4ZD-06

個人の考え方に対応したヘイトスピーチ遮断手法の提案

烏丸　正大†　　寺澤　卓也†
東京工科大学　メディア学部†

1.はじめに

日本国内では、国内に居住する外国籍の人らへの差別、宗教の差別や排除の扇動をするような言動（ヘイトスピーチ）が古くから続けられてきた。2019 年 12 月、川崎市で「川崎市差別のない人権尊重のまちづくり条例」[1]が可決、成立し、道路や公園などでのヘイトスピーチに刑事罰が科されることが決定した。しかし、ネット上でのヘイトスピーチは罰則対象から外れており課題が残っている。本研究では、ネット上でのヘイトスピーチ問題の解決策として、ブラウザ上でヘイトスピーチを遮断し、情報の受け手に不快な情報を見せないようにする手法を提案する。この手法は、個人の考え方に合わせ、必要以上な情報遮断を行わないため、情報発信者の表現の自由を侵さない利点を持っている。

2.パーソナルヘイトストップフィルタ(PHSF)

本研究では、個人の考え方に対応した情報遮断を行うブラウザの拡張機能を作成し、評価することを研究のゴールとし、この拡張機能をパーソナルヘイトストップフィルタ(PHSF)と名付ける。

PHSF は、ユーザごとに割り当てた、情報遮断レベルを利用し、個人の考え方に合わせた情報遮断を行う。情報遮断レベルとは、ユーザのヘイトスピーチに対する許容範囲を表した値のことで、情報遮断レベルが低い人ほど、ヘイトスピーチに対する許容範囲が狭い人であることを表している。つまり、情報遮断レベルが低い人ほど、ヘイトスピーチだと感じる情報の量が多いので、PHSF が遮断する情報も多くなる。

PHSF の情報遮断機能には、ヘイトスピーチによく使われる単語をまとめた辞書を利用している。単語には、TF-IDF のアルゴリズムによって計算される、ヘイトレベルが付与されている。ヘイトレベルとは、単語がヘイトスピーチに用いられる可能性を表したもので、ヘイトレベルの高い単語を含む文章ほど、多くの人が不快だと感じるヘイトスピーチであることを表す。

A method to block hate speech contents in the Internet based on each person's philosophy
†Masahiro Karasumaru †Takuya Terasawa
†School of Media Science, Tokyo University of Technology

3.PHSF の実装

PHSF は Google Chrome の拡張機能として実装した。PHSF を初めて利用する際は、初回登録として、情報遮断レベル決定アンケートに回答する。情報遮断レベル決定アンケートは、ブラウザのタスクバーに表示されている PHSF のアイコンをクリックすることで利用できる。全 25 問の文章に対して、文章から受ける不快度を 6 段階で評価することで、情報遮断レベルが決定され、情報遮断機能が利用可能になる。ユーザがニュースサイトのコメント欄を表示する際に PHSF が作動し、ユーザの目に不快な情報が届く前に、ヘイトスピーチをブロックする。図 1 は PHSF がコメントをブロックした際の様子である。

図 1 PHSF 利用時の様子

4.評価実験

PHSF がネット上のヘイトスピーチのような、個人の人権と表現の自由の両方に関わる情報を遮断することに適していることを示す為に評価実験を実施した。

今回の評価実験では、PHSF とすべてのユーザに対して一律の情報遮断を行う、ベイジアンフィルタ[2]を用いたスパム文判定システムの適合率、再現率を比較する。

適合率とはシステムがヘイトスピーチであると予測したデータの内、実際にユーザがヘイトスピーチであると考えているデータの割合、再現率とは、実際にユーザがヘイトスピーチであると考えているデータの内、システムが正しく

予測し、ヘイトスピーチであると予測したものの割合のことで、それぞれ式(1)、式(2)で求められる。式(1)、式(2)の TP はユーザ、システムともにヘイトスピーチであると判定したデータの個数、FP は、ユーザはヘイトスピーチでないと判定し、システムはヘイトスピーチであると判定したデータの個数、FN は、ユーザはヘイトスピーチであると判定し、システムはヘイトスピーチでないと判定したデータの個数を表している。本来であれば、図 2 の上側の手順の様に実際に PHSF を利用した被験者に、ブロック箇所が適切であったかを回答してもらった上で、適合率、再現率を計算することが望ましい。

$$適合率 = TP/TP + FP \quad (1)$$

$$再現率 = TP/TP + FN \quad (2)$$

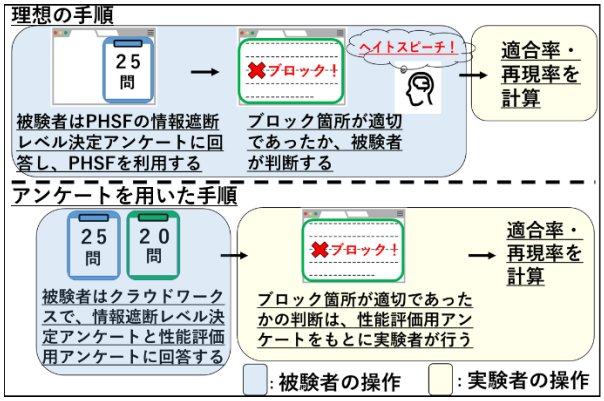

図 2 実験手順

しかし、この方法では、ユーザごとに PHSF を利用する状況を同じにすること、PHSF が不要な情報を遮断していないか、ユーザに判断してもらう際の判断基準を統一することが難しいなどの問題がある。その為、図 2 の下側の手順で、被験者が行う作業の一部を、2 種類のアンケートの回答結果をもとに、実験者が代行する形式で実験を実施した。

アンケートはクラウドソーシングサイト「クラウドワークス」で計 258 名の被験者に対して、情報遮断レベル決定アンケート 25 問と、適合率、再現率を測るテストに用いられる性能評価用アンケート 20 問の 2 種類を実施した。後者のアンケートは、ニュースサイトのコメント欄からランダムに抜き出した文章で、前者のアンケートと同様に文章から受ける不快度を 6 段階で評価する。「非常に不快である」、「不快である」、「やや不快である」と回答した文章をヘイトスピーチであると被験者が考えていると定め、適合率、再現率を計算した。

図 3 から、PHSF とベイジアンフィルタを用いたスパム文判定システムを比較すると、適合率は PHSF の方が高く、再現率はベイジアンフィルタを用いたスパム文判定システムの方が高くなっていることがわかる。式(1)から、適合率が高いシステムであるほど、ユーザがヘイトスピーチではないと考えている文章を誤って遮断する可能性が低いシステムであることが読み取れる。情報発信者の表現の自由を侵さないことを念頭に、ネット上のヘイトスピーチを遮断する場合、ベイジアンフィルタを用いたスパム文判定システムのように、ユーザの考え方を考慮しない、機械的な情報遮断を行うシステムより、PHSF のように、ユーザの考え方に合わせて柔軟な情報遮断を行うシステムの方が適していると言える。

		PHSF	
		ヘイトスピーチである	ヘイトスピーチでない
ユーザの考え	ヘイトスピーチである	TP:465	FN:1249
	ヘイトスピーチでない	FP:525	TN:2921

		ベイジアンフィルタを用いたスパム文判定システム	
		ヘイトスピーチである	ヘイトスピーチでない
ユーザの考え	ヘイトスピーチである	TP:1383	FN:331
	ヘイトスピーチでない	FP:2745	TN:701

	適合率	再現率
PHSF	0.470	0.271
ベイジアンフィルタを用いたスパム文判定システム	0.335	0.807

図 3 実験結果

5.おわりに

本研究では、ネット上でのヘイトスピーチの解決策として、情報遮断システムの提案を行った。評価実験の結果、PHSF はヘイトスピーチのような、人それぞれ許容範囲が異なる情報を遮断するときに効果的なシステムであることが分かった。今回の実装分では、ヘイトスピーチの中でも、限定的な範囲のみを扱ったシステムとなったが、今後は単語辞書の強化、情報遮断を行える Web ページの拡大を行っていく予定である。また、ヘイトスピーチの問題と同様に、セクシャルハラスメントの様な、人それぞれの考え方が異なることによっておこる問題での PHSF の有効性を確かめる実験を予定している。

6.参考文献

[1] 川崎市,川崎市役所ホームページ,"川崎市差別のない人権尊重のまちづくり条例", www.city.kawasaki.jp/250/page/0000113041.html,2019, (参照日 2019 年 12 月 26 日)

[2] Paul Graham, "Better Bayesian Filtering", Proceedings of the 2003 Spam Conference, 2003