれるが，人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」に使われているため誹謗中傷であるとは思わない．また「最高」や「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 4.1 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．

4.3.1　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内の単語とのパターンマッチによって判定が行われる．表 4.1 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 全てのツイートを誹謗中傷と判定する．この手法には人間の感覚を擬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．

4.3.2　手法 2：ポジティブ・ネガティブ（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」[6] を用いた．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内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，誹謗中傷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 4.1 の II のツイートを解析させると，「馬鹿」が −1，「最高」「楽しい」などのポジティブな単語が 2 つあるので +2 となり，全体の文章のスコアは 1 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誹謗中傷と判定している．よって手法 2 を用いることで，手法 1 により判定された誹謗中傷の候補から II のツイートを除外することができる．

4.3.3　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 4.1 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[7] を用いて解析する．この手法では，あるツイート内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，誹謗中傷であると判定する．

手法 3 で誹謗中傷と判定される例文を，CaboCha に係り受け解析させ，出力された結果を図 4.2 に示す．この結果では 1 つの文章が 5 つの文節に分けられ，文節毎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 4.2 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発見すると，トゲワードが「人間」に対して使われていると判断し，誹謗中傷と判定を下す．図 4.2 の例では，3 つ目の文節でトゲワードの “馬鹿” を発見し，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，誹謗中傷と判定している．ここで，何 k ホップまで辿るのが最

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 4.2　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 4.3　トゲワードと固有名詞・人称代名詞に係り受けの関係がない例

良か議論する必要があるので，6 章の考察で検証する．表 4.1 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誹謗中傷と判定する．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，誹謗中傷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）にある文節を辿ることで，固有名詞または一人称を除く人称代名詞を発見するまで k ホップ辿るといった方式を採っているが，一方で，図 4.3 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」という固有名詞を発見できない．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，誹謗中傷と判定を下す．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 4.3 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，誹謗中傷と判定している．

第 5 章

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の比較実験を，手製トゲワードリストと拡張トゲワードリストを用いてそれぞれ行う．

5.1 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基に，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. 提案手法がトゲあり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 5.1 に示す．ツイート例のように本実験で扱うツイートの中には必ず「K さん」という固有名詞が含まれている．これはツイートを抽出する際に「K さん」をキーワードとしたからであり，意図としてはエゴサーチによって受ける誹謗中傷による被害を減らすことを目的としたためである．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．

5.2 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く，いくつかの種類に分けられると考えている．例えば，「お前のこと嫌いだわ」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「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

表 5.1　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
K さん存在からして無理なんだが。
ここまで貶されてるの K さん惨めでめちゃくちゃ面白いｗｗ

傷の対象が相手自身であるか，または相手の所有物や所属している組織であるか，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依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

5.3　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[8] 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，ネットタレントの活動名や「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，解析精度の低下，あるいは偶然に正解の結果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．

5.4　実験結果

本研究で行った実験は，手製トゲワードリストを用いた誹謗中傷の判定，WordNet による類義語の自動追加，拡張トゲワードリストを用いた誹謗中傷の判定の 3 つである．精度評価に用いる式を以下に示す．

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.4.1　手製トゲワードリストを用いた実験結果

提案手法の実験結果を表 5.2 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 で誹謗中傷であると誤検出されたトゲなしツイートを，ツイートのネガティブ度や，トゲワードの対象が何かを考慮することによって，誹謗中傷であると検出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，再現率の低下を可能な限り抑えることを目標としている．しかし，表 5.2 の手法 1+2 の結果を確認すると，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，再現率が手法 2 と同様に下がっているが，適合率は上がっている．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分かる．

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，4.3.3 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 5.1〜5.3 にそれぞれ示す．

図 5.1 では，手法 1+3a の方が適合率が高くなっていることが分かる．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 5.2 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分かる．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発見す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分かる．

図 5.3 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分かる．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分かる．

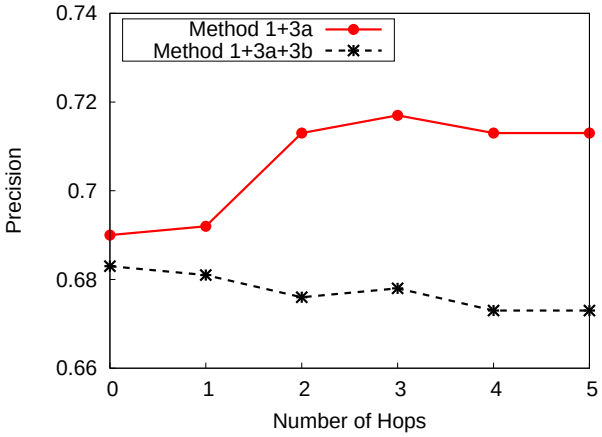

図 5.1　手法 1+3 のホップ数と適合率

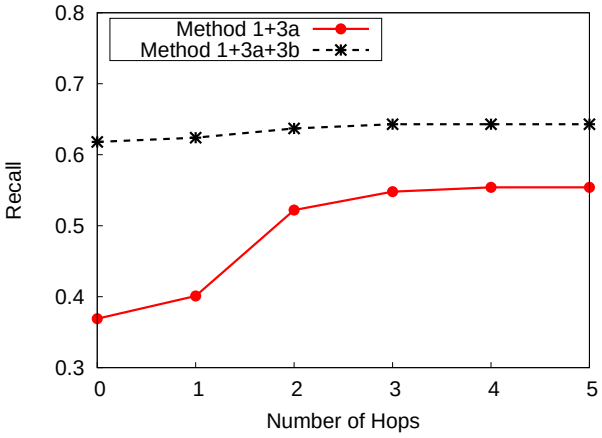

図 5.2　手法 1+3 のホップ数と再現率

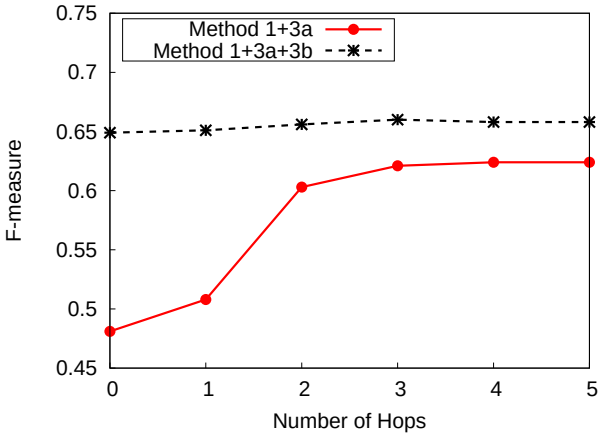

図 5.3　手法 1+3 のホップ数と F 値

5.4.2　WordNet によって出力された単語

拡張トゲワードリストは，手製トゲワードリストに含まれる 138 単語のうち類義語を除く 40 単語のトゲワードを，WordNet に入力し得られる類義語を追加したものである．入力するトゲワードの詳細は付録 B に示す．WordNet の入力において，命令形のワードは出力されないため「死ね」や「消えろ」などの単語は基本形の「死ぬ」や「消える」に変換する．収録されている単語数の詳細は，WordNet によって出力された類義語のうち重複を除いた 680 単語を追加した全 818 単語である．以下に WordNet によって出力されたトゲワードの類義語の例を示す．

「馬鹿」の類義語として出力された「脳たりん」や「痴れ者」や，「殺す」の類義語である「ぶち殺す」などは，手製のトゲワードに追加する条件を満たしているが，一方で「ごみ」の類義語である「破片」や，「臭い」の類義語である「アロマ」などの，誹謗中傷として使われるとは考えられない単語も含まれる．ここで著者が，自動追加された 680 単語に対してトゲワードとして条件を満たすかどうかの評価を行った．結果，トゲワードとして適する単語が抽出された割合（適合率）は 0.57 であった．

破片，がらくた，脳たりん，痴れ者，あんぽんたん，アロマ，異臭，眠らす，ぶち殺す

5.4.3　拡張トゲワードリストを用いた実験結果

WordNet により拡張されたトゲワードリストを用いた結果を表 5.3 に示す．手法 1 の再現率は，拡張前の手製トゲワードリストを用いた実験よりも 0.07 上がっている．一方で適合率が 0.124 下がってしまっている．手法 1 により，手法 2，3a，3b で解析するツイートが増えるため，他の実験の精度も変わっている．表 5.3 の 3 ホップ目まで辿るときの手法 1+3a の実験結果では，拡張前の辞書を用いた実験よりも再現率は 0.083 上がり適合率が 0.078 下がってしまったが，再現率の上がり幅の方が大きかったため F 値は 0.014 と若干上がっている．手製トゲワードリストを用いた実験における組み合わせの手法の中で最も精度の良かった手法 1+3a+3b の実験結果では，手製トゲワードリストの結果とは異なり，拡張トゲワードリストを用いた場合，手法 1+3a に手法 3b を追加することで F 値が下がっている．

表 5.2　手製トゲワードリストを用いた実験結果

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

表 5.3　拡張トゲワードリストを用いた実験結果

手法					評価		
1	2	3a	#hop	3b	P	R	F
✓			–		0.533	**0.828**	**0.648**
✓	✓		–		0.536	0.567	0.551
✓		✓	0		0.657	0.414	0.508
✓		✓	1		0.648	0.446	0.528
✓		✓	2		0.639	0.599	0.618
✓		✓	3		0.639	0.631	**0.635**
✓		✓	4		0.629	0.637	0.633
✓		✓	5		0.629	0.637	0.633
✓		✓	0	✓	0.568	0.688	0.622
✓		✓	1	✓	0.567	0.701	0.627
✓		✓	2	✓	0.557	0.713	0.626
✓		✓	3	✓	0.554	**0.720**	0.626
✓		✓	4	✓	0.546	0.720	0.621
✓		✓	5	✓	0.546	0.720	0.621
✓	✓	✓	0		0.620	0.280	0.386
✓	✓	✓	1		0.608	0.306	0.407
✓	✓	✓	2		0.596	0.414	0.489
✓	✓	✓	3		0.600	0.439	0.507
✓	✓	✓	4		0.598	0.446	0.511
✓	✓	✓	5		0.598	0.446	0.511
✓	✓	✓	0	✓	0.566	0.490	0.526
✓	✓	✓	1	✓	0.564	0.503	0.532
✓	✓	✓	2	✓	0.549	0.503	0.525
✓	✓	✓	3	✓	0.545	0.503	0.523
✓	✓	✓	4	✓	0.541	0.503	0.521
✓	✓	✓	5	✓	0.541	0.503	0.521

第 6 章

考察

本章では，評価実験で正しく検出されたトゲありツイートや検出されなかったトゲありツイート，誤検出されたトゲなしツイートを基に各手法の考察を行う．また，手製トゲワードリストと拡張トゲワードリストを用いた実験結果を比較し，WordNet を用いることについて考察を行う．

6.1　手製トゲワードリストを用いた実験結果の考察

表 5.2 から全体的な結果を確認すると，手法 1 だけを用いた実験結果における適合率を上げるために，手法 2，3a，3b を追加したにもかかわらず，手法 1 だけの解析精度が一番高くなっている．各手法毎の結果から，手法 1+2 は期待していた精度とはならず，手法 1 だけの実験結果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，3b は，F 値を上げられなかったものの，適合率は手法 1 だけを用いた実験結果を超えることができた．また，ここに手法 3b を加えた手法 1+3a+3b の実験結果からは，3 ホップ目まで辿った時の F 値が 0.660 となっており，組み合わせた手法の中では精度が一番高いことが分かる．

6.1.1　手法 1 の精度について

本実験の目的は誹謗中傷をどれだけ取りこぼさずに検出できるかである．よって，手法 1 ではまず再現率を上げる必要がある．手法 1 の解析結果では，ツイート 500 件のうち 181 件を誹謗中傷であると検出している．このうち，正しく誹謗中傷と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分かる．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．再現率を上げるためにはこのようなトゲワードにも対応しなければならないが，まずは拡張トゲワードリストを用いた実験結果を 6.2 節にて確認した後に考察することとする．

また誤検出されたトゲなしツイートは 62 件であった．誹謗中傷を減らすシステムとして，

誹謗中傷でないツイートを削除したり投稿を防ぐことは避けたいところなので，誤検出を減らすことで適合率を向上させる手法を考える必要がある．後の考察で手法 2 や手法 3 が，この 62 件のトゲなしツイートの排除にどれ程貢献し適合率を上げることができたか詳述する．

6.1.2　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 2 は，トゲワードを含むか否かで判断する手法 1 とは異なり，そのツイートがポジティブとネガティブのどちらであるかを考慮することで，より正確に誹謗中傷だけを抽出できる可能性が高いという仮説に基づいて考案した手法である．しかし結果は，手法 1 が誤検出した 62 件のトゲなしツイートのうち，手法 2 が正確に排除したツイートはわずか 19 件であるのに加え，誤って排除したトゲありツイートが 40 件となった．

まずは排除できなかったトゲなしツイートについて考察する．排除できなかったトゲなしツイートは全部で 43 件であるが，その内容は全て意見や感想，同情などであった．例文を表 6.1 に示す．本実験でツイートを抽出したタイミングが，K さんの不祥事で炎上した時期であったため，トゲなしツイートの中には表 6.1 のような内容が多かった．つまり手法 2 は，感謝や祝福などの内容のトゲなしツイートがデータに含まれる場合には，正しく排除し適合率を上げる可能性も考えられるため，本実験では排除できたトゲなしツイートは少なかったが一概に間違った手法であるとは言えない．

一方で，誹謗中傷であるのにも関わらず誹謗中傷でないと分類してしまった 40 件のトゲありツイートについて考察する．例文を表 6.2 に示す．40 件のトゲありツイートの特徴は 4 つに分類できる．1 つ目は K さんに対しては誹謗中傷しているのに，文章全体で見ればスコアが 0 未満にはならないためポジティブな内容であると判定される場合である．このようなツイートは 40 件中 3 件確認された．2 つ目が否定語の解析ミスによる失敗である．本提案手法の PN 判定には，日本語評価極性辞書に登録されている単語に否定語が用いられていた場合にはスコアを反転させている．しかし実験結果では「ない」や「ず」などには正しく反応するが，「人権なくて」の「なくて」のように反応しない場合があることが分かった．このようなツイートは 5 件確認され，原因は MeCab を用いた分かち書きのミスであり，「なくて」が否定語として認識できなかったためである．3 つ目は「あたおか」のような造語が日本語評価極性辞書にないためスコアが付与されずに判定されなかった場合である．このようなツイートは 24 件確認された．4 つ目は皮肉や煽りとして用いられている「さすが」や「おもしろい」などの単語がポジティブとしてスコア化されてしまう場合である．「さすが」や「おもしろい」などの単語は使い方によってはポジティブになる単語であることは分かるが，文脈を考慮せずに単語だけで判定すると表 6.2 の最後のツイート例のようにポジティブと判断され誹謗中傷でないと判定されてしまうことが分かった．このようなツイートは 8 件確認された．

現状，PN 判定の改善点としては，造語を含めた日本語評価極性辞書の拡張や，単語別に付与する数値を変えたりすることでツイートのスコア（ネガティブ度）を複雑にすることを検討し，上述した 4 つの失敗例を減らそうと考えている．

表 6.1　手法 2 で排除できなかったトゲなしツイート

陰湿ないじめ K さんが流石に可哀想だわ（同情）
K さんの虚言癖ヤバい（批判）
A さん達と K さんの喧嘩怖い（感想）
A さんのこと陥れようとしたのに，K さん可哀想って思う奴いるのか（意見）

表 6.2　手法 2 で検出できなかったトゲありツイート

A さんは良かったのに，K さんは不快だった
K さん人権なくて草
K さんあたおかすぎる
K さんさすがにひどすぎて面白いわ

6.1.3　手法 3a の精度について

手法 3a の精度は期待とは少し遠かったが，トゲワードの係り受け先を認識することで適合率を上げることが可能になるという仮説に基づいた精度を出したと言える．表 5.2 から，手法 1+3a の実験結果では，手法 1 に比べて適合率が 0.06 上がり，再現率が 0.21 下がっている．

適合率が上がった原因は，手法 1 で誤検出された 62 件のツイートを 28 件排除したことである．一方で排除できなかった 34 件の例を表 6.3 に示す．誤検出の原因の多くは，係り受け解析は正しく行えたものの内容は誹謗中傷でなかったことである．「可哀想」が擁護や同情として使われる場合と，惨めであると罵倒目的で使われる場合を区別させることは難しいため，トゲワード検出された単語の係り受け先を認識するだけでは困難であると考えられる．また，「やばい」の対象が「虚言癖」というその人が嘘をついた行動に対しての発言でも，係り受け解析では最終的に「K さん」という固有名詞を発見するため誤検出することが分かった．「虚言癖」が「動画」や「顔」などの対象の作品（所有物）であったり容姿であると誹謗中傷となるという点から，何が「やばい」のかということを機械が認識しなければならない．本実験ではトゲワードの係り受け先が，固有名詞あるいは人称代名詞であるかを考慮した分類を行ったが，今後の実験では「行動」に対する意見や批判のツイートの分析には，トゲワードと結びつく単語が何であれば誹謗中傷となるのかを考慮した分類方法が必要であると考えられる．

再現率を大幅に下げてしまった原因は，手法 3a が検出できなかったトゲありツイート（誤排除されたツイート）が 33 件あったことである．例文を表 6.4 に示す．まずは一番上の例文では図 4.3 でも説明したように，文章が句点で区切られることでトゲワードの対象を発見できなくなる場合である．同じ例が 33 件中 8 件確認された．このような例が手法 3b を加えることで検出可能になるかを 6.1.5 節で考察する．また二番目の例のように，「K さんのこと私は嫌い」のような逆に一人称の主語を正確に書いてある文章には，解析方法の性質上，「嫌い」が「私」に係ってしまうため，トゲワードの対象を発見できない場合も確認された．このよう

表 6.3　手法 3a で排除できなかったトゲなしツイート

K さんが可哀想だ.
まじ K さんの虚言癖がやばい
結局 K さんが嫌いなのか、A さんがいじめてるから炎上してるのか、どっちなの

表 6.4　手法 3a で検出できなかったトゲありツイート

A さんと K さんの動画見てないけど、なんだこれ。きっしょいわ
K さんって初めて知ったけど、私本当に大嫌いだわこの人
K さんさすがにヤバいわ??おもしろいけど??

な例はこの 1 件のみであった．こちらの例文も同様に 3b を加えることで検出可能となる事を期待する．最後のツイート例が，手法 3a が良い精度を出せなかった一番の原因である．手法 3a の働きでは，この例のツイートは誹謗中傷と判定可能であるが，実際にこのようなツイートが誤排除されている．原因は係り受け解析のミスである．係り受け解析のミスによる失敗例は 24 件である．この例文では本来，「K さん←さすがに←ヤバ」という係り受け関係を持つ構造になるのだが，解析結果では「ヤバ←おもしろいけど」となっており，何に対しての「ヤバ」なのか係り受け先を発見できないというミスが起こっている．原因はツイート内の「??」にある．この記号はツイートを抽出する際，正しく抽出することのできない絵文字などがクエスチョンマークに変換されたものである．この記号を句点に直すことで正しく解析されることは確認済みなので，句読点以外の記号を句点に修正したり，記号のある場所で文を区切り，文章を分けて係り受け解析させるなど工夫が必要である．他のツイートの解析ミスの中には，助詞が抜けていたり，形態素解析をする際の MeCab の辞書に登録されていない単語（未知語）などが原因となる場合が多かった．このようなツイートを正しい文章に修正し提案手法に入力すれば，精度は上がる可能性があるが，助詞の脱字や未知語などの自動修正は難しく，解決策は未だ検討できてない．

6.1.4　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，この仮説が正しいのかを確かめる．

図 5.2 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，誹謗中傷と判定されたトゲありツイートの例を表 6.5 に示す．実際に誹謗中傷と判定された，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

表 6.5　ホップ数毎の検出されたトゲありツイート

#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。

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，誹謗中傷と判定される．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 5.1 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” に対しての批判や感想であるため，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．

6.1.5　手法 3b の精度について

手法 3b の目的は，手法 3a の解析の失敗により検出できなかったトゲありツイートを，真の対象語を特定する “主題語の選択” を行うことで誹謗中傷と判定することを可能にし，再現率を向上させることである．図 5.2 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分かる．手法 3b を加えた実験結果では，手法 1+3a の 3 ホップ目まで辿る時の精度と比べてトゲありツイートを 15 件多く検出することが可能になった．正しく検出できた例を表 6.6 に示す．手法 3a で課題となった，句点で区切られ主語が省略されているツイートや，対象よりもトゲワードに近い位置に一人称があるツイート例において、真の対象語である主題語を求めることで検出を可能にした．また，改善であるとは言えないが，係り受け解析の失敗により検出されなかったツイートにも主題語を求めることによって検出が可能になった．これらの点においては主題語を設けた意味があったと言える．一方で検出できなかったトゲありツイートは 18 件である．これらの例文は 140 字近くのツイートが多く文章が長いという事と，解析が複雑という理由で説明しにくいという点から表として提示しない．失敗例の特徴としては主題語の選択が人間ではなかったという点であり，選択された主題語の例を挙げると「ため」や「方」,「詐欺」などであった．「ため」については名詞と判定されることが問題である可能性がある．「詐欺」が主題語になった例文について詳述する．ツイート内容は，「K さんのレギンスの件で何が詐欺なのか逆に教えてほしいわ本当に。何でも粗探しして詐欺と言えばいいと思ってる。ばかじゃん。」

表 6.6　手法 3b で検出可能になったトゲありツイート

A さんと K さんの動画見てないけど、なんだこれ。きっしょいわ
K さんって初めて知ったけど、私本当に大嫌いだわこの人
K さん凸してくる人に話被せまくってうざい口悪いし
K さんさすがにヤバ??おもしろいけど??

表 6.7　手法 3b で誤検出されたトゲなしツイート

酔っ払っていたのかしらないけど，K さんの誤字に恐怖を感じる
K さんは身を張って怖さを教えてくれたんだね
返金されないのは変だけど，K さんの自叙伝を買った私がバカだと思う．

というものである．このツイートのトゲワードの「ばか」の真の対象語は「詐欺と言っている人」に対してなので，対象がツイート内に登場していないという点で起こった失敗の可能性が考えられる．実際「詐欺」という単語はこのツイートにおいては主題語に選択されても間違いではないと思えるが，提案したシステム上，「詐欺」は人間（固有名詞）ではないので誹謗中傷とは判定されなかった．このように選択は正しいがトゲワードの対象とはならないというツイートも失敗例に含まれていることが分かった．

また手法 1+3a+3b では手法 1 が誤検出した 62 件のトゲなしツイートのうち，48 件のツイートを排除することができなかった．手法 1+3a の時点では排除できなかったツイートが 34 件であったことから，手法 3b を加えることで 14 件多く誤検出したということが分かる．手法 3b を加えることで誤検出されるようになった 14 件のツイートを表 6.7 に示す．原因は，行動に対する批判や意見であるツイートと誹謗中傷との区別が難しいことや，主題語の選択ミスである．表 6.7 の一番下のツイート例においては，主題語に「K さん」が選択されているが，人間の感覚では「私」が主題語であると思えるため，間違った選択であると考えている．他の上 2 つのツイート例においては，「K さん」が主題語と選択されるのは正しいが，行動に対する意見や批判あるいは感想であるため，この失敗例を解決するのは係り受け解析や主題語の選択だけでは困難であると考えている．

6.2　拡張トゲワードリストを用いた実験結果の考察

拡張トゲワードリストに WordNet を用いたことについての考察と，辞書を拡張する事で各手法の精度がどのように変わったか考察する．結論から言うと，WordNet を用いることは間違いであったとは思わないが拡張の仕方に工夫が必要である．

6.2.1　類義語の選択精度について

まずは WordNet が出力した類義語について考察する．5.4.2 節より，出力された単語がトゲワードの条件を満たしている割合は 0.57 である．トゲワードの類義語であるにも関わらず

数値が低い原因は，基本形に直すことでトゲワードの条件から外れた「消える」や「死ぬ」の類義語がほぼ全てトゲワードではなかったためである．一方で「馬鹿」や「恥」などの類義語はそれぞれ 54 単語と 34 単語であり割合的にも類義語を多く持つ単語であったが，全てがトゲワードの条件を満たしていた．また「ごみ」や「くそ」,「くず」などの類義語はトゲワードとなる割合が低かった．

結果をまとめると「馬鹿」や「恥」などの人に対して使われる可能性の高いトゲワードの類義語は，かなり高い確率でトゲワードとなり，「ごみ」や「くそ」などの人以外に対しても使われる可能性の高いトゲワードの類義語は，トゲワードになりにくいということが分かった．「消える」や「死ぬ」の類義語は，命令形に直せばほぼ全ての単語がトゲワードの条件を満たすため，命令形に直す作業を行うことで精度の向上が考えられる．

6.2.2　各手法の精度について

WordNet を用いることで再現率を上げることができたが，一方で適合率の下がり幅が再現率の上がり幅よりも大きく，各手法で精度が下がる結果となった．WordNet を用いたことで検出が可能になったトゲありツイートを表 6.8 に示す．検出されたトゲワードは，手製トゲワードリストではあえて登録しなかった単語が検出された例が多い．主観ではあるが手製トゲワードリストを作成する上で，例えば「駄目」や「悪い」は,「死ね」や「馬鹿」に比べて一つの単語だけでは誹謗中傷になりにくいと考えている．例えば「馬鹿」は相手に対して用いられていれば誹謗中傷と取れるが,「悪い」や「駄目」は相手に用いられても行動に対してならば意見や批判となる可能性があり，その単語一つでトゲとなる性質を持たないと判断しているからである．これは著者の感覚であるが，表 6.8 の例において正しく検出されるべきトゲワードは「性格悪い」や「黒歴史」であったり,「駄目」に関してはトゲワードとは考えにくく，実際は「K さんと関わると良いことがない」という内容が誹謗中傷に当たるのであって正解ではない．しかしながら，このような単語（トゲワード）が含まれていれば誹謗中傷の可能性があるという推測としては，正しく反映された結果と捉えることもできる．一方で，誤検出してしまったツイートは 52 件であった．例を表 6.9 に示す．この例では「頑」や「すごい」や「悪い」などがトゲワード検出され，批判や意見などのトゲなしツイートにも反応することが分かる．ワードを増やすと誤検出が増えるという予想の通りの結果となったと言える．

拡張トゲワードリストを用いた手法 1 の結果では，検出された単語に疑問が生じたが，ここで手法 2 や手法 3a，3b の精度にどれだけ影響を与えたか結果を確認する．手法 2 や手法 3 が解析するツイートは手法 1 によってトゲワードが検出されたツイートであるため，まずはこの解析するツイートの件数を比較する．拡張トゲワードリストを用いてパターンマッチ（手法 1）を行った結果が，表 6.10 より，トゲありツイート 130 件とトゲなしツイート 114 件の合計 244 件となっており，手製トゲワードリストを用いた実験結果と比べて手法 2 や手法 3 が解析するツイートが 63 件増えている．誹謗中傷でないツイートの含有率は手製トゲワードリストを用いた手法では 34 %，拡張トゲワードリストを用いた手法では 47 %となっており，拡張トゲワードリストを用いた手法 2，3a，3b の重要性が高まったと言える．各手法がどのような働

表 6.8　WordNet を用いたことで検出可能になったトゲありツイート

K さんと関わったら全員駄目になるな
K さんの動画投稿見てかなり性格悪いなって思いました。
K さんすごいな w こういう人絶対関わりたくねえ
ほんと病院行きなよ、K さんって検索してみ？黒歴史しか載ってないよ

表 6.9　WordNet を用いたことで誤検出されたトゲなしツイート

K さん側の動画に出演を頑なに拒否が事実ならちょっと首を傾げる
登録者伸びてて K さんやっぱすごいなって思った
K さんの A さんに出て来る時の立ち回りが下手すぎて勿体なさすぎる
K さんは自分に都合悪いところもひっくるめて全部載せなよ

きをしたのか，正しく検出したトゲありツイート（TP）と誤検出したトゲなしツイート（FP）を比較した棒グラフを作成した．棒グラフを図 6.1 に示す．図 6.1 における手法 3 を含むデータは，手製トゲワードリストの実験において精度が一番高かった 3 ホップまで辿った時の結果である．また拡張トゲワードリストを用いた手法を以下，手法 Ex（Extend）と呼ぶ．

拡張トゲワードリストを用いた手法 1+2Ex は，手法 1Ex で誤検出したトゲなしツイートを 37 件排除できているが，同時にトゲありツイートも 41 件排除しており，目的の働きは果たせなかった．誤検出されたトゲなしツイートがかなり多くなっている原因は，6.1.2 節でも考察したように，誹謗中傷ではないがネガティブな内容であるツイートが多かったためである．

手法 1+3aEx ではトゲなしツイートを 58 件排除できており，手法 1+2Ex と比べると 21 件多く排除できていることが分かる．またトゲありツイートを 31 件排除してしまっているが，手法 1+2Ex と比べると 10 件だけ少なく抑えられていることが分かる．手法 1+2Ex の結果とは異なり，解析するツイートの中にトゲなしツイートが増えたのにも関わらず正確に排除することができた上，誹謗中傷を正しく検出したという傾向が確認できるため，係り受け解析を用いた手法 3a は，ある程度目的通りの働きをしていることが分かる．

一方で手法 1+3a+3bEx では，拡張トゲワードリストを用いることによって誤検出されたトゲなしツイートをあまり排除することができず，誹謗中傷と判定されていることが分かる．手製トゲワードリストを用いた結果では，手法 1+3a に比べて再現率の低下を抑えられていたが，拡張してもその働きは変わらずトゲありツイートの誤排除は他の手法よりも少なかったが，一方で偽陽性（FP）が多かったことから，本実験によって手法 3b はトゲありツイートとトゲなしツイートのどちらとも誹謗中傷と判定する可能性があることが分かった．しかし抽出された主題語を確認すると，そのツイートの主題語となる固有名詞（K さん）を正しく選択している例が多かった．本実験の 3b の精度が悪くなった原因としては，プロセスに問題があったというよりは，全てのツイートに固有名詞（K さん）が含まれている点であり，手法 3b にとっては誹謗中傷と判定するのが難しい内容であったと考えられる．

表 6.10　手法 1 の TP（陽性）と FP（偽陽性）の比較

	TP	FP	合計
手法 1(手製)	119 件	62 件	181 件
手法 1(拡張)	130 件	114 件	244 件

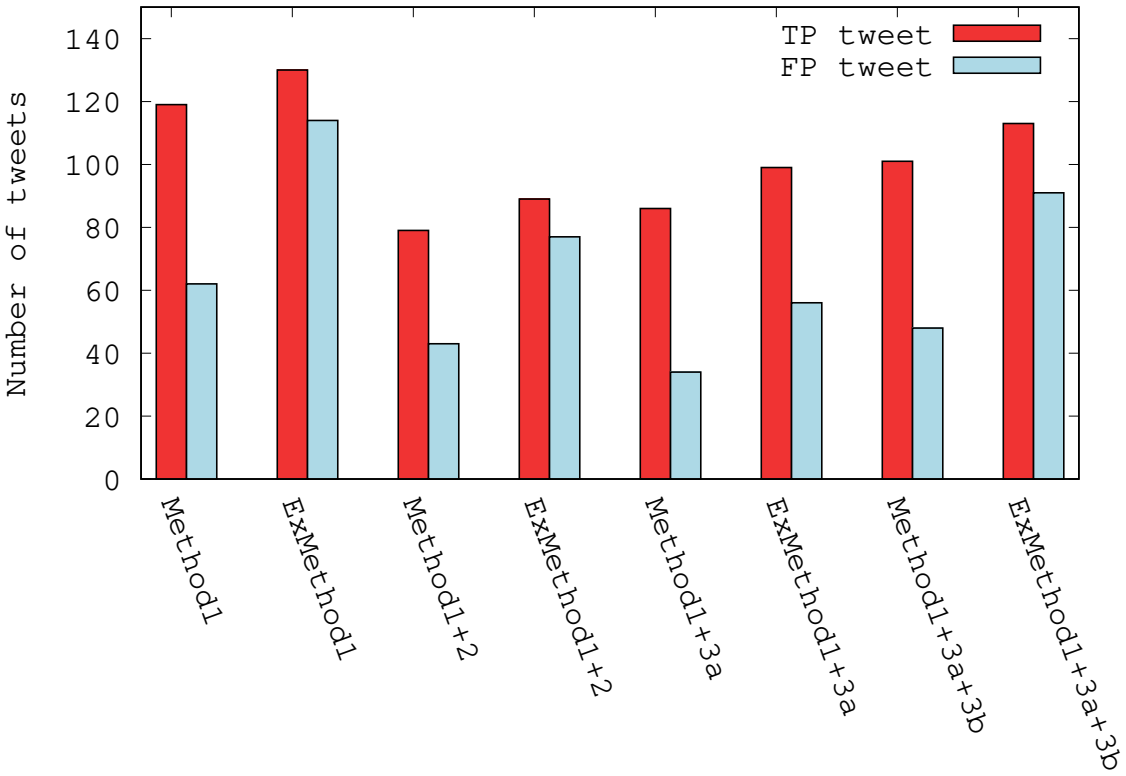

図 6.1　各手法の TP（陽性）と FP（偽陽性）の比較

第 7 章

結論

本章では各手法の結果から問題となった部分を簡潔にまとめ，今後の課題を述べる．最後に本研究の社会的貢献及び技術的貢献を述べ，本稿を終える．

7.1　まとめと今後の研究課題

本稿では，Twitter に投稿された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実験では，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．また「頭が悪い」や「顔が汚い」などの 2 つの単語が組み合わさってできるトゲワードについては，本実験で扱うトゲワードリストとの単語とは区別した．このように単語がペアとなってトゲワードとなる言葉には，単純なパターンマッチではなく一方の単語が出現した時，もう一方の単語が結びついて出現した場合にトゲワードと検出する仕組みを作りたいと考えている．仕組みを変える理由は単純なパターンマッチにおいて，「頭が悪い」というトゲワードは多数の表現が考えられるため，網羅するのが難しいことや効率が悪いからである．

また WordNet を用いたトゲワードの自動追加では，再現率を上げることに成功したが，辞書を拡張することによって検出可能になったトゲワードは「悪い」や「すごい」などの誹謗中傷だけでなく，批判や意見あるいは普通の内容であるトゲなしツイートでも使われるような単語も少なくなかった．この結果から追加する単語を選別する必要があるが，選別方法については思案中である．

PN 判定は，誹謗中傷でなくともネガティブな内容であるツイートが多かったため分類するのには適さない場合があることが分かった．一方で，ネガティブな内容であるのに正しく誹謗中傷と判定できないという問題も存在するため，この部分は日本語評価極性辞書の拡張，または，単語別に細かい数値が付与されている単語感情極性対応表 [9] などの他の辞書を用いることで精度向上を目指す．

係り受け解析においては，ホップ数が 3 の時に精度が一番高いという結果となったが，他のデータで実験したり，ホップ数の最大値を 5 から増加させたりするなど，さらに検討する．また 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．

最後に，本実験とは別の方法で抽出するツイートで解析を行うことで精度評価が変わる可能性について考える．本実験では，扱ったデータが「不祥事に対するユーザのツイート」あるいは「エゴサーチをした場合に見つかるツイート」と分類されるツイートへのトゲワード検出を行ったが，例えば「炎上しているツイートの返信欄」からトゲワード検出を行うと，トゲワードの対象がツイート内に含まれていない可能性があり手法 3 の精度が下がる可能性があると考えられたり，「アンチによる誹謗中傷ツイート」に対しては，本来祝福や激励のコメントが多くなるはずの内容のツイートに対しても誹謗中傷が含まれることがあるため，誹謗中傷のコメントだけが浮き彫りになる可能性があるため，このようなデータにおいては誤検出を避けつつトゲありツイートだけを誹謗中傷と判定することが可能となり精度が上がる可能性があると考えられる．このように扱うデータによって精度が変わってくるため，他のデータでも実験を行い足りない点を確認しながら，手法の改善や新たな手法を考案していきたい．

7.2　社会的貢献及び技術的貢献について

インターネット上における誹謗中傷は，インターネットが家庭に普及した 20 年以上も前から問題視されている．スマートフォンの普及やネット回線の高速化，使いやすいコミュニケーションツール（SNS）などが増え，現在インターネットを利用するユーザは急激に増えたと感じられる．誹謗中傷に関する問題は，こうした時代の変化とともに大きくなっている．世界中で多数のユーザが利用している SNS では自由に思うことを発信できる反面，メッセージの受け取り方は人それぞれ異なるため，時に危険な投稿をしてしまうユーザも少なくない．このような社会において，相手を傷つけてしまう可能性のある投稿はなくすべきであるが，様々な技術が進んでいる現在でも，インターネット上における誹謗中傷による被害はなくならない．

本研究の社会的貢献としては，誹謗中傷を自動で検出することが可能になれば，相手を傷つける可能性のある文章の投稿を防いだり，閲覧者側から見えなくすることが可能になり，誹謗中傷による被害を受けて辛い思いをする人が減り平和なネット社会に繋がると考えている．

本研究の技術的貢献としては，実験を通じて誹謗中傷の特徴や判定の難しさについて部分的であるが知見を得ることができたところである．誹謗中傷の特徴は，受け取り手に依ってダメージが異なる点で，その人の行動に対する批判であったとしても「棘のある言葉」が含まれていれば傷つける可能性がある上，「棘のある言葉」は相手を攻撃する強い言葉だけでなく，状況や文脈に依って一般的に相手を傷つけることのない言葉でも人を傷つける場合があるところである．用いられている単語の情報だけでも，PN 判定や係り受け解析を用いることで，その文章の特徴を擬似的に認識させ分類する手法を提案したが，精度向上の余地はあるものの誹謗中傷の判定をするためには，状況や文脈を判断させることが必要なため，本実験で行った手法だけでは不十分であることが分かった．インターネットの明るい未来のために，このような研究が進むことに貢献していきたい．

謝辞

本研究に際して，様々なご指導を頂きました服部峻助教と荒澤孔明先輩に厚く御礼申し上げます．また，日常の議論を通じて多くの知識や示唆を頂いた研究室の皆様にも深く感謝の意を表します．

参考文献

[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,” 言語処理学会第 16 回年次大会, pp.383–386 (2010.3)

[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] Guangwei Wang and Kenji Araki, “Modifying SOPMI for Japanese weblog opinion mining by using a balancing factor and detecting neutral expressions,” In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, pp.189–192 (2007).

[4] 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).

[5] 日本語 WordNet, http://compling.hss.ntu.edu.sg/wnja (2020).

[6] 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).

[7] CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer, https://taku910.github.io/cabocha (2020).

[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab (2020).

[9] PN Table, http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html (2020).

付録 A

手製トゲワードリスト

消えろ
キエロ
きえろ
ごみ
ゴミ
塵
馬鹿
バカ
ばか
雑魚
ザコ
ざこ
ざっこ
恥晒し
恥さらし
はじさらし
くそ
クソ
糞
くず
クズ
屑
あほ
安保
アホ
死ね
シネ
しね
４ね
氏ね
死刑
不細工
ブス
ぶさいく
ブサイク
ぶす
ｂｓ
老害
ろうがい
気持ち悪い
キモイ
ｋｍ
きしょい
きっしょ
きんも
キモ
きも
きめえ
きもちわるい
きもい
気持悪い
キモチ悪い
気持ちワルイ
キモチワルイ
気持ち悪
爺
ジジイ
じじい
ばばあ
BBA
JJI
老害
奴隷
問題児
基地外
キチガイ
害悪
ガイジ
害児
情弱
弱者
変態
痴漢者
犯罪者
殺す
コロス
ころす
轢き殺す
５６す
陰キャ
インキャ
クサイ
臭う
くさい
障害持ち
障害者
貧乳
ぺちゃぱい
ペチャパイ
デブ
でぶ
汚い
汚らしい
きたない
貧乏人
ホームレス
浮浪者
肉便器
ヤリマン
でべそ
短足
たんそく
無理
むり
ムリ
嫌い
きらい
嫌われ
きらわれ
キライ
嫌わ
うざい
うっざ
ウザい
ウザイ
ヤバ
やば
やんば
やっば
ヤッバ
不快
不愉快
怖い
怖
こわ
イキる
いきる
いきんな
イきる
可哀想
かわいそう
無能
むのう
低俗
頑固
がんこ
あたおか
腹黒

付録 B

WordNet に入力したトゲワード

消える
ごみ
馬鹿
雑魚
恥
くそ
くず
あほ
死ぬ
不細工
老害
気持ち悪い
爺
奴隷
障害者
情弱
犯罪者
殺す
陰キャ
臭い
貧乳
デブ
汚い
貧乏人
短足
無理
嫌い
うざい
ヤバ
不快
怖い
イキる
可哀想
無能
低俗
頑固
あたおか
腹黒