誹謗中傷表現辞書・プログラムを利用した インターネットの書き込みにおける誹謗中傷の対策

広瀬研究室 3 年
C1182369 吉野凌太

2020 年 11 月 24 日

概要

インターネット上における発言内容は，発言者が読み手を傷つけるつもりがなかったとしても誹謗中傷にあたることがあり，そのようなコメントによる苦悩から命を絶ってしまう人もいる。そうした被害者を出さないために，人を不快させたり，苦しませたりするような言葉の一覧である誹謗中傷表現辞書をつくり，その中のものに一致したら,「不快にさせるメッセージが含まれています」などと表示し，書き手には書いた内容の意味を確認させ，読み手には自分が不快になる恐れのある書き込み内容を読みたいかを確認させ，精神的な苦痛をなくすシステムを提案する。

1　インターネットでの誹謗中傷

インターネット上では，誰でも簡単に好きな情報を発信したり，得ることができるといったメリットがあるが，デメリットもある。それは，匿名性が高く,「他の人もしているから」という集団心理が働くことで攻撃性が高まるため，他人の気持ちを考えない自分勝手な発言がしやすいことだ。さらに，そのような根拠のないコメントを鵜呑みし，便乗して誹謗中傷をする人もいる。

そのせいで苦しむ被害者が年々増えていることが問題となってる。また，読み手に嫌な思いをさせるつもりがなく，知らない間に加害者になってしまっている人もいる。

誹謗中傷を行ってしまう理由としては，嫉妬やストレスの解消，自分の弱い部分を隠したり，自分の強さ，賢さ（優位性，正当性）を示すため，また，相手がどのように反応するかをみてを楽しむためにしている人もいる。

2　誹謗中傷件数

平成 29 年度の法務省の人権擁護機関の取り組み [1] によると，インターネット上の人権侵害情報に関する事件数が，2,217 件（対前年比 16.1 %増加）で，5 年連続して過去最高件数となっている。

2.1　SNS の例

- Twitter（ツイッター）
- Facebook（フェイスブック）
- LINE（ライン）

2.2　インターネットの書き込みの例

- YouTube のコメント欄
- Yahoo!のコメント欄
- 電子掲示板（例: 5 ちゃんねる）

3　研究内容

まず，SNS や電子掲示板上のテキストを対象とし，関連研究にある SO-PMI という方法を用いて誹謗中傷にあたる言葉・文章である可能性の度合いを表した「誹謗中傷度」を算出し，それに基づいて誹謗中傷の言葉・文章の辞書（以降誹謗中傷表現辞書）を作成する。

そして，誹謗中傷表現辞書に基づいて，誹謗中傷に成り得る言葉が含まれていたら，書き手，読み手にそれぞれ「不快なメッセージが含まれいています」などと警告し，書き手にはどういう意図で書いたのか，読み手にはそれでもどういう内容か知りたいかを確認させるようなシステムを作成する。

3.1 対象とする問題群

- 例えば,「死」という単語だけに反応してしまうと「『死』について考えさせられた」や「あの蜂に刺されると最悪死ぬんだよな」といった感想や意見を誹謗中傷の表現があると誤認する恐れがあるので，単語と単語の有害度・違法度を抽出し，それを使用する。
- 単語と単語を合体させてできた造語であっても，単語ごとに分け，どれかに誹謗中傷に成り得るものがあれば，それを作成するプログラムで判断し，その造語を省く。また，誹謗中傷に成り得るインターネットスラング（以降ネットスラング）も辞書に入れる。
- 運動会の動画に対して「おでん食べたい」といったようなと記事や動画とまったく関係ない発言内容も省けるようにする。

3.2 単語の説明と例

以下は，ネットスラングの説明とその例，研究で収集して対象にしようと考えている情報である。

3.2.1 インターネットスラング

インターネットで使用されるスラング（隠語，略語，俗語）であり，インターネット利用者を中心に通用する，チャットや電子掲示板，あるいは電子メールでユーザー間の交流から生まれた言語表現である。

目で文章や文字を読むインターネットコミュニティの特性から，声に出して読むことはあまり想定されていないものがほとんどある。このような背景から原則として文字言語の一種と解釈できるが，なかには口語として一般に浸透し音声言語も存在する。

3.2.2 造語・ネットスラングの例

- 競馬鹿（「競馬」と「馬鹿」を合わせたもの）
- バ課金（「バカ」と「課金」を合わせたもの）
- タヒね（「死」の代わりに「タヒ」という文字をあてている）
- マジキチ（マジでキチガイじみてるの略，本当に気が狂っているという意味）

3.2.3 収集する情報

- Twitter の投稿内容
- 5 ちゃんねるの書き込み内容

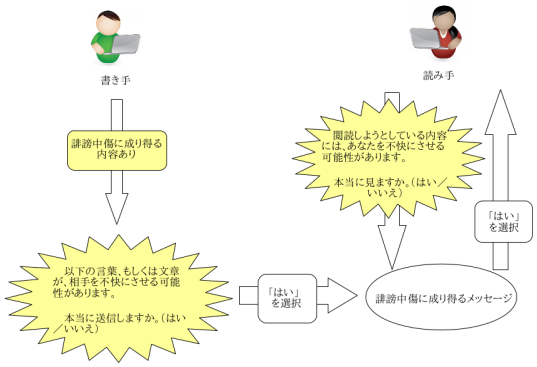

図 1: 提案手法

4 関連研究

- 石坂らの研究 [2] では，電子掲示板サイト”2 ちゃんねる”の書き込みを入力文や学習データに使用し，高いほど悪口単語である可能性が高いという意味のみを持った，単語の「悪口度」を算出して，悪口文/非悪口文の文分類していた。
- 池田らの研究 [3] では，違法性・有害性のある単語だけではなく，文書も抽出していた。キーワードリストを生成した後，例えば,「爆破」と有害性が低いのは「炭鉱」で，有害性が高いのは「学校」ということがわかる。
- 大友らの研究 [4] では，Twitter 上のテキストを検証する情報として集め，TF-IDF という値で単語の重要度を評価し，SO-PMI という方法でいじめの表現と成り得るいじめ度を算出し，それらに基づいていじめの表現と成り得る辞書を作成した。そして，作成したいじめ表現辞書，N グラム，Word2vec，Doc2vec を特徴量として使用し，機械学習手法を選択し，ネットいじめの自動検出を試みていた。

5 目的

電子掲示板や SNS などに導入できるような誰もが気分を害さず，発言ができ，意見をきくことができるシステムを考案，構築する。

6 提案手法の概要

図 1 は，提案手法の概要である。

6.1 誹謗中傷表現辞書の作り方

今回は，Google の検索エンジンで「ネットスラング」と調べ，その中で誹謗中傷に成り得る単語を情報

として CSV ファイルに記録する。

6.2 作成した CSV ファイル

```
"キモオタ","気持ち悪いオタク"
"香ばしい","頭おかしい","痛々しい"
"マジキチ","マジでキチガイじみてるからやめろ，マジでキチガイじみてる"
"池沼","知的障害"
"沼","知的障害"
"ピザ","太っている"
"厨","中毒"
"自宅警備員","ひきこもり，ニート"
"ks","カス"
"LOL","（馬鹿にして）大声を出して笑う"
"lol","（馬鹿にして）大声を出して笑う"
"馬鹿","人を傷つける可能性のある言葉"
"カス","人を傷つける可能性のある言葉"
"死ね","人を傷つける可能性のある言葉"
"クソ","人を傷つける可能性のある言葉"
```

実際に CSV ファイルを作成した。

各行を「[誹謗中傷に成り得る単語],[その意味]」としている。

今後は，これに「組み合わせによって有害性が高くなる・低くなる単語」なども加えていく。例えば「クソ」であれば,「野郎（クソ野郎）」が有害性が高くなる単語で,「可愛い（クソ可愛い）」が有害性が低くなる単語である。また，単語の有害性が高いか低いかは SO-PMI，AIC 等を用いて数値化する。

6.3 SO-PMI

SO-PMI(Se-mantic Orientation Using Pointwise Mutual Informa-tion) は，Wang and Araki が提案した手法で，2 つの基本単語を用意し，対象の単語がその 2 つのどちらと文書内で共起しやすいかを計る。この手法を用いる理由としては，誹謗中傷に成り得る単語同士は同一 Web ページ内において共起しやすいという性質を持っていて，Web 検索ヒット件数を使用するためである。

6.4 AIC

AIC（赤池情報量規準）は，統計モデルの良さを評価するための指標である。モデルの複雑さとデータとの適合度との均衡を保つために使用される。多くの場合，AIC 最小のモデルを選択すれば，良いモデルが選択できる。

6.5 誹謗中傷の確認をするプログラム

プログラミング言語 Ruby の正規表現を使って，ユーザがインターネット上に書き込もうとしている，もしくは SNS に投稿しようとしている文章の中に，作成した CSV ファイルの誹謗中傷表現辞書にある単語が含まれていたら警告文と単語の意味を表示する。

7 まとめと課題

誹謗中傷辞書にあるネットスラングとマッチしたら，その単語の意味，有害性を表示することができた。

しかし，今回はもともとインターネット上にあった誹謗中傷に成り得る単語とその情報を CSV ファイルに入れただけで，単語の有害性や違法性は測らなかったので，先行研究の SO-PMI，AIC などを用いてそれらを測ることや形態素解析を試し，誹謗中傷に成り得る単語や文章を自動的に検出できるようにすることが目標である。

また，ネットスラングの有害性を調べることは，造語であったり，比喩表現であったりするため，先行研究から SO-PMI，AIC では正確に測ることは困難であることがわかっている。そのため，検索エンジンで「ネットスラング」と調べ，インターネットからその情報を拾うしかなく，新しくネットスラング等が生み出される度に，誹謗中傷表現辞書を更新しなければならず，手間がかかるため，克服するための何らかの手段が必要になると考えている。

参考文献

[1] 法務省. 平成２９年における「人権侵犯事件」の状況について（概要）～法務省の人権擁護機関の取組～.

[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, E1-6, pp. 131-134, 2011

[3] 池田和史, 柳原正, 松本一則, 滝嶋康弘. 格要素の抽象化に基づく違法・有害文書検出手法の提案と評価. 情報処理学会第 72 回全国大会, 5D-4, pp. 2-71-2-72, 2010

[4] 大友泰賀, 張建偉, 中島伸介, 李琳. 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. 第 12 回データ工学と情報マネジメントに関するフォーラム（第 18 回日本データベース学会年次大会）, C7-1, p22, 2020