インターネット上の書き込みによる誹謗中傷の対策システムの提案

廣瀬研究室 4 年
C1182369 吉野凌太

概要

インターネット上の掲示板やコメント欄での発言内容は，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，人を不快にさせたり，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」などと表示する。そして，書き手には書いた内容の意味，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。(267 文字)

目次

第 1 章

はじめに

はじめに，本研究にするに至った背景などを以下に述べる。

1.1 背景

インターネットを介して行われる会話では，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である。さらに，そのような根拠のないコメントを鵜呑みにし，便乗して誹謗中傷をする人もいる。

それによって苦しむ被害者が年々増えていることが問題となっている [1]。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。

1.2 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（図 1.1）。

1.3 インターネット上の会話の例

以下は，インターネットを介して会話がなされているアプリケーションや掲示板の例である。

- Twitter（ツイッター）
- Facebook（フェイスブック）
- LINE（ライン）
- YouTube のコメント欄
- 5 ちゃんねる

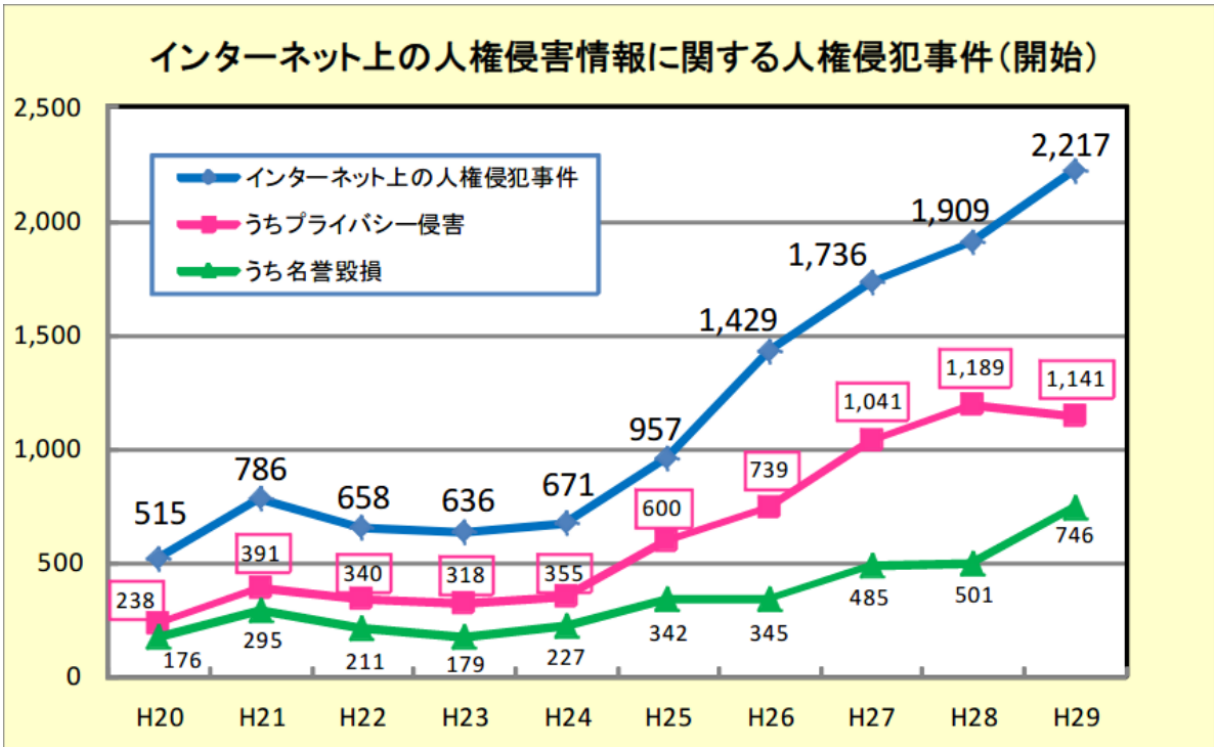

図 1.1　インターネット上の人権侵害情報に関する事件数 [1]

1.4　目的

1.5　目的

特にインターネットをよく使い，その影響を受けやすい人が気分を害さず，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。

そして，最終的には，既存のアプリケーションや電子掲示板などに導入できるようにする。

第 2 章

過去の事例とそれに対する問題提起

以下に，先行研究の重点と課題点を述べる。

2.1 「Web 上の誹謗中傷を表す文の自動検出」・石坂達也，山本和英

- ポイント: 電子掲示板サイト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単語の「悪口度」を算出して，悪口文と非悪口文の分類をしている。
- 課題点: 「競馬鹿」という単語の場合，単語分割をする際に「競馬」と「鹿」という悪口度の低い単語に分割される。全ての造語に対して対応はできていない。

2.2 「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」・池田和史，柳原正，松本一則，滝嶋康弘

- ポイント: キーワードリストを生成した後，例えば「爆破」をキーワードとすると「炭鉱」は有害性が低く「学校」は有害性が高いということを算出していた。
- 課題点: キーワードリストを作成しなければ，参照できない。

第 3 章

提案

このシステムでは誹謗中傷に成り得る単語・文章があったら，書き手だけでなく読み手にも警告文を出すようにする（図 3.1）。

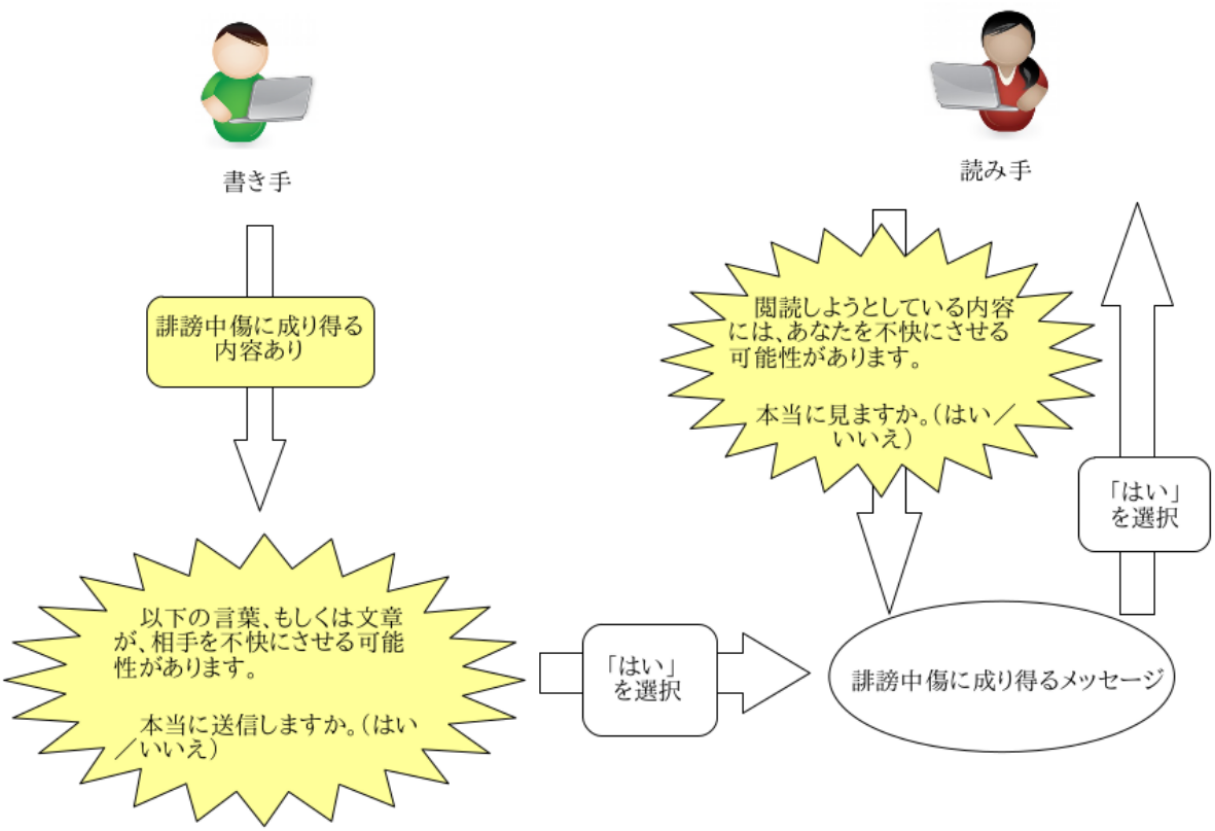

図 3.1　提案手法

第 4 章

システムの設計

以下に，システムの要件を述べる。

- 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。

第 5 章

システムの開発

5.1 開発環境

Ruby
HTML
CSV

5.2 誹謗中傷表現辞書のファイル

誹謗中傷表現辞書は，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 ファイルに手動で記録したものとする。

以下がその CSV ファイル「list.csv」の一部である。

```
word,meaning
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
```

5.3 誹謗中傷表現の有無を確認をするプログラム

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。

以下は，そのプログラム「check.rb」の一部である。

```
#!/usr/bin/env ruby
#-*- coding: utf-8 -*-

require 'csv'

printf("試しに何か入力:\n")
input = STDIN.gets.chomp

list = CSV.table('list.csv', headers: true).map(&:to_h)

check = false

list.each do |value|
  if input =~ /#{value.dig(:word)}/
    check = true
    puts("\n 言葉: #{value.dig(:word)}\n 意味: #{value.dig(:meaning)}")
  end
end

if check == true
  puts("\n 警告！！\n あなたの入力内容には、上記の誹謗中傷表現があります。")
end
if check == false
  puts("\n あなたの入力内容には、誹謗中傷表現はありません。")
end
```

以下は，上記のプログラムの実行結果である。

```
$ ruby check.rb
試しに何か入力:
お前は香ばしいキモオタだな ww

言葉: キモオタ
意味: 気持ち悪いオタク

言葉: 香ばしい
意味: 頭おかしい、痛々しい

警告！！
あなたの入力内容には、上記の誹謗中傷表現があります。
```

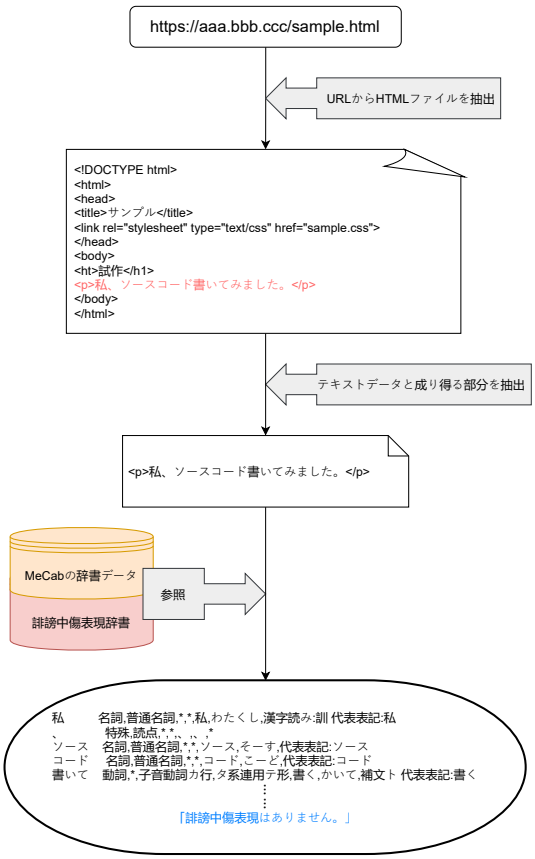

図 5.1　システムの動作概念図

5.4　テキストデータの収集方法

まず，Ruby のパッケージ「open-uri」を用いて，URL から HTML ファイルを取得する。次に，パッケージ「nokogiri」を用い，取得した HTML ファイルから書き込みの要素を抽出し，テキストデータとして収集する。そして，**??**節のプログラムによって，誹謗中傷表現が含まれているかどうかを確認する（図 5.1)。

以下は，その収集を行うプログラムである。標準入力の URL から HTML ファイルの p 要素を抽出している。

```
require 'nokogiri'
require 'open-uri'

url = STDIN.gets.chomp

charset = nil
html = open(url) do |f|
  charset = f.charset
  f.read
end

doc = Nokogiri::HTML.parse(html, nil, charset)
p doc.css('p').text.strip
```

第 6 章

実験

第 7 章

考察

第 8 章

まとめ

最後に，本研究の課題と今後の展望を以下に述べる。

8.1 課題

課題としては，形態素解析をし，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である。

また，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。

8.2 今後の展望

現状から，「お前の知能はサル以下だ」といった比喩表現を用いた文章の検出ができるようにすることが今後の展望である。

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～. http://www.moj.go.jp/content/001253506.pdf.（参照 2020-12-04）.

[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

[5] 荻野敏樹. SNS の歴史. 京都学園大学人間文化学会 学生論文集編集委員会. 2019 年 3 月. 人文学部学生論文集 第 17 号. https://lab.kuas.ac.jp/~jinbungakkai/pdf/2018/p2018_01.pdf.（参照 2020-12-29）.