係り受け解析を用いた誹謗中傷判定に関する研究

情報科学科　檜谷 麻由　　指導教員：辻 孝吉 教授

1　はじめに

日本国内における SNS の普及率は，昨年 5 月時点で 82%となっており[1]，SNS 上での誹謗中傷が現在深刻な問題となっている． 昨年 5 月に総務省が発表したアンケートによると，「他人を傷つけるような投稿」を SNS ユーザの約半数が目撃している[2].このような投稿に対して，投稿の削除や，アカウントの凍結などの対応を人手で行うには限界がある．先行研究では，文章が誹謗中傷であるかを判定する手法がいくつか提案されている．

本研究では，係り受け解析を用いることで，従来手法の精度を向上させることを目的とする．

2　提案手法

本研究では，他者を誹謗中傷する単語を「バッドワード」として，バッドワードリストを作成する．作成方法では，三宅ら[3]が提案した word2vec による学習を行い，辞書の拡張を行う手法を改良している．そして，他者を誹謗中傷する文章を作成し，リストと文章中の単語のパターンマッチによって内容が誹謗中傷であるか否かを判定する．また，バッドワードリストと対をなす表現として「グッドワードリスト」を作成し，2 つのリストを学習データとして二値分類の学習を行い，二値分類器を作成する．パターンマッチによる判定及び二値分類器による判定によって誹謗中傷であると判定された文章に対して，さらに係り受け解析を用いて，誹謗中傷単語の係り元が固有名詞か代名詞であった場合，文章が誹謗中傷であると判定する．拡張前のバッドワードリストとのパターンマッチによる判定を手法 1，拡張後のリストとのパターンマッチによる判定を手法 2，二値分類器による判定を手法 3 とする．

実験データは，次の 3 種類の文章を用いる．①他者を誹謗中傷している文章(96 文)，②バッドワードリストの単語を使っているが，他者を誹謗中傷していない文章(30 文)③，他者を誹謗中傷していない文章(30 文).

次に評価方法について説明する．手法 1, 2, 3 では，誹謗中傷であると判定してほしい単語（正解単語）を選定（189 語）し，それらが正常に判定されているか否かで評価する．係り受け解析を用いた判定では，文章の誹謗中傷判定がどれだけ正確にできているのかで評価する．

3　実験結果

手法 1, 2, 3 の結果を表 1 に示す．適合率は誹謗中傷であると判定した単語のうち正解単語であるものの割合，再現率は正解単語のうち正しく判定できたものの割合，F 値は適合率と再現率の調和平均である．

表 1：手法 1, 2, 3 の結果

	手法 1	手法 2	手法 3
適合率	1	0.96875	0.60274
再現率	0.148148	0.492063	0.931217
F 値	0.258065	0.652632	0.731809
判定成功	28	93	176

手法 1 では，リストの単語数が少なく，誹謗中傷であると判定してほしい単語が 28 単語しか判定されなかったため，再現率，F 値共に低い結果となった．

手法 2 では，手法 1 よりもリストの単語数が多いため，その分正常に判定される単語も増加し，再現率，F 値共に手法 1 よりも高い結果となった．

手法 3 では，二値分類器の判定によって，より多くの単語を判定することができたため，手法 1, 2 と比較して再現率が高くなった．しかし，誤って誹謗中傷であると判定されてしまった単語が多く存在したため，適合率は手法 1, 2 と比較して低い結果となった．

次に係り受け解析を用いた判定の結果を表 2 に示す．適合率は誹謗中傷であると判定した文章のうち実際に誹謗中傷文である割合，再現率は実際に誹謗中傷文であるもののうち正しく判定できた割合である．

表 2：係り受け解析を用いた判定の結果

	手法 1	手法 2	手法 3
適合率	0.666667	0.705882	0.705882
再現率	0.105263	0.226415	0.382979
F 値	0.181818	0.342857	0.496552

どの手法においても，誤って誹謗中傷であると判定された文章が，係り受け解析を用いた判定によって正常に判定される場合が増加した．その一方で，係り受け解析を用いた判定によって，正常に判定されていた文章が誹謗中傷でないと判定されてしまう場合も増加した．手法 1 では，ほとんどの誹謗中傷文が誹謗中傷ではないと判定されたため，再現率は低くなった．手法 2 も同様に再現率が低かったが，手法 1 よりもリストの単語数が多いため，精度もやや高くなった．手法 3 も同様に精度が高いわけではないが，手法 1, 2 よりも良い結果となった．

4　むすび

従来手法を改良した手法 2, 3 を提案した．また，実験により手法 3 が有効であることを示した．二値分類器を使用することで，パターンマッチによる判定よりも正常に判定できる単語が増加したが，その一方で誹謗中傷でないのに誹謗中傷だと判定されてしまう単語が増えた．これは，リストの作成方法に原因があると考えられる．また，使い方によっては誹謗中傷とはならない単語の判定も問題となった．単語の係り元が固有名詞か代名詞であれば，文脈的にその単語が誹謗中傷ではないとしても誹謗中傷であると判定されてしまうため，そのような単語に対しては何かラベル付けをして，文脈によって判定を変えられるようにすることが今後の課題である．

参考文献

[1] ICT 総研，”2022 年度 SNS 利用動向に関する調査”, https://ictr.co.jp/report/20220517-2.html,（参照 2022-11- 23）.

[2] 総務省：インターネット上の誹謗中傷情報の流通実態に関するアンケート調査結果，プラットフォームサービスに関する研究会（第 36 回），p.7，2022.

[3] 石坂 達也，山本 和英: 2 ちゃんねるを対象とした 悪口表現の抽出, 言語処理学会第 16 回年 次大会, pp.178 – 181, 2010.