人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第23回)
SIG-AM-23-04

SNS上の悪口を含む投稿に対する取り下げを促すフィードバック文の自動生成方法の検討

A Study on Automatic Feedback Massage Generation for Asking to Withdraw SNS Message Including Toxic Expression on SNS

藤堂悠杜[1*]　西原陽子[1]　山西良典[1]
TODO, Yuto[1]　NISHIHARA, Yoko[1]　YAMANISHI, Ryosuke[1]

[1] 立命館大学 情報理工学部
[1]College of Information Science and Engineering, Ristumeikan University

Abstract: In this paper, we conducted experiments to obtain the most effective feedback sentence to ask for withdrawing SNS messages including toxic expressions. In our study, Cognitive Response Model and use self-talk feedback to SNS users. We prepared four types of self-talk feedback sentence and experimented the four with questionnaires on the Web. We set four types of condition in obtaining a feedback sentence: whether or not a SNS message includes explicit toxic expression, and whether or not the feedback sentence points out the toxic expression. Experimental results showed that the most effective feedback was a sentence that presented a sender to a receptor of toxic expression. We also found that the self-talk feedback was more effective when a SNS message included an explicit toxic expression and the feedback pointed out the toxic expression.

1　はじめに

近年のインターネットの普及やスマートフォンの発達などの影響より，小・中学性からインターネットに触れる機会が増加している．総務省の調べ [1] によると，2011 年では 13 歳から 19 歳におけるスマートフォン所有率は 14.6 パーセントだったのに対し，2016 年には 81.4 パーセントにまで上昇している．こうしたネット社会の普及によって，利便性が増す一方で，インターネットの危険な側面も広がってしまっている．その一例がインターネットでのいじめであり，新聞記事などでもその被害が報告されている．

インターネット上でのいじめについての研究，それを防ぐ研究は数多く行われている [2, 3, 4, 5]．インターネット上でのいじめにおいては，被害者にとって有害な情報を投稿することにより行われることがある [6]．インターネット上では匿名で発言できることや，相手の顔が見えないことから，攻撃的な投稿が生まれやすくなってしまう，いわゆるフレーミング現象が起こりやすくなっている [7]．有害表現を含む投稿をフィルタリングする技術が研究されているが [8]，有害表現を含む投稿がなされてしまうことは防ぐことは難しい．2019 年 6 月には Twitter において，有害な表現を含むリプライが送られると，リプライを受けたユーザとそのフォロワーを含む全員からリプライが隠される機能が実装された [9]．これにより，有害表現を含むリプライは投稿者以外からは見ることができない状態になった．さらに，Instagram でも攻撃的な内容を含む投稿を検知すれば，“Are you sure you want to post this?” と，ユーザを諭す機能を実装した [10]．有害表現を含む投稿が，被害者やその周りの人の目に触れないようにする取り組みが進められている．

ただ一方で，有害表現を含む投稿を一方的に規制してしまうと，より一層攻撃的な投稿をするユーザの気持ちを刺激してしまい，周りから見えないからこそ言いたい放題な投稿をしてしまう可能性も考えられる．それではインターネットを利用する情報モラルの観点で考えると根本的解決に至っていない．ネットいじめに対して情報モラル教育の拡充の必要性も示されており [11]，ユーザ自身の判断で，投稿を取り下げさせることも必要と考えられる．さらに，Instagram などの手法では，示されるフィードバック文が常に同じものである．投稿内容に応じてフィードバック文を生成し，示すことにより，有害な表現を含む投稿の取り下げに対して効果が高まると考えられる．

本研究では SNS 上の有害表現を含む投稿に対し取り下げを促す方法を提案することを目的とする．このために，本稿においては，有害表現のうち他者への悪口

*連絡先：立命館大学情報理工学部
滋賀県草津市野路東 1-1-1
E-mail: nisihara@fc.ritsumei.ac.jp

を扱い，悪口を含む投稿の取り下げを促すことに有効なフィードバック文について実験により明らかにする．フィードバック文は単に投稿を禁止するものではなく，Greenwald の「説得の認知反応プロセス」[12] を参考にし，投稿者にセルフトークを起こさせるものとする．本研究が想定する投稿者は，他者に対して攻撃的な投稿（脅迫のような違法な情報 [13]）をしようとしている人ではなく，他者からの指摘を受けると行動を省みる人である．これにより，不用意に他者を傷つける発言をしてしまう投稿者自身も保護したい．

2 関連研究

はじめに，態度変容についての関連研究を紹介し，本研究の位置付けを行う．Greenwald は「説得の認知反応プロセス」において，人はメッセージを受け取ることで態度変容を起こすのではなく，メッセージを受け取った後に行うセルフトークにより態度変容を起こすとしている [12]．セルフトークとは自己会話であり，自分自身に対して話しかけることである．既存の SNS サイトでは投稿の確認をする際に，セルフトーク形式のフィードバック文を提示するものは少ない．本研究では，セルフトーク形式のフィードバック文を採用し，最も効果の高い文面について明らかにする．

五十嵐らの研究においては，ドライビングシュミレータを用いて速度違反を抑止する効果的なメッセージについて検証が行われた [14]．速度違反や遵守した際にメッセージが提示される場合と，速度にかかわらず，常時メッセージが提示される 2 パターンの提示方法を用いられた．メッセージの文面としては禁止型と感謝型の 2 パターンが用いられた．検証した結果，禁止型および感謝型のいずれにおいても，速度抑止の効果に違いは見られないが，禁止型よりも感謝型のメッセージの方が，時間が経過しても効果が持続するということがわかった．禁止型だと心理的リアクタンス，すなわち制限されると反発したくなる気持ちが働いたことが影響したためと考えられる．清の研究 [15] においても，医師が患者の行動を制する場合には,「～ないでください」と直接的な禁止表現ではなく,「～はやめておきましょうか」と共感性の高い禁止表現を用いることが多いことが確認されている．これらのことから，文面によっても，態度変容への影響が変わると考えられる．本研究ではセルフトーク形式の文面を，他者への自己投影型と未来予想を提示する型の 2 つに分け，その効果を検証する．

3 取り下げに有効なフィードバック文の調査実験

取り下げに有効なフィードバック文の調査を行うべく，被験者実験を行なった．実験では，被験者を募り，アンケートに回答することを依頼した．アンケート結果を分析し，取り下げに有効なフィードバック文を調査した．

3.1 実験目的

本実験では，以下の 3 点を明らかにすることを目的とする．

1. 悪口の具体的な指摘は取り下げに寄与するか？
2. 悪口の種類は取り下げに寄与するか？
3. セルフトークの種類は取り下げに寄与するか？

1 つ目で明らかにしたいことは，SNS への投稿文の中に悪口が含まれるときに，その悪口の単語またはフレーズを指摘することで取り下げ率が上がるかどうかである．情報システムにより提示されるメッセージは無視されがちであるが，悪口の単語を具体的に指摘することで，無視されることが減り，結果的に取り下げ率が上がる可能性がある．

2 つ目で明らかにしたいことは，SNS への投稿文の中に含まれる悪口が露骨な悪口の場合に，取り下げ率が上がるかどうかである．露骨ではない悪口（隠語など）の場合は意図して書いていると考えられるが，露骨な悪口は思いもよらず書いてしまうこともあり，指摘をすると書いてしまったことに気づけ，結果的に取り下げ率が上がる可能性がある．

3 つ目で明らかにしたいことは，セルフトークを起こさせるメッセージの形式が，他者へ自己を投影する形式か，未来の利益を提示する形式かで，取り下げ率が上がるかどうかである．他者へ自己投影することにより，他者への共感が生まれ，悪口投稿を自分が受け取った場合を考える，つまりセルフトークを起こすことができ，結果的に取り下げ率が上がる可能性がある．一方で，未来の利益を提示する場合は，自己の利益を受け取った場合を考える．こちらもセルフトークを起こすことができ，結果的に取り下げ率が上がる可能性がある．他者への共感は生まれにくいため，セルフトークは起こりにくいと考えられる．

以上の 3 つのことを明らかにすることを，本実験の目的とする．

表 1: セクションごとの条件設定

セクション	セルフトーク	露骨な悪口	悪口の指摘
1	○	○	○
2	○	○	×
3	○	×	○
4	○	×	×

3.2 実験手順

実験は次の手順で行なった．初めに実験者（第一著者）が被験者を募集する．被験者が SNS へ悪口を含む投稿を試みていると仮定し，被験者には悪口の投稿文と投稿文の取り下げを促すフィードバック文が示される．被験者は投稿文とフィードバック文を見て，投稿文を取り下げるかどうかを回答する．

投稿文の取り下げについての質問は，悪口の指摘の有無と，露骨な悪口か否かで条件を変え，４回のセクションに分けて行われた．表 1 に各回のセクションでの条件設定を示す．全てのセクションにセルフトーク形式のフィードバック文が提示される部分は共通しているが，悪口の指摘の有無と露骨な悪口か否かの条件がそれぞれ異なっていた．セクション 1 からセクション 4 まで，被験者はそれぞれ異なる順番で回答をした．つまり，ある人はセクション 1,2,3,4 とたどり，別の人は 1,3,4,2 と辿るなど，被験者ごとにたどるセクションの順番が異なっていた．

3.2.1 悪口の投稿文とフィードバック文の提示方法

SNS でグループチャットを行っているときに，被験者自身が他のメンバーの 1 人に向けて悪口の投稿を試みているという状況を仮定した．フィードバック文が提示された時に，投稿を取り下げるかどうかを回答してもらった．

グループチャットの人数は 3 名とした．グループチャットの投稿文は，チャットにおけるネットいじめが体験できる Web サイトにあるログを参考に作成した[1]．図 1 に，実験で用いたグループチャットの例を示す．図は，Web 上でチャット画面を作成することができるサイトを用いて作成した.[2] 右側から出ている吹き出しが被験者の投稿文，ユーザ A が一緒に悪口を投稿しており，ユーザ B が悪口を言われている人となる．

悪口の投稿文は露骨な悪口を含むものと，隠語の悪口を含むものを用意した．露骨な悪口を含む投稿文は図 1 の最後に,「今まで黙ってたけどマジでお前馬鹿すぎてもう無理．死ねよ.」で,「馬鹿」「死ね」と露骨な悪口の単語を含んでいた．こちらの投稿文も同じサイト

[1]http://linelog.jpn.org/（2019 年 10 月 31 日確認）
[2]https://sp.mojimaru.com/ (2019 年 11 月 2 日確認)

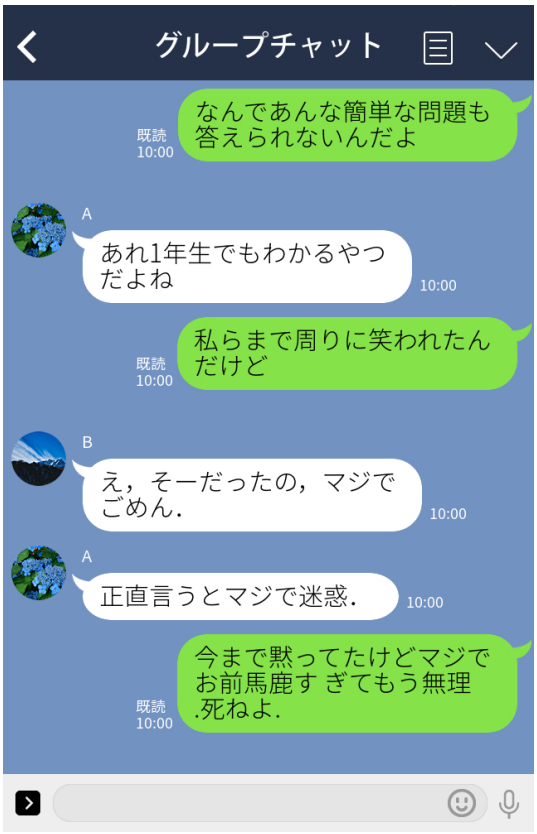

図 1: 実験で用いたグループチャットの例．

のログを参考に作成をした．隠語の悪口を含む投稿文は図 2 の最後に,「この集合写真の A さん，目が半開きで笑う（笑）」で，露骨な悪口の単語やフレーズは含まれないが，文全体として A さんへの悪口になっていた．

提示するフィードバック文は 1 つは他者への自己投影形式であり，もう 1 つは未来の利益を提示する形式である．これらはさらに 2 つずつ分け，合計 4 種類のフィードバック文を用意した．表 2 に実験で用いたフィードバック文を示す．

- 他者へ自己投影する形式：
  - **1** 悪口の対象を投稿者に置換したフィードバック
  - **2** 周りや相手への影響を示唆するフィードバック
- 未来の予想を提示する形式：
  - **3** 投稿者の未来を予想するフィードバック
  - **4** グループチャットの未来を予想するフィードバック

フィードバック文を見て取り下げるかどうかについて，５段階の選択肢を用意し，選択肢の中から 1 つ選ぶことで回答をしてもらった．５段階の選択肢は，投

図 2: 実験で用いたグループチャットの例.

稿を取り下げたいと「強く思う：+2」「少し思う：+1」「どちらとも言えない：0」「あまり思わない：-1」「全く思わない：-2」とした.

3.2.2 被験者の情報

本実験の被験者は情報理工学部に所属する 20 代の男女 16 名で，内訳は男性が 11 名，女性が 5 名であった．共感的関心についての質問，およびセクション 1 から 4 までの質問に回答し終えた後，被験者から実験の感想をもらった.

表 2: 実験で用いたフィードバック文

分類	フィードバック文
1（自己投影）	「もしその投稿をあなたが受けたら，あなたは不快に思いませんか？」
2（自己投影）	「あなたの投稿は，トークルームの人や，相手を不快にしてしまいませんか？」
3（未来提示）	「もし投稿を取り下げれば，周囲や相手を傷つけずに済みますよ?」
4（未来提示）	「あなたの投稿の後，その先のトークはどう進んでいくと思いますか？」

4 実験結果と考察

表 3 に，アンケートの調整を行なった後の評価値の平均を取ったものを示す．各行，被験者 16 名分の評価値の平均となっている．評価値が高いほど，取り下げに有効であったことを示している．評価値が高かったものは，セクション 1 におけるフィードバック 1，つまり露骨な悪口が投稿文に含まれ，その悪口を具体的に指摘し,「もしその投稿をあなたが受けたら，あなたは不快に思いませんか？」とフィードバック文を提示した時であった．続いて評価値が高かったものは，セクション 2 におけるフィードバック 2，つまり露骨な悪口が投稿文に含まれるが，その悪口の指摘はせず,「あなたの投稿は，トークルームの人や，相手を不快にしてしまいませんか？」とフィードバック文を提示した時であった．フィードバック 2 はセクション 1 においても高い評価値を示した．フィードバック 1 と 2 は，セクション 3 を除き，正の評価値が得られており，取り下げに有効な可能性があるとわかった.

反対に評価値が低かったものは，セクション 4 におけるフィードバック 4，つまり露骨な悪口も，悪口の指摘もなく,「あなたの投稿の後，その先のトークはどう進んでいくと思いますか?」であった．フィードバック 4 はセクション 3 においても低い評価値を示した．続いて評価値が低かったものは，セクション 4 におけるフィードバック 3，つまり露骨な悪口ではないが，悪口を指摘し,「もし投稿を取り下げれば，周囲や相手を傷つけずに済みますよ？」であった．フィードバック 3 と 4 は全てのセクションにおいて負の評価値が得られており，取り下げに有効ではない可能性が高いとわかった.

4.1 悪口の具体的な指摘に対する取り下げの効果

実験目的の 1 に対応し，悪口の具体的な指摘に対する取り下げの効果について検証する．セクション 1 と 2 は，露骨な悪口を含む投稿文であることは共通していて，悪口の指摘の有無が異なっている．セクション 1 の評価値の平均は 0.187 であり，セクション 2 の評価値の平均は 0.047 であった．指摘をする方が評価値の平均は高くなった．また，セクション 3 とセクション 4 は，露骨ではない悪口を含む投稿文であることは共通していて，悪口の指摘の有無が異なっている．セクション 3 の評価値の平均は-0.422 であり，セクション 4 の評価値の平均は-0.250 であった．指摘をしない方が評価値の平均は高くなった．このことから悪口に相当する単語を具体的に指摘することで，投稿の取り下げに効果があるが，露骨な悪口に限定される可能性が明

表 3: 各設問におけるアンケートの評価値

セクション	フィードバック	評価値
1	1	**0.812**
	2	**0.375**
	3	-0.437
	4	0
	平均	0.187
2	1	0.312
	2	**0.500**
	3	-0.375
	4	-0.250
	平均	0.047
3	1	-0.250
	2	-0.187
	3	-0.562
	4	**-0.687**
	平均	-0.422
4	1	0.250
	2	0.312
	3	**-0.687**
	4	**-0.875**
	平均	-0.250

らかになった．被験者からの感想でも,「有害ワードが具体的に示されるほうが，なにがいけなかったのかがわかりやすいため.」や,「具体的な忠告によって，一旦メッセージを全て読み，考えることができる．ありきたりの文であると．すぐにその警告文を無視してしまう.」とあった．

4.2 悪口の種類に対する取り下げの効果

実験目的の 2 に対応し，悪口の種類に対する取り下げの効果について検証する．セクション 1 と 3 は悪口の指摘は共通しているが，悪口の種類が異なっている．セクション 1 の平均値は 0.187 であり，セクション 3 は-0.422 であった．露骨な悪口の方が評価値が高くなった．セクション 2 と 4 は悪口の指摘をしない点で共通しているが，悪口の種類が異なっている．セクション 2 の平均値は 0.047，セクション 4 の平均値は-0.250 であった．こちらも露骨な悪口の方が評価値が高くなった．このことから，露骨な悪口の方がフィードバック文を提示することにより，取り下げられる可能性が高いことがわかった．自由回答からは，悪口かそうでないか微妙なものに対してフィードバックを行っても，ユーザ自身にそもそも悪い投稿を行っているという自覚がない場合には,「何が投稿してはいけなくて注意されているのかわからない」という意見や,「機械の誤認識も考えられるから効果が無い」という意見が見られた．

表 4: セクション 1 で得られた評価値に対し，平均値の差の検定を行い得られた p 値．bonferroni 法による補正後値

	FB1	FB2	FB3
FB2	0.8990	-	-
FB3	**0.0017**	0.1629	-
FB4	0.1933	1.0000	0.1770

4.3 セルフトークの種類に対する取り下げの効果

実験目的の 3 に対応し，セルフトークの種類に対する取り下げの効果について検証する．ここまでの実験結果に対する考察において，露骨な悪口に対し，具体的な指摘をすることで，フィードバック文を示すと，取り下げに対し効果があることがわかった．そこで，本節ではこの条件に合致するセクション 1 で得られた評価値に対し，4 種類のセルフトークの中で取り下げに効果があるものについて考察をする．

セクション 1 で得られた調整後の評価値に対し，フィードバック文のペアごとに平均値の差の検定を行った．多重比較となるため bonferroni 補正を行なった．検定の結果を表 4 に示す．有意差が見られたのは，フィードバック文 1 と 3 の間であった（$p=0.0017 < 0.05$）．この結果からフィードバック文 3 と比較すると，フィードバック文 1 は取り下げに対し，効果が高いことが明らかになった．フィードバック文 1 は自己投影型で，フィードバック文 3 は未来提示型であった．悪口を受ける対象を自己に置き換えることで，相手の気持ちに共感することができたため，取り下げに効果があったと考えられる．

5 おわりに

本稿では，SNS 上における悪口を含む投稿に対し，取り下げを促すフィードバック文を自動生成するために，取り下げ効果の高いフィードバック文を実験を用いて明らかにした．フィードバック文は投稿を単に禁止する文面ではなく，Greenwald の「説得の認知反応プロセス」で挙げられているセルフトークを起こさせる文面とした．効果の異なる文面を 4 パターン用意し，最も取り下げの効果が高いものを実験により明らかにした．実験の結果，他者への自己投影を起こす形式の文面の取り下げに対する効果が高いことがわかった．また，実

験において，投稿文に含まれる悪口の種類が取り下げに寄与するか，フィードバック文内で悪口の具体的な指摘が取り下げに寄与するかについても調べた．結果として，投稿文に露骨な悪口が含まれた時に，フィードバック文を提示すると取り下げ率が高くなることがわかった．さらに，フィードバック文で悪口を具体的に指摘することでも取り下げ率が高くなることがわかった．一方で，この結果は大学生から得られた結果であり，小中学生など，発達段階が異なる層に対しては異なる結果が出る可能性がある．今後，同様の実験を行い，発達段階の違いを考慮した上で，取り下げ効果の高いフィードバック文について明らかにする．

謝辞

本研究の一部は，科研費（17K13254），安心ネットづくり促進協議会の助成を受けて行われました．記して謝意を申し上げます．

参考文献

[1] 情報通信白書平成 30 年版. 総務省, 2018.

[2] 笹川喬介, 和泉順子. 誹謗中傷問題のインターネットによる影響に関する考察. 情報処理学会研究報告（グループウェアとネットワークサービス）, 第 2013-GN-89 巻, pp. 1–6, 2013.

[3] 藤桂, 吉田富二雄. ネットいじめ被害者における相談行動の抑制. 教育心理学研究, Vol. 62, No. 1, pp. 50–63, 2014.

[4] 内海しょか. 中学生のネットいじめ, いじめられ体験 —親の統制に対する子どもの認知, および関係性攻撃との関連—. 教育心理学研究, Vol. 58, No. 1, pp. 12–22, 2010.

[5] Michele L Ybarra and Kimberly J Mitchell. Youth engaging in online harassment: associations with caregiver-child relationships, internet use, and personal characteristics. *Journal of adolescence*, Vol. 27, No. 3, pp. 319–336, 2004.

[6] 田代光輝, 服部哲. 情報倫理 ネットの炎上予防と対策. 共立出版, 2013.

[7] 加藤由樹, 加藤尚吾, 杉村和枝, 赤堀侃司. テキストコミュニケーションにおける受信者の感情面に及ぼす感情特性の影響：電子メールを用いた実験による検討. 日本教育工学会論文誌, Vol. 31, No. 4, pp. 403–414, 2008.

[8] Ryuichi Omi, Yoko Nishihara, and Ryosuke Yamanishi. Extraction of paraphrases using time series deep learning method. In *International MultiConference of Engineers and Computer Scientists 2019*, pp. 276–278, 2019.

[9] *Giving you more control over your conversations.* https://blog.twitter.com/enus/topics/product/2019/morecontrolofconversation.html, 2019.

[10] *Our Commitment to Lead the Fight Against Online Bullying.* https://instagram-press.com/blog/2019/07/08/our-commitment-to-lead-the-fight-against-online-bullying/, 2019.

[11] 三島浩路, 本庄勝. 技術的観点からのネットいじめ対策. 電子情報通信学会　通信ソサイエティマガジン, Vol. 9, No. 2, pp. 102–109, 2015.

[12] Anthony G. Greenwald. *Cognitive Learning, Cognitive Response to Persuasion, and Attitude Change*, chapter 6, pp. 147–169. Academic Press Inc., 1968.

[13] 小向太郎. インターネット上の青少年犯罪被害対策の動向. 情報処理学会研究報告（グループウェアとネットワークサービス）, 第 81 巻, pp. 1–6, 2011.

[14] 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. 交通科学, Vol. 46, No. 1, pp. 13–20, 2015.

[15] ルミ清. 禁止の場面における現実の言語表現：医師と美術館員の場合. 世界の日本語教育：日本語教育論集, Vol. 16, pp. 107–123, 2006.