研究指導 中澤 真 教授

Twitterにおける誹謗中傷抑止システムの精度向上に有効な特徴量の検討

佐藤　千夏

1. はじめに

近年，インターネットの普及によりSNSを介して行われる誹謗中傷[1]が社会問題となっている[2]．特に2020年，誹謗中傷が原因とされる著名人の自殺が多くのメディアで取り上げられ，そのことによって社会の誹謗中傷に対する関心が高まったと推測できる[2]．そのため，SNS上の誹謗中傷に関する対策が現在練られている．対策の例として，SNSの利用についての啓発活動や，投稿内容の規制が挙げられる[3][4]．また，近年注目されているのが誹謗中傷抑止システムであり，システムの開発や精度向上を目的とした研究がされている[5][6][7]．

このようにSNS上の誹謗中傷問題の改善に向けた取り組みがされているものの，十分な効果を上げているとは言い難い．その理由として，誹謗中傷ツイートの判別材料になる特徴語・特徴量の選定が難しいことが挙げられる．例えば伊藤ら[5]では，誹謗中傷に用いられる単語とその対象となる語の係り受け関係から特徴量を選択したが，誹謗中傷の判別精度が不十分であった．よって，精度を高めるためにも新たな特徴量を検討する必要がある．

そこで本研究では，Twitterを対象にした誹謗中傷抑止システムの精度向上を目指し，単語だけではなく文字数や品詞情報に着目した特徴量を検討する．

2. Twitterにおける誹謗中傷抑止システム

2.1 Twitter上の誹謗中傷の現状

誹謗中傷とは誹謗と中傷という二つの語で成り立つ語である．誹謗とは他人の悪口を言ったり罵ったりすることを指し，中傷とは根拠のないことを言いふらすことで他人を傷つける行為と定義される[1]．このような行為が主にSNSを介して行われており，現代の社会問題となっている[2]．

SNSの一つであるTwitterは，ユーザの投稿に返信できるアカウントを制限する機能や，通報を受けたアカウントの削除や停止などの誹謗中傷対策を行っている[3][4]．しかしこのような対策が取られているにもかかわらず，違法・有害情報相談センターに寄せられる相談件数の割合は，TwitterがSNSの中で最も多いというのが現状である[3]．

2.2 誹謗中傷抑止システムの現状課題

誹謗中傷の対策として，投稿内容の規制，SNSの利用に関する啓発活動が実施されている[3][4]．中でも近年注目されているものは誹謗中傷抑止システムである[5][6]．これらのシステムは，SNS利用者が投稿する際に投稿内容が適切なものか否かを判別し，誹謗中傷となる内容が含まれている場合には警告表示をする機能を持つ．ここで重要なのは，誹謗中傷であるか否かを判別する精度であり，これを左右するのが投稿内容の特徴量や特徴語である．このため，高い精度を得るためにどのような特徴量・特徴語を使うべきかという研究が取り組まれている．

例えば伊藤ら[5]の研究では，人手によって誹謗中傷によく使われる単語を判別のための特徴語としてリストを作成し，これらの単語との係り受け関係を特徴量として，誹謗中傷ツイートと誹謗中傷ではないツイートを判別する手法を示した．一方，川上ら[7]は誹謗中傷か否かではなく，炎上する内容か否かの判別を目的にシステム構築している．この研究では，単語の共起関係に着目して100語程度の語を選び出して特徴量としている．ただし，これらの研究は判別の精度にまだ改善の余地がある．伊藤らの研究では，句点や記号で本文が区切られた場合の対応ができず，川上らの研究では対象語に着目していないため，結果として抽出漏れが多くなってしまっている．

以上のことを踏まえて本研究では，誹謗中傷抑止システムの精度向上を目指して，誹謗中傷ツイートと誹謗中傷ではないツイートの判別精度を高めるために，ツイート全体の文字数や品詞情報といった幅広い要素を対象に，特徴量の検討・評価に取り組む．

3. 誹謗中傷判別のための特徴量

3.1 検証概要

検証のために，まず誹謗中傷に用いられやすい5つの単語「死ね」「タヒね」「消えろ」「くたばれ」「きしょい」(以後，誹謗中傷用語)をキーワードとしてツイートを収集した．収集にはTwitterの投稿データをCSV形式でダウンロードできる「ついすぽ[1]」を用いて1783件のデータを入手した．次に収集したデータを誹謗中傷ツイートと誹謗中傷ではないツイート(以後，NOT誹謗中傷ツイート)の2つのツイート群に分けるラベル付けを人手によって行った．判別精度を適切に検証できるようにするために，これら2つの群が同数になるようにサンプル数を揃え，結果としてどちらの群も557件のデータとした．

[1]https://tilde.afonomics.com/TweetExport/

ラベル付けでは以下の二つの条件を満たすものを誹謗中傷ツイートと定義した.

① 誹謗中傷用語がツイート内に含まれている.

② 誹謗中傷の対象が示されている.

なお, 誹謗中傷の対象としては, 組織や国など不特定多数を示すもの, あるいは特定のアカウントへのリプライやリンク[2]も含めている. また, キャラクター名のようなものが対象の場合も, 実在する人物のニックネームの可能性を排除できないため一律に誹謗中傷ツイートとした. 一方, 対象が「あいつ」や「お前」のような不明確な場合, 投稿者自身, 人以外の生物名の場合は, 誹謗中傷ではないツイートとして扱った.

次に検証に用いる特徴量について以下に示す.

1. ツイートの文字数
2. 記号間の平均文字数
3. 品詞別の出現の有無
   - 3-1. ツイート内全体を対象
   - 3-2. 誹謗中傷用語から一定語数以内を対象
   - 3-3. 誹謗中傷用語と係り受け関係があるものを対象(先行研究)

これらの情報を分析するために, 係り受け関係には「CaboCha[3]」, 形態素解析には「MeCab[4]」を用いた.

3.2 特徴量の検証

まず初めに, ツイートの文字数を特徴量とした検証について説明する. 感情に任せて勢いで誹謗中傷ツイートを投稿するユーザは短い文になりがちで, 一般のツイートと比べて文字数が少なくなると考えたことが, これを特徴量候補として扱う理由である. 検証方法としては, ユーザIDとスペースを除いて, 1ツイートに用いられる文字数をカウントする. これをTwitterの文字数上限である140文字を4分割した階級で, 誹謗中傷とNOT誹謗中傷の2つのツイート群で該当するツイート件数を表 1のように整理した.

表 1 文字数別のツイート件数の比較

階級値(字)	誹謗中傷(557件)	NOT誹謗中傷(557件)
1~34	429	309
35~69	85	132
70~104	21	71
105~	22	45

記号間の平均文字数を特徴量の候補とした理由は, 誹謗中傷ツイートが短文になる傾向にあるならば, 句読点などの記号によって区切られたフレーズも同様に短くなり, 両群の傾向に差異が現れると考えたためである. そのことも踏まえ本研究で対象とした記号は,「、」「, 」「。」「・」「;」「！」「!」「？」「?」「:」「;」「＃」「#」「“」「”」「^」「()」の17種類の記号とした.「. 」と「:」を含めなかった理由としては, これらはリンクのURLに使われる記号であり, Excel上でカウントする際に正確な値を導けなくなることを考慮したためである. ここでは, 対象の記号のいずれかによって文が区切られた場合, その記号間の平均文字数を特徴量としている. 差異を比較するために, 区間幅を10文字とする7階級に分け, それぞれの群における該当ツイート件数を表 2のようにまとめた.

表 2 記号間の平均文字数別のツイート件数の比較

階級値(字)	誹謗中傷(557件)	NOT誹謗中傷(557件)
0~10	282	163
11~20	177	205
21~30	50	92
31~40	22	32
41~50	9	27
50~60	2	13
61~	7	17

次に品詞別の出現の有無を特徴量とする理由について述べる. 伊藤らの研究[5]では, 誹謗中傷の対象語となる固有名詞と代名詞だけに着目していたが, 他の品詞の出現傾向を分析し, 精度への影響を明らかにすることで, 新たな特徴量の候補になり得ると考えたからである.

そこでまず, 2つのツイート群における品詞別の出現頻度を算出し, これらを説明変数, 誹謗中傷とNOT誹謗中傷のラベルを目的変数とする判別分析[8]によって, どの品詞が判別に有用であるかを有意差検定によって決定する(表 3参照). この結果, P値が5%有意水準で有意な形容詞, 助詞, 助動詞, 動詞を特徴量の候補と決定した. 以後は各品詞の出現の有無を特徴量とし, 先行研究で用いられていた品詞も加えて分析することとした. ただし, 一部の特徴量はカテゴリーデータであるため, 判別分析にかける際にはダミー変数[8]として処理するものとする.

表 3 判別分析による各品詞の有意性検定の結果

品詞名	P 値	品詞名	P 値
名詞	0.0940	副詞	0.9660
形容詞	**0.0245**	連体詞	0.0690
助詞	**0.0089**	接続詞	0.3915
助動詞	**P < 0.001**	感動詞	0.9889
動詞	**P < 0.001**		

本研究では, これらの品詞が判別精度に与える影響を分析する際に, ツイート内の任意の位置での出現の有無で考えるケースの他に, 誹謗中傷用語の近辺での出現の有無を特徴量とすることも考える. 誹謗中傷用語との関係性を考慮することにより, さらに適切な特徴量となる可能性があるからである.

係り受け関係に着目した伊藤ら[5]では句点や記号によって文が区切られると対象語を判別できない

[2] リンク先の内容については問わない.

[3] http://taku910.github.io/cabocha/

[4] https://taku910.github.io/mecab/

という課題があった. よって本研究では, 係り受け関係ではなく, 誹謗中傷用語から一定の語数の範囲内に出現する品詞の有無を考える. こうすることで, 句点や記号を超えた語句の関係も扱えるようになるからである. なお, 先の有意差検定の結果では名詞を除外することになったが, 先行研究の特徴量である固有名詞と代名詞の有無については比較のために検証対象に加える.

これら5つの特徴量を単体または組み合わせて分析に使用し, 最終的に誹謗中傷抑止システムの精度向上に適した特徴量を明らかにすることを本研究の目的とする.

4. 特徴量の分析と結果

4.1 ツイートの文字数

表 1に示したように誹謗中傷群のツイートは, NOT誹謗中傷群のものよりも文字数が少ない傾向にあることが読み取れる. また, 2つの群の文字数に関する代表値を算出すると, 誹謗中傷群の中央値は14, 最頻値は3, 平均値25.3となり, NOT誹謗中傷群の中央値は30, 最頻値は21, 平均値は43.4となった. いずれの値も誹謗中傷群のツイートのほうが約20字少なく, 2つの群に差異が生じていることが判明した. このことから, 文字数は誹謗中傷ツイートを判別するのに有効な特徴量候補になると考えられる.

4.2 記号間の文字数

表 2に示したように誹謗中傷群のツイートは, NOT誹謗中傷群のものよりも記号間の平均文字数が少ない傾向にあることが読み取れる. また, 2つの群の文字数に関する代表値を算出すると, 誹謗中傷群の中央値は10, 最頻値は3, 平均値13.7となり, NOT誹謗中傷群の中央値は15.5, 最頻値は15, 平均値は20. 4となった. いずれの値も誹謗中傷群のツイートのほうが少なく, 2つの群に差異が生じていることが判明した. このことから, 記号間の平均文字数は誹謗中傷ツイートを判別する特徴量候補として適切であると考えられる.

4.3 ツイート全体を対象とした品詞別の出現有無

収集した全ツイートを形態素解析し, 対象とした6種類の品詞それぞれの出現の有無を説明変数, 誹謗中傷か否かを目的変数として判別分析を実施した. この特徴量による判別的中率は67.15%となり, 精度向上に有効な特徴量だと考えられる.

4.4 誹謗中傷用語から一定語数以内を対象とした品詞別の出現有無

誹謗中傷用語から一定語数以内を対象とした品詞別の出現有無に関する検証では, 誹謗中傷用語からの範囲を1語から6語まで変化させ, それぞれに対して判別分析を実施した. ここで対象とした品詞は前述した固有名詞・代名詞・形容詞・助詞・助動詞・動詞の6つである. 結果として, 固有名詞および代名詞は5単語の範囲で最も判別的中率が高く68.49%となった. ほかの4つの品詞に関しては6単語の範囲で最も高い的中率となり59.78%であった. また, 1語および2語の範囲で判別的中率を求めた際の判別的中率が3語から6語より明らかに低かったため, ほかの特徴量と組み合わせる分析には用いないこととした.

4.5 誹謗中傷用語と係り受け関係があるものを対象とした品詞別の出現有無(先行研究)

先行研究で用いられた特徴量である係り受け関係についての結果としては, 表 4に示したように判別的中率は60.14%であり, ほかの特徴量よりも低いことがわかる. このことから, 誹謗中傷用語と関連性を考慮した各品詞の出現傾向を考える場合には, 係り受け関係よりもシンプルに出現位置に着目したほうが精度に良い影響を与えるといえる.

5. 特徴量の判別精度の検証

5.1 単体での判別的中率

まず, それぞれの特徴量を単体で用いた場合に判別精度に及ぼす影響を検証した. 判別分析では, 誹謗中傷か否かを目的変数, 各特徴量単体を説明変数としている.

表 4に示したように, 誹謗中傷用語から前後4単語の範囲内に対象品詞が出現しているか否かを特徴量とした場合の的中率が最も大きくなったが, まだ7割には到達できていない. そこで, これらの特徴量を組み合わせて, 判別精度を高めることを試みる.

表 4 特徴量単体の判別的中率

特徴量	判別的中率
係り受け解析(先行研究の特徴量)	60.14%
①範囲が1単語以内の6品詞の出現の有無	56.46%
②範囲が2単語以内の6品詞の出現の有無	61.85%
③範囲が3単語以内の6品詞の出現の有無	66.43%
④範囲が4単語以内の6品詞の出現の有無	**68.40%**
⑤範囲が5単語以内の6品詞の出現の有無	68.13%
⑥範囲が6単語以内の6品詞の出現の有無	68.04%
⑦ツイート全体での6品詞の出現の有無	67.15%
⑧ツイートの文字数	66.70%
⑨記号間の平均文字数	59.96%

5.2 組み合わせによる判別的中率

特徴量のあらゆる組み合わせを試し, 判別的中率の上位10種類を抜粋したものが表 5である. 上位10種類すべてに文字数が含まれていることから, 文字数は精度を高めるために必要な特徴量であるということが明らかとなった. また, 組み合わせたことによって判別的中率は7割を超え, 精度向上を目指す際に特徴量を複数組み合わせることが有用であるということが判明した. 特徴量の組み合わせとしては, ツイート文字数, 誹謗中傷の対象となり得る固有名詞・代名

詞が誹謗中傷用語から5単語以内に出現しているか否か，形容詞・助詞・助動詞・動詞のいずれかがツイート内に出現しているか否かの3種類が適切であると結論づけられる．

表 5 特徴量組み合わせによる判別的中率

特徴量	判別的中率
①ツイートの文字数＋ツイート全体での形容詞・助詞・助動詞・動詞の出現の有無＋範囲が5単語以内の代名詞・固有名詞の出現の有無	**70.74%**
②ツイートの文字数＋範囲が6単語以内の 6 品詞の出現の有無	69.57%
③ツイートの文字数＋範囲が6単語以内の助詞・形容詞・助動詞・動詞の出現の有無＋範囲が5単語以内の固有名詞・代名詞の出現の有無	69.48%
④ツイートの文字数＋記号間の平均文字数＋範囲が5単語以内の 6 品詞の出現の有無	69.21%
⑤ツイートの文字数＋範囲が5単語以内の 6 品詞の出現の有無	69.03%
⑥ツイートの文字数＋記号間の平均文字数＋範囲が4単語以内の 6 品詞の出現の有無	68.85%
⑦ツイートの文字数＋範囲が3単語以内の助詞・形容詞・助動詞・動詞の出現の有無＋範囲が5単語以内の固有名詞・代名詞の出現の有無	68.85%
⑧ツイートの文字数＋範囲が4単語以内の 6 品詞の出現の有無	68.49%
⑨ツイートの文字数＋範囲が4単語以内の助詞・形容詞・助動詞・動詞の出現の有無＋範囲が5単語以内の固有名詞・代名詞の出現の有無	68.40%
⑩ツイートの文字数＋記号間の平均文字数＋範囲が3単語以内の 6 品詞の出現の有無	67.68%

最も判別的中率が高くなった特徴量の組み合わせである①については，適合率と再現率[9]についても分析し，先行研究と比較した(表 6参照)．

表 6 特徴量①の再現率と適合率と F 値

	再現率	適合率	F値
本研究	66.4%	72.6%	69.4%
先行研究	71.7%	54.8%	62.1%

本研究の特徴量が，再現率と適合率を総合的に評価するF値[9]でも先行研究より高い精度であることが示された．また，再現率は先行研究と比較して若干劣るものの，適合率が大幅に向上した．よって，これらの特徴量を用いた場合，NOT誹謗中傷ツイートを誤って誹謗中傷ツイートと判別してしまうことが少なくなり，判別精度の向上につながると考えた．

さらに，使用した特徴量の誹謗中傷ツイートへの影響や役割を確かめるために，標準化判別係数で重要度を把握することにした．表 7の判別係数の絶対値が大きいものに注目することで，5単語範囲の固有名詞・代名詞の有無，ツイート内の助動詞の有無の影響度が大きいことがわかる．また，誹謗中傷の重心は負の値のため，標準化判別係数が相対的に大きな負の値となっている固有名詞・代名詞が誹謗中傷用語の近くにあると，誹謗中傷ツイートと判別される可能性が高くなることも明らかになった．一方，相対的に大きな正の値の係数に注目すると，助動詞が含まれていて文字数が多いツイートは，NOT誹謗中傷ツイートと判別される可能性が高くなることが読み取れる．

表 7 特徴量①の標準化判別係数

説明変数(特徴量)	標準化判別係数
ツイートの文字数	0.2691
ツイート全体での形容詞の出現の有無	0.1976
ツイート全体での助詞の出現の有無	0.1161
ツイート全体での助動詞の出現の有無	0.3386
ツイート全体での動詞の出現の有無	-0.0897
範囲が5単語以内の代名詞の出現の有無	-0.3737
範囲が5単語以内の固有名詞の出現の有無	-0.6538

6. 研究のまとめ

本研究では，誹謗中傷抑止システムの精度向上に有用な特徴量の種類と，その組み合わせ方を明らかにした．その結果，ツイートの文字数，ツイート内全体を対象とした品詞別の出現の有無，誹謗中傷用語から一定語数以内を対象とした品詞別の出現有無といった特徴量を組み合わせることが，最も判別的中率を高めるということが確認できた．また，本研究で扱った特徴量は先行研究で用いた特徴量である係り受け解析よりも高い判別精度となることも確認できた．

しかし本研究で用いた誹謗中傷用語は5つのみであり，ほかの誹謗中傷用語を対象とした場合の判別精度は確認できていない．このことから，新たな誹謗中傷用語を追加して特徴量の判別精度がどのように変化するかを明らかにすることが今後の課題である．

参考文献

[1] 株式会社 Hew One’s Way，誹謗中傷・ネット削除ガイド，https://hibou-tyusyou.help/guide/1279（参照:2021-07-01）

[2] 毎日新聞取材班，SNS 暴力―人はなぜ匿名の刃をふるうのか―，毎日新聞出版，2020.

[3] 総務省，インターネット上の誹謗中傷への対策，http://urx3.nu/YyvN (参照:2022-02-06)

[4] 総務省，総務省におけるインターネット上の誹謗中傷対策の取り組みについて，http://urx3.nu/n5vv (参照:2021-12-13)

[5] 伊藤ら,“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”，信学技報，Vol.120,No.311,IN2020-42,pp.7-12,2021.

[6] Rethink - Before the Damage is Done，https://www.rethinkwords.com/(参照:2021-07-01）

[7] 川上幹,瀰冨仁，“Twitter への投稿テキストによる炎上警告システムの構築”,第 32 回ファジィシステムシンポジウム講演論文集,pp.705-708,2016.

[8] 菅民郎，例題と Excel 演習で学ぶ多変量解析―回帰分析・判別分析・コンジョイント分析編，オーム社，2016.

[9] 金明哲，テキストアナリティクスの基礎と実践，岩波書店，2021.

[10] 鈴木美羽，“テキストの難易度に基づく子供向け Web ページ判定手法の提案”，会津大学短期大学部産業情報学科経営情報コース卒業研究論文要旨集，2019.