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、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がいる。
- 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る。
- 誹謗中傷表現辞書の中のものに一致したら、読み手と書き手にそれぞれ警告文を表示する。
- 精神的な苦痛をなくすシステムを提案する。

背景１（メリット･デメリット）

メリット：

- 誰でも簡単に好きな情報を発信できる。
- また、それらを得ることができる。

デメリット：

- 匿名性が高く、集団心理が働き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。
- 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がいる。

背景２（誹謗中傷を行う理由・事件数）

理由：

- 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（優位性や正当性）を示すため。
- 相手がどのように反応するかをみてを楽しむため。

誹謗中傷件数：

- 2217 件で、5 年連続して過去最高件数となっている（平成 29 年度の法務省の人権擁護機関の取り組み）。

研究内容

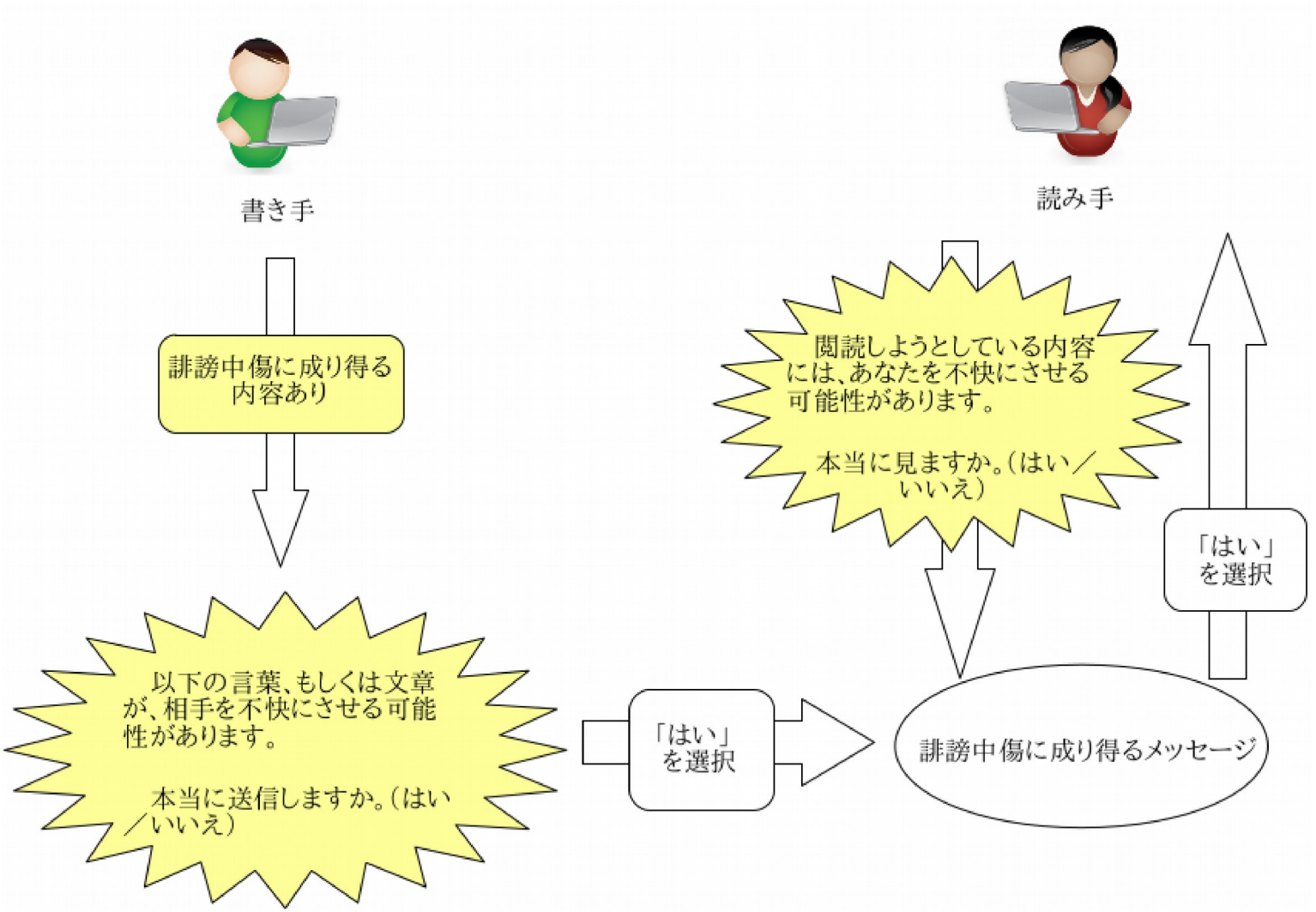
- SNSや電子掲示板上のテキストを対象とし、SO-PMIという方法を用いて誹謗中傷にあたる言葉・文章である可能性の度合いを表した「誹謗中傷度」を算出し、誹謗中傷の言葉・文章の誹謗中傷表現辞書を作成する。

- 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」といった警告をし、書き手にはどういう意図で書いたのか、読み手には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。

目的

- 誰もが気分を害さず、発言でき、意見を聞くことができる。
- 電子掲示板やSNSに導入できる。

提案手法


書き手
読み手
誹謗中傷に成り得る内容あり
閲読しようとしている内容には、あなたを不快にさせる可能性があります。
本当に見ますか。（はい／いいえ）
「はい」を選択
以下の言葉、もしくは文章が、相手を不快にさせる可能性があります。
本当に送信しますか。（はい／いいえ）
「はい」を選択
誹謗中傷に成り得るメッセージ

まとめと今後の課題

- 誹謗中傷表現辞書にあるネットスラングと一致したら、その単語の意味、有害性を表示することはできる。
- 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。
- 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。
- 新しくネットスラングなど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。

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