

攻撃的コメントの発信抑制に向けたオンラインコメントの受け手の精神ダメージ推定

Mental Damage Estimation of Online Comment Recipients to Suppress Sending Offensive Comment

伊藤 誠基^{1*} 砂山 渡² 服部 峻²
Masaki Ito¹ Wataru Sunayama² Shun Hattori²

¹ 滋賀県立大学大学院 工学研究科

¹ Graduate School of Engineering, The University of Shiga Prefecture

² 滋賀県立大学 工学部

² School of Engineering, The University of Shiga Prefecture

Abstract: 近年、オンラインコミュニケーションが盛んに行われるようになるにつれ、炎上やネットいじめなどの問題が起こるようになってきている。このような問題を避けるためには、コメントの発信者が発信相手の状況を客観的に理解し、発信したコメントによる影響を考えられるようになることが大事と考えられる。そこで本研究では、コメントの受け手の精神ダメージの量を推定し可視化するとともに、そのダメージ量に応じた攻撃的コメントの発信を抑制するメッセージを提示できるシステムを提案する。実験により、システムが推定するダメージ量と、提示するメッセージによる攻撃的コメントの発信抑制効果との関係を検証した。

1 はじめに

近年、オンラインコミュニケーションが盛んに行われるようになるにつれ、炎上やネットいじめなどの問題が起こるようになってきており、特にSNS上での誹謗中傷等の深刻化が問題となっている[1]。インターネット上における誹謗中傷の場面においては、意図的もしくは無意識に攻撃的コメントが発信されることがある。既存のオンラインコミュニケーションツールの多くは、あらかじめ発信することが適切でない単語を用意して、その単語を含むコメントを発信できないようにしている場合が多い。しかし、あらかじめ想定していない単語が含まれるコメントは防ぐことができないため、例えば「死ね」という単語は検出できるが、「氏ね」という単語を検出できない。また、攻撃の意図を持ってしまった時点で、そのような禁止ワードを避けた表現を考えることは難しいため、インターネット上における誹謗中傷問題の根本的な解決につながらない。

このような問題の解決を図るためには、コメントの発信者が発信相手の状況を客観的に理解し、発信したコメントによる影響を考えられるようになることが大事と考えられる。「赤信号みんなで渡れば怖くない」の

ように、冷静に考えると避けるべき事柄であっても、周囲の状況によってはそれが見えなくなることがある。

そこで本研究では、インターネット上のコメントの受け手の精神ダメージの量を推定し可視化するとともに、攻撃的なコメントを発信しようとした際に、相手のダメージ量に応じた攻撃的コメントの発信を抑制するメッセージを提示することで、その攻撃的なコメントの発信を防止するシステムを提案する。

以下、本論文では、2章で関連研究について述べ、3章で攻撃的コメントの発信抑制システムについて述べる。4章で提案システムの効果を検証した実験について述べた上で、5章で本論文を締めくくる。

2 関連研究

2.1 オンライン投稿における攻撃的コメントの検出に関する研究

入力されたコメントに含まれる単語の悪口度を算出して、そのコメントが悪口文であるか自動判別する研究[2]や、コメントに含まれる単語がどれだけいじめと関連するかの程度を数値で表した、いじめ表現辞書を用いてTwitter上のネットいじめを検出する研究[3]がある。

*連絡先：滋賀県立大学大学院工学研究科 電子システム工学専攻
伊藤 誠基
〒522-8533 滋賀県彦根市八坂町 2500
E-mail: ov23mito@ec.usp.ac.jp

これらの研究では、個々のコメントにおいて、人を傷つける単語に重みをつけ、入力されたコメントが人を傷つける程度を判定しているが、それ以前の過去のコメントの影響を考慮していない。本研究では、単一のコメントだけではなく、コメントの時系列データを対象として精神ダメージの推定を行う。

2.2 オンライン投稿におけるコメントの分析に関する研究

Word2Vec や fastText などの単語の分散表現を用いて、ニュースサイトのコメント分析し炎上を検出する研究 [4] や、YouTube のコメントのベクトル化によりコメントを分析して炎上を検出する研究 [5] がある。

これらの研究では、発信された全てのコメントに数値を与え、その値から炎上の程度、すなわちコメントの受け手をどのくらい傷つけるか評価しているが、ネガティブなコメントの量を評価しており、実際に人を傷つけるかどうかまで踏み込んではいない。本研究では、人を傷つける可能性がある単語とその程度を用いて、精神ダメージの推定を行う。

2.3 攻撃的コメントの投稿取り下げに効果的なメッセージに関する研究

ネットいじめに関わる攻撃的コメントの投稿を防ぐために、Web アンケート調査結果を分析し、相手の立場に立ち考えることを促すセルフトークメッセージの効果を検討した研究 [6] がある。この研究では、攻撃的コメントを発信する際に、受け手が攻撃的コメントによって傷つくことを共感させる「共感」、攻撃的コメントを発信しないことのメリットを提示する「メリット」、攻撃的コメントを発信することのデメリットを提示する「デメリット」、攻撃的コメントの発信を禁止する「禁止」の4種類のメッセージによって、攻撃的コメントの発信を抑制するメッセージを提示する。また、車の速度違反抑制に効果的なメッセージについて検討した研究 [7] においては、説得において反発心を喚起させにくい「感謝」のメッセージの効果が高いことが示されている。

本研究においては、これらの5種類のメッセージを発信を抑制するメッセージとして利用することに加えて、これらのメッセージが効果的となるタイミングについて検証を行う。

2.4 本研究における攻撃的コメント

Vangelisti らは、人が傷つく要因とは、自分との関係が軽視されること、屈辱を受けること、言語的な又

は非言語的な攻撃であること、変えることのできない欠陥に関すること、不適切、悪意が感じられるユーモアであること、誤解されること、落胆させられること、ショックを受けさせられることとまとめている [8]。そこで本研究では、非言語的な攻撃を除く、人の傷つく要因を満たす単語を「相手の存在を否定する単語、相手との関係を軽視する単語、物理的に危害を加えることを示唆する単語、人格や容姿を貶す単語、相手を貶める単語」と定義した上で、これらの単語を「攻撃単語」と定義する。

実際的には、単純に攻撃単語を含んでいても、対話相手を攻撃する意図を含まないコメントも存在する。たとえば、軽口が言い合える仲の良い友人同士で、少し失敗した相手に対して、「おまえバカだよな」とコメントしたり、「周りはブスばかりだけど、君は美人だね」のように、相手に対しては攻撃しないが、その他の人を攻撃する場合などが考えられる。

しかし前者については、どこまで相手が許してくれるか不透明な可能性があり、無意識的にでも傷つける可能性がある単語の使用は避けた方が良くと考えられること、また後者については、第三者への攻撃についても、見ていて気分を害する人がいる可能性や、そのような単語に触発された人が、攻撃単語を使ってしまう可能性なども考えられるため、本研究においては、攻撃単語の使用そのものを抑制の対象とする。

3 攻撃的コメントの発信抑制システム

本研究においては、複数のユーザがお互いにオンラインコミュニケーションを取る場面において、各ユーザが攻撃的コメントによってダメージが蓄積されない環境の構築を目指す。本章で述べる攻撃的コメントの発信抑制システムにおいては、一人のユーザが他のユーザに攻撃的コメントの投稿を行う場面に着目し、そのコメントの受け手となるユーザの精神ダメージを推定し、投稿を行おうとしているユーザの攻撃的コメントの発信を抑制する場面を想定している。

3.1 攻撃的コメント発信抑制システムの構成

図1に攻撃的コメントの発信抑制システム¹の構成を示す。

図中の「外部サイト」は、ユーザがオンラインでコミュニケーションを取るための環境を表す。あるユーザがそのサイト内で、別のユーザに対してコメントの

¹本システムの実装に際しては、テキストマイニングのための統合環境 TETDM(Total Environment for Text Data Mining)[11, 12]を利用した。

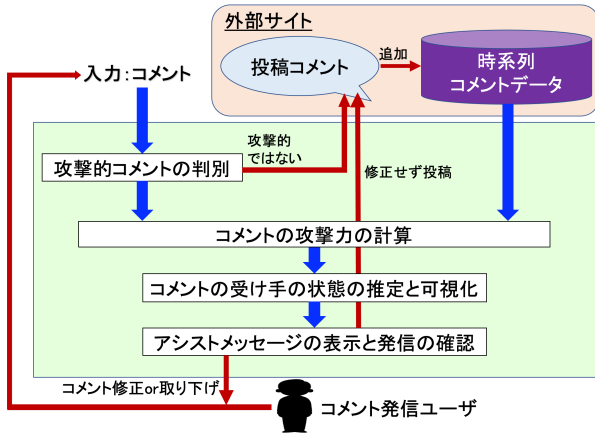


図 1: 攻撃的コメント発信抑制システムの構成

発信を行う場面において、発信しようとしているコメントをシステムへの入力とする。システムは、コメントが攻撃的コメントに該当するかどうかを判定し、攻撃的コメントでなければ、そのコメントは外部サイトに投稿される。攻撃的コメントとして判定された場合は、そのコメントを受信するユーザがそれまでに受けたコメントによる蓄積ダメージを表す状態 (HP: 精神ダメージ許容量) を推定し、コメントの発信を抑制するアシストメッセージを、ユーザにフィードバックしてコメントの修正や取り下げを促す。

3.2 攻撃的コメントの判別

ユーザによって入力されたコメントに、以下で定義するいずれかの攻撃単語が含まれているコメントを「攻撃的コメント」として判定する。攻撃単語の一覧を表1に示す。これらの分類は、2.4節の関連研究に基づいて行い、その攻撃の強さを3段階に分けた。これらに該当する単語は、日本語評価極性辞書 [9] においてネガティブと判定される単語の中から、上記分類に当てはまる単語を手動で選択するとともに、それらの単語の類義語を、類語辞典 [10] を用いて用意した。

3.3 コメントの攻撃力の計算

前節の単語の攻撃力に基づき、コメントの攻撃力の計算を行う。コメントの攻撃力は、あるユーザが受けた時系列コメント全体による精神ダメージの推定量、ならびに投稿されようとしている攻撃的コメントが、これから与える精神ダメージの推定量に用いる。

コメントの攻撃力は、コメントに含まれる攻撃的単語の攻撃力の和で与える。ただし、単語 w の攻撃力 $atk(w)$ は、コメントを受け取る人の直近の 100 コメン

トにおける、単語 w の出現回数を $count(w)$ として、単語の基礎攻撃力 $base_atk(w)$ を用いて式 (1) により与える。これは、同一単語が複数回使われたときには、その精神ダメージとなる感覚量は対数に比例して減少する「ウェーバー・フェヒナーの法則」に基づいている。また式中の、 hp_w は後述のユーザの精神ダメージ許容量 (HP) の最大値に応じた重みを表し、本研究では $hp_w = 14$ に設定している。

$$atk(w) = base_atk(w)(1 - \log_9 count(w)) \times hp_w \quad (1)$$

続いて、コメント c の攻撃力 $attack(c)$ は、コメント c に含まれる攻撃単語 w_i の攻撃力の和と、そのコメントの攻撃意図の程度 $atk_int(c)$ を用いて、式 (2) により与える。

$$attack(c) = \sum_i atk(w_i) \times atk_int(c) \quad (2)$$

ここで、コメント c の攻撃意図の程度 $atk_int(c)$ は、式 (3) で与える値として、0 から 70 の範囲で設定される値となる。

$$atk_int(c) = 0.01 \min\{-0.7atk_dis(c) + 70, 70\} \quad (3)$$

攻撃意図の程度 $atk_int(c)$ は、直前の攻撃的コメントとの間隔に基づいて、攻撃的コメントが連続する状況においては、攻撃的なコメントが発生しやすく、直前の攻撃的コメントからの距離が遠いほど、攻撃的なコメントが発生しにくいという傾向に基づいて定めた値である。この式は、YouTube Data API [13] を用いて収集した攻撃的コメントを多く含む炎上動画と、そうでない動画のコメントを各 4 万コメント (合計 8 万コメント) 収集した上で、それぞれで攻撃的コメントの発生間隔の分布を調べ、炎上動画のコメントを実際の攻撃意図を含むコメント、それ以外の動画のコメントを攻撃意図を含まないコメントとして算出した割合となっている。

たとえば、攻撃間隔 $atk_dis(c)$ が 0 のときの攻撃意図の程度 $atk_int(c)$ は 70%、攻撃間隔 $atk_dis(c)$ が 100 のときの攻撃意図の程度 $atk_int(c)$ は 0% となり、その他の間隔の値は両者を直線的に補間する値となる。

3.4 コメントの受け手の状態の推定と可視化

一連のコメントを受けているユーザの状態を推定する値として、精神ダメージの許容量 (HP) を設定する。HP は 100 を初期値として、攻撃的コメントを受けるたびに、前節で定めたコメントの攻撃力の値が減少し、大きな精神ダメージにより、これ以上コミュニケーション

表 1: 攻撃単語の分類と点数

攻撃単語の分類	単語数	単語例	基礎攻撃力
・相手の存在を否定する単語 ・相手との関係を軽視する単語 ・物理的に危害を加えることを示唆する単語	48	「死ね」,「消えろ」,「殺す」 「邪魔」,「絶交」,「勘当」	3
・相手の人格を貶す単語 ・相手の容姿を貶す単語	306	「クズ」,「ケチ」,「意地汚い」 「ブス」,「デブ」,「ハゲ」	2
・上記以外の相手を貶める単語	303	「馬鹿」,「アホ」,「雑魚」 「ボケ」,「ドジ」,「低脳」	1

表 2: コミュニケーションの状態に応じた 1 コメントあたりの HP の回復量と最大 HP

状態	HP 回復量	最大 HP
攻撃モード	0.1	100
通常モード	0.5	110

ンを継続することができない状態を、HP が 0 の状態として設定する。

計算された HP は、次の 3 段階で色分けして、棒グラフで可視化する。

- 通常状態：HP が 71 以上、緑色で表示
- 注意状態：HP が 31 以上 70 以下の時、黄色で表示
- 危険状態：HP が 30 以下、赤色で表示

また、HP が減少するだけでは、いずれすべてのユーザの HP が 0 になってしまうことと、人は時間経過によってある程度ダメージを回復することができると考えられるため、攻撃的でないコメントが与えられるたびに、表 2 に示す量の HP を回復させる。ただし、精神ダメージの回復量は、攻撃的コメントが発生している状況と、そうでない状況とは異なると考えられるため、攻撃的コメントが発生している状況（攻撃モード）とそれ以外（通常モード）とで、1 コメントあたりの HP 回復量と回復後の最大 HP に差を設ける。

攻撃モードの判定条件は、直近の 100 コメントの中にある、攻撃コメント集合 ATK と攻撃コメントの数 $count_atk$ を用いて、式 (4) の条件を満たすときに攻撃モード、それ以外の時を通常モードとして判定する。すなわち直近の 100 コメントにおいて、式 (3) で計算される攻撃意図の程度が、平均 50 を超えた場合に攻撃モードと判定される。

$$\sum_{i \in ATK} atk_int(c_i) > 50 \times count_atk \quad (4)$$

これらの値は、Youtube の炎上動画のコメントとそうでない動画のコメントを元に、炎上動画では HP が 0 になり、それ以外の動画では HP が 100 付近で安定する値として設定した。

表 3: アシストメッセージとその種類

種類	アシストメッセージ
共感	逆の立場なら、あなたはそのコメントに傷つきませんか？
感謝	いつもありがとうございます。お互いに気持ちの良いやりとりができるようにご協力をお願いいたします
メリット	そのコメントを送らないことが、あなた自身の成長につながりますよ
デメリット	そのコメントを送ると、あなた自身の評価も下がることになりますよ
禁止	そのコメントを送ってはいけません

3.5 アシストメッセージの表示と発信の確認

本研究では、攻撃的コメントを発信しようとしているユーザの意識に変化を与え、攻撃的コメントの発信を抑制するメッセージをアシストメッセージと呼ぶ。

本研究で用意したアシストメッセージとその種類を、表 3 に示す。アシストメッセージには、2.3 節で述べた内容に基づいて攻撃的コメントにより、相手が傷つくことを「共感」させるメッセージ、「感謝」を伝えることで心理的リアクタンスを喚起させないメッセージ、攻撃的コメントを発信しないことによる「メリット」「デメリット」を伝えるメッセージ、および発信を「禁止」するメッセージを用意した。

これらのメッセージを表示するタイミングについて、次章で述べる検証結果に伴って設定する。このアシストメッセージを表示した後に、コメントを発信しようとしているユーザに対して、「本当にこのメッセージを送っても大丈夫ですか？」と確認を行うダイアログを表示させ、「はい」を選択した場合はコメントが投稿される。また、HP が 30 以下の場合には、「禁止」のメッセージを出して、コメントの発信を強制的に防ぐ。

3.6 攻撃的コメント発信抑制インタフェース

図 2 に、攻撃的コメントの発信抑制支援システムの画面を示す。

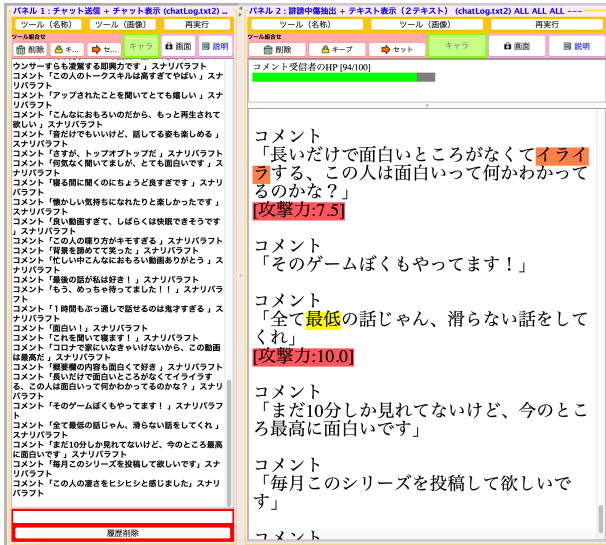


図 2: 攻撃的コメントの発信抑制支援システムの画面

左側のパネルの下部に、コメントを投稿するテキストフォームがある。右側のパネルの上部に、コメントを発信しようとしている相手の HP が表示され、その下に、そのユーザがそれまでに受けたコメントと攻撃単語が可視化して表示される。

攻撃単語は、基礎攻撃力 3 の単語が紫色、基礎攻撃力が 2 の単語がオレンジ色、基礎攻撃力が 1 の単語が黄色で表示される。またコメントの攻撃力も表示される。

4 攻撃的コメントの発信抑制システムの検証

本章では、提案する攻撃的コメントの発信抑制システムが提示する HP の妥当性とアシストメッセージの効果、ならびにアシストメッセージを表示するタイミングを検証するために行った実験について述べる。本検証は以下で述べるアンケート形式で 200 名の被験者に対して行った。被験者は 20 代、30 代、40 代、50 代の男女それぞれを 25 名ずつとした。また本調査は株式会社クロス・マーケティングに依頼して実施した。また、題材としてある動画投稿サイトに投稿された 7 件の動画に与えられた各 100 件のコメントをもとに実験を行ったが、実施に際して動画の投稿者やコメント発信者への影響を考慮して、動画のタイトルやコメントの内容は実験の本質に影響がない範囲で一定の改変を行って実施した。

表 4: 検証課題 1 に用いた動画のタイトルとシステムの推定最終 HP および投稿者のメンタルの強さ

動画のタイトル	HP	メンタル
ひたすら面白い話をする	94	普通
モデルのメイク紹介	78	普通
自分の動画の再生数があまり増えない	67	普通
お店のドライブスルーでドッキリをしにかけてみた	31	強い
いつも勝手に撮ってるけど自分を勝手に撮ってくる人に激怒してみた	22	強い
自身の炎上について歌ってみた	-41	強い

4.1 検証課題

4.1.1 課題 1: HP の妥当性の確認

提案システムが提示する HP の妥当性を確認するための課題を実施した。表 4 に示す 6 つの動画に与えられたコメント 100 件ずつを抽出し、そのコメントを読んでもらい、動画の投稿者がどの程度傷ついているかを、「特にダメージを受けていない」「少しダメージを受けているかもしれないが気にするほどではない」「多少ダメージを受けているがこれ以上耐えられないほどではない」「かなりダメージを受けていてこれ以上の傷つくコメントは見られない」の 4 択で回答してもらった。表中の HP は、各動画コメント 100 件を本システムに入力したときの最終の HP の値を示す。HP が 71 以上の動画を 2 つ、31 から 70 の範囲の動画を 2 つ、30 以下の動画を 2 つ用意して実施した。また動画の投稿者の立場で回答してもらうに際して、人によって精神ダメージへの強さが異なると考えられたため、動画の投稿者のメンタルの強さについて、著者の合議で「強い」と「普通」の切り分けを行った。

4.1.2 課題 2: 精神ダメージ量と効果的なアシストメッセージの種類の関係の確認

「街の女性に声をかけて勝手に容姿を評価してみた」というタイトルの動画に与えられた 100 件のコメントを用いて、被験者が「周りに迷惑をかけるようなアホなことをするな」とコメントを発信しようとしている状況を設定し、100 件のコメントによる精神ダメージと攻撃コメントを提示した上で、コメントの発信を取りやめるメッセージを、表 3 の 5 つと、「メッセージの内容によらず、このコメントを送る」「メッセージの内容によらず、このようなコメントを送る気にはならない」の中から選択して回答してもらった（複数回答可）。本課題は、一定の攻撃的コメントが発生しても止むを得ないと考えられる状況の中で、その攻撃的コメントの発信の取りやめに効果的なメッセージを探る目的で

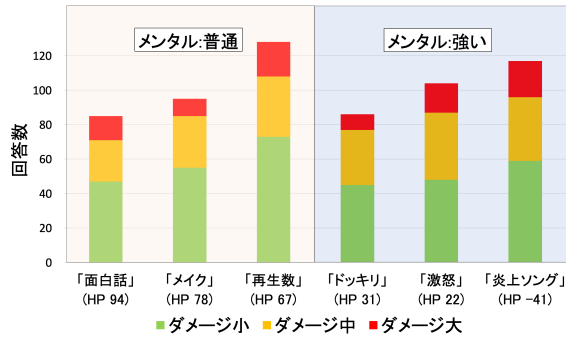


図 3: 各動画コメントに対して投稿者が傷つくと回答した人数

行った。提示する精神ダメージは、提示する攻撃コメントが発生したときに提案システムが算出する HP の減少量を用いて、0 から 100 まで 10 刻みで 11 パターンを用意し、ダメージ量を表す棒グラフとして提示した。一人の被験者にこの 11 パターンをランダムで提示し、各被験者には 11 回回答してもらった。

4.2 検証結果と考察

4.2.1 課題 1 の結果と考察

図 3 に、各動画コメントに対して投稿者が傷つくと回答した人数を示す。左半分のメンタルが普通の投稿者と、右半分のメンタルが強い投稿者のそれぞれにおいて、HP が減少するにつれ傷つくと判断した人の割合が多くなった。このことから、システムが推定する HP に一定の妥当性があることが確認できる。

一方で、前回答者数に対して、ダメージ中（多少ダメージを受けている）やダメージ大（かなりダメージを受けている）と回答した人数の割合は高くなく、また全く傷ついていないと回答した人が約半数（100 人前後）いた。これは、自分ではない他者の精神ダメージ量を過少に見積もってしまいやすいことや、それまでになされている 100 件（多量）のコメントを十分に読まずに勝手に大丈夫だろうと思ってしまうことなどが原因と考えられる。そのため、本システムによる精神ダメージの推定量を表す HP の提示は、攻撃的コメントの発信抑制に向けて、大きな意味を持つ可能性が考えられる。

4.2.2 課題 2 の結果と考察

表 5 に、メッセージの内容によらず、「このコメントを送る」または「このようなコメントを送る気にはならない」と回答した人数とその割合を示す。全体の 50%

表 5: メッセージの内容によらず、「このコメントを送る。」または「このようなコメントを送る気にはならない」と回答した人数と割合

ダメージ	送る	割合 (%)	送らない	割合 (%)
0	6	3.0	114	57.0
10	9	4.5	114	57.0
20	9	4.5	106	53.0
30	8	4.0	102	51.0
40	10	5.0	104	52.0
50	7	3.5	105	52.5
60	6	3.0	107	53.5
70	10	5.0	105	52.5
80	11	5.5	102	51.0
90	9	4.5	106	53.0
100	7	3.5	105	52.5

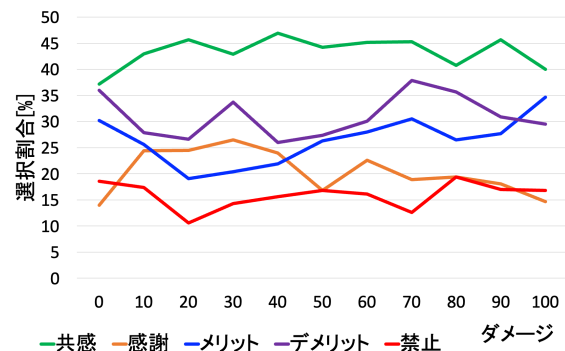


図 4: 攻撃的コメントの発信を取りやめるアシストメッセージの割合（攻撃的コメントの発信の意思がない人を除く）

強の被験者は、「このようなコメントを送る気にはならない」と回答したことから、約半数の人からは攻撃的コメントの発信の意思がないことが伺える反面、残りの半数弱の人は攻撃的コメントを発信してしまう可能性を示唆している。また全体の約 5% の人は、メッセージの内容によらず「このコメントを送る」と回答しており、これらの人に対しては提案システムによるアシストでは対応することができない。本研究における提案システムは、強い攻撃意志を持っている人ではなく、無意識やなんとなくで攻撃的コメントを発信してしまう人を対象としているが、今後このような人の一部にでも、意識の変化を促すことができる方法を模索することは重要と考えている。

図 4 に、表 5 で攻撃的コメントの発信の意思がなかった人について、攻撃的コメントの発信を取りやめるアシストメッセージの割合を示す。全ダメージ量において、相手の立場に立って考える「共感」の効果が最も高く、続いて「デメリット」の効果が高く、「禁止」が

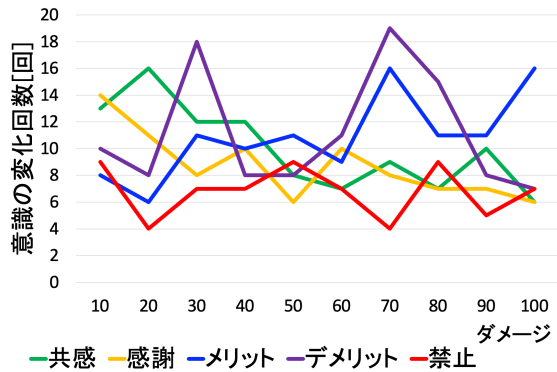


図 5: 攻撃的コメントの発信を取りやめる意識の変化の回数（ダメージ n-10 では攻撃的コメントの発信を行っていたが、アシストメッセージによってダメージ n でその発信を取りやめた人数）

最も効果が低い結果となった。「共感」「デメリット」の効果が高かったのは、間接的または直接的に自身がダメージを受けることを避ける保身的な考えが働いたためと考えられる。これは、「感謝」「メリット」を上回っていることから、日本人特有の自身が得をすることよりも、損をすることを嫌う傾向が反映されたものとも考えられる。また「禁止」の効果が低かったのは、理由もなく一方的に制限されることは、大きな心理的リアクタンスを生むためと考えられる。

図 5 に、各ダメージにおける攻撃的コメントの発信を取りやめる意識の変化があった回数（ダメージ n-10 では攻撃的コメントの発信を行っていたが、アシストメッセージによってダメージ n でその発信を取りやめた人数）を示す。

「共感」「感謝」のグラフは、いずれも右肩下がりのグラフとなっており、ダメージ量が少ない領域において、意識の変化をもたらしやすいことがわかる。これは、まだあまり攻撃的コメントが与えられていない状況においては、周りに流されることなく、相手の立場に立って客観的に物事を捉えたり、感謝の気持ちを受け入れやすかったためと考えられる。

また、「メリット」のグラフは右肩上がりとなっており、ダメージ量が多い領域において、意識の変化をもたらしやすいことがわかる。これは、一定の攻撃的コメントが与えられている状況においては、周りに流されやすい状況が発生しており、自分にメリットがなければ投稿を取りやめるきっかけにはなりにくいことを示唆している。そのため、先の「メッセージの内容によらず、このコメントを送る」と回答した人たちに対しても、より具体的に現実的なメリットを付与すること、たとえば、攻撃的でないコメントを送ることでポイントが付与されるなどの、強いメリットを提示

することが有効な可能性が考えられる。

「デメリット」のグラフは、左右にそれぞれ大きな山があり、デメリットによる意識の変化のタイミングは人によって 2 つのパターンに分かれたことが伺える。デメリットは、リスクとも捉えることができるため、リスクマネジメントにおいて、早い段階からリスクを避けようとする人と、一定のリスクを受容する人とに分かれた可能性が考えられる。

「禁止」は、ダメージ量によらず、意識の変化をもたらす回数は最も低い結果となった。

これらの結果、および図 4 においては全体的に「共感」のアシストメッセージの効果が高かったことから、以下のダメージ量に応じたアシストメッセージの提示方法が提案される。

- ダメージ量が少ない段階においては、まずは「共感」と「感謝」のアシストメッセージを提示し、続いて「デメリット」のメッセージを提示する。
- ダメージ量が中程度の段階においては、「共感」と「デメリット」のアシストメッセージを提示し、続いて「メリット」のメッセージを提示する。
- ダメージ量が多い段階においては、まずは「共感」と「メリット」のアシストメッセージを提示し、続いて「デメリット」のメッセージを提示する。

また、各種類のアシストメッセージとしては、表 3 に示したものの以外にもメッセージ例を用意することで、同一のメッセージによる飽きが来ないように配慮することも必要と考えられる。

5 おわりに

本論文では、インターネット上のコメントの受け手の精神ダメージの量を推定し可視化して、攻撃的なコメントを発信しようとした際に、相手のダメージ量に応じた攻撃的コメントの発信を抑制するメッセージを提示することで、その攻撃的なコメントの発信を防止するシステムを提案した。

検証実験の結果、システムが提示する精神ダメージの許容量（HP）が人間の感覚に近い一定の妥当性を持つことを確認した。また、5 種類のアシストメッセージについて、ダメージ量に応じた攻撃的コメントの発信抑制効果を検証した。その結果、ダメージ量が少ない場合においては、「共感」「感謝」のメッセージ、ダメージ量が多い場合においては「メリット」のメッセージが意識の変化をもたらしやすいことを確認した。

参考文献

- [1] 総務省：総務省におけるインターネット上の誹謗中傷対策の取組について, (URL) <http://urx3.nu/n5vv> (2023 年 2 月 16 日確認)
- [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 発表論文集, pp.131 – 134 (2011)
- [3] 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 (DEIM2020), C7-1 (2020)
- [4] Praboda Rajapaksha, Reza Farahbakhsh, Noel Crespi, Bruno Defude: Uncovering Flaming Events on News Media in Social Media, 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC2019), pp.1 – 8 (2019)
- [5] 堀雄之介, 竹内幹太, 伊東栄典：コメントを利用した炎上動画検出に関する検討, 情報処理学会研究報告, Vol.2021-ICS-203, No.9, pp.1 – 5 (2021)
- [6] 西原陽子, 藤堂悠杜, 山西良典：セルフトークメッセージの呈示による悪口投稿取り下げの効果分析, 電子情報通信学会論文誌 A, Vol.J104-A, No.2, pp.19 – 26 (2021)
- [7] 五十嵐彩那, 臼井伸之介：速度違反抑制に効果的なメッセージと提示タイミング, 交通化学, Vol.46, No.1, pp.13 – 24 (2015)
- [8] Anita L. Vangelisti, Stacy L. Young, Katy E. Carpenter-Theune, and Alicia L. Alexander: Why Does It Hurt?: The Perceived Causes of Hurt Feelings, Communication Research, Vol.32, No.4, pp.443 – 477 (2005)
- [9] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一：意見抽出のための評価表現の収集. 自然言語処理, Vol.12, No.3, pp.203 – 222 (2005)
- [10] 類語辞典 weblio：(URL) <https://thesaurus.weblio.jp> (2023 年 2 月 16 日確認)
- [11] Total Environment for Text Data Mining (テキストデータマイニングのための統合環境)：(URL) <https://tetdm.jp> (2023 年 2 月 16 日確認)
- [12] 砂山渡, 高間康史, 徳永秀和, 串間宗夫, 西村和則, 松下光範, 北村侑也, 統合環境 TETDM を用いた社会実践, 人工知能学会論文誌, Vol.32, No.1, pp.NFC-A.1 – 12 (2017).
- [13] YouTube Data API: (URL) <https://developers.google.com/youtube/v3> (2023 年 2 月 16 日確認)