SNS 上の誹謗中傷投稿への同調に関する分析
～ナッジメッセージによる誹謗中傷抑制の試み～*

秋田航汰[a]　小野祐紀[b]　河合求真[c]　神橋龍ノ介[d]　村田拡介[e]　森実輝[f]

要約

SNS 上の誹謗中傷は深刻な社会問題であり，法律や情報技術などの分野で対策が取られているが，費用や権利の側面からの制約が存在し，必ずしも対策が万全であるとは言えない．その点において，ナッジは金銭的負担を用いず，選択の自由も確保できるため，誹謗中傷の対策として有用である．そこで，誹謗中傷への同調を抑制するナッジメッセージの効果検証を行った．利己性，利他性，社会規範の３つのメッセージを作成し，オンラインアンケートでの RCT を実施して検証したところ，利己性メッセージと利他性メッセージに有意な抑制効果が見られた．ただし，これらメッセージは誹謗中傷にあたらない単なる意見の対立をも抑制している可能性があるほか，抑制効果は投稿内容により変動することが示唆される．また，有意な効果が見られなかった社会規範メッセージは，多くの回答者にとって社会規範を意識させるのに十分な情報ではなかったことが考えられる．

JEL 分類番号：D91，Z13
キーワード：ナッジ，ソーシャルネットワーク，損失回避，利他性，社会的規範

---

* 本研究は，大阪大学大学院経済学研究科の倫理委員会の承認を得て行われている．（承認番号：R51108-1）なお，本論文に関して，開示すべき利益相反関連事項はない．
[a] 大阪大学経済学部 u070147d@ecs.osaka-u.ac.jp
[b] 大阪大学経済学部 u759787a@ecs.osaka-u.ac.jp
[c] 大阪大学経済学部 u856501h@ecs.osaka-u.ac.jp
[d] 大阪大学経済学部 u014969c@ecs.osaka-u.ac.jp
[e] 大阪大学経済学部 u238898d@ecs.osaka-u.ac.jp
[f] 大阪大学経済学部 u845230k@ecs.osaka-u.ac.jp

1. はじめに

昨今のオンラインコミュニケーションの急速な普及に伴い，ソーシャルネットワーキングサービス（以下，SNS）上の誹謗中傷がその対象に引き起こす精神的苦痛は広く問題視されるようになっている．現状を受けて，様々な分野において対策が講じられている．例えば，法的な対策としてのプロバイダ責任制限法改正（小向，2021）や，情報技術による対策としての攻撃的文章訂正システムの構築（吉田・松本・吉田・北，2022）があげられる．しかし，これら対策には課題も存在する[1]．

これらの各対策とその課題を踏まえると，行動経済学の知見を活かしたナッジは誹謗中傷問題の対策として有用であると考えられる．行動変容によって誹謗中傷が抑制されれば，情報開示請求訴訟に係る費用の改善が可能である上，選択の自由を確保しているのだから，言論の自由等にも抵触せずに対策ができると考えられる．

そこで本研究では，ナッジを活用したメッセージが誹謗中傷投稿への同調行動に及ぼす効果を検証する．本研究での同調行動の定義は，SNS 上での「いいね」「再投稿」を行うことである．また，メッセージの効果検証とともに，誹謗中傷への同調に影響を及ぼす属性についても分析し，考察を加える．

2. 検証する仮説とメッセージ

2.1. 仮説

ナッジを活用したメッセージによって，以下の３つの仮説を検証する．

【仮説１】（利己性）誹謗中傷により自己が被る損失を強調するメッセージは，誹謗中傷投稿への同調を抑制する．

【仮説２】（利他性）誹謗中傷により他者が被る損失を強調するメッセージは，誹謗中傷投稿への同調を抑制する．

【仮説３】（社会規範）誹謗中傷を行う者が少数であることを強調するメッセージは，誹謗中傷投稿への同調を抑制する．

2.2. メッセージ

以上の仮説から，それぞれメッセージを作成した．メッセージは以下の通りである．

---

[1] 改正プロバイダ責任制限法については，情報開示請求は海外の事業者に対して行うことが多いことから，本当に時間やコストの削減になるのか懸念の声が上がっている（小向，2021）．文章訂正システムについては，コーパスに含まれないような攻撃的表現に対応する際に文章の意味が変わったり不自然な文章になってしまったりする（吉田・松本・吉田・北，2022）ほか，特定の文章の発信を予め規制するシステムが広く普及すれば，言論の自由といった自由権と衝突してしまう恐れも考えられる．

【利己性メッセージ】SNS 上で根拠のない悪口を投稿すると，名誉毀損罪や侮辱罪に問われたり，高額な慰謝料を請求されたりすることがあります.

【利他性メッセージ】SNS 上での誹謗中傷に苦しみ，亡くなる方がいます．あなたの行動が人を殺してしまうかもしれません.

【社会規範メッセージ】たった 1.5%の人しか炎上に加担していないという研究報告があります.

3. 調査と推定の方法

3.1. 調査の方法

オンラインアンケートにより RCT を実施した．アンケートはマイボイスコム株式会社に依頼し，2023 年 11 月 24 日から 2023 年 11 月 27 日までの期間で行った．回答者は全国在住の満 20～69 歳の男女 2,500 人である．回答者を無作為に 4 グループに分け，内 3 グループを介入群，1 グループを統制群としてそれぞれ処置を施した．介入群は利己性メッセージ，利他性メッセージ，社会規範メッセージを表示し，統制群にはいかなるメッセージも表示しなかった.

SNS 上の誹謗中傷投稿への同調は，アンケート内に架空の SNS 投稿とその投稿に対する誹謗中傷的返信を提示し，それぞれの返信に「いいね」及び「再投稿」をするかどうかを質問することで測定した．投稿の内容は，男女の交際における費用の分担についての主張（以下，「デート投稿」）と，転売の是非についての主張（以下，「転売投稿」）である.

また，誹謗中傷的返信と同時に，誹謗中傷にはあたらない返信も回答者に提示し，「いいね」及び「再投稿」をするかどうか質問した．誹謗中傷にあたらない返信は「デート投稿」では「世の中」「経済力」「コスメ」，「転売投稿」では「資本」「地方」「現地」というそれぞれ 3 種類の返信を提示した.

3.2. 推定の方法

各メッセージの誹謗中傷的返信/誹謗中傷にあたらない返信への同調に及ぼす効果を検証するために，以下の推定式に基づき回帰分析を行った.

$$Y_i = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i + \beta_4 X_i + \varepsilon_i, \qquad (1)$$

iは回答者を示す．被説明変数Y_iは，各投稿に対する返信へ，いいね/再投稿をしていれば 1，そうでなければ 0 をとるダミー変数である．A_i，B_i，C_iは，各メッセージを受け取っていれば 1，そうでなければ 0 をとるダミー変数である．すなわち，β_1，β_2，β_3はメッセージが同調行動に及ぼす限界効果を表している．また，説明変数X_iは，各グループでX_iの平均値がバランスしていない場合のコントロール変数である.

4. 分析結果

4.1. 誹謗中傷的返信についての推定結果

推定結果を表１に示した．デート投稿のいいねにおいてはどのメッセージも有意な効果を示さなかった一方で，デート投稿の再投稿においては利己性メッセージと利他性メッセージが，転売投稿のいいね及び再投稿においては利己性メッセージが，それぞれ有意な効果を示した．利己性メッセージはデート投稿に対する誹謗中傷的返信への再投稿を 2.7%ポイント，転売投稿に対する誹謗中傷的返信へのいいねを 4.0%ポイント，再投稿を 2.1%ポイント抑制した．利他性メッセージはデート投稿に対する誹謗中傷的返信への再投稿を 3.3%ポイント程度抑制した．社会規範メッセージはどの推定においても有意な効果を示さなかった．

表１　誹謗中傷的返信についての分析結果

	Dependent variable:			
	デート投稿		転売投稿	
	いいね	再投稿	いいね	再投稿
利己性	-0.004	-0.027**	-0.040**	-0.021**
	(0.018)	(0.011)	(0.016)	(0.011)
利他性	-0.022	-0.033***	-0.023	-0.017
	(0.018)	(0.011)	(0.016)	(0.011)
社会規範	-0.003	-0.012	-0.011	-0.005
	(0.018)	(0.011)	(0.016)	(0.011)
Constant	0.026	-0.005	0.069**	0.011
	(0.036)	(0.022)	(0.031)	(0.021)
Observations	2,385	2,385	2,385	2,385
R^2	0.045	0.063	0.043	0.073
Adjusted R^2	0.035	0.054	0.033	0.064
Residual Std. Error (df = 2361)	0.309	0.188	0.271	0.182
F Statistic (df = 23; 2361)	4.784***	6.906***	4.568***	8.075***
Note:	*p<0.1; >**p<0.05; >***p<0.01			

4.2. 誹謗中傷にあたらない返信についての推定結果

デート投稿，転売投稿それぞれの推定結果をそれぞれ表２，３に示した．利己性メッセージは転売投稿の「現地」へのいいねと再投稿を，利他性メッセージはデート投稿の「経済力」へのいいねと転売投稿の「現地」への再投稿を，それぞれ抑制した．利己性メッセージは転売投稿の「現地」へのいいねを 3.1%ポイント，再投稿を 2.0%ポイント程度抑制した．利他性メッセージはデート投稿の「経済力」へのいいねを 3.2%ポイント程度，転売

投稿の「現地」への再投稿を 2.7%ポイント程度抑制した.

表２　誹謗中傷にあたらない返信についての推定結果（デート投稿）

	Dependent variable:					
	世の中		経済力		コスメ	
	いいね	再投稿	いいね	再投稿	いいね	再投稿
利己性	-0.017	-0.007	-0.015	-0.015*	-0.009	-0.009
	(0.011)	(0.007)	(0.015)	(0.009)	(0.011)	(0.007)
利他性	-0.017	-0.010	-0.032**	-0.008	-0.012	-0.010
	(0.011)	(0.007)	(0.015)	(0.009)	(0.011)	(0.008)
社会規範	-0.002	0.001	-0.018	-0.002	0.002	0.001
	(0.011)	(0.007)	(0.015)	(0.009)	(0.011)	(0.007)
Constant	0.046**	-0.008	0.092***	0.011	0.006	-0.006
	(0.022)	(0.014)	(0.029)	(0.017)	(0.021)	(0.015)
Observations	2,385	2,385	2,385	2,385	2,385	2,385
R^2	0.042	0.061	0.048	0.074	0.047	0.068
Adjusted R^2	0.033	0.052	0.038	0.065	0.038	0.059
Residual Std. Error (df = 2361)	0.191	0.125	0.255	0.149	0.182	0.129
F Statistic (df = 23; 2361)	4.537***	6.684***	5.122***	8.149***	5.116***	7.507***
Note:	*p<0.1; **p<0.05; ***p<0.01					

表３　誹謗中傷にあたらない返信についての推定結果（転売投稿）

	Dependent variable:					
	資本		地方		現地	
	いいね	再投稿	いいね	再投稿	いいね	再投稿
利己性	-0.007	-0.009	0.0002	0.001	-0.031**	-0.020**
	(0.011)	(0.008)	(0.012)	(0.009)	(0.015)	(0.010)
利他性	-0.013	-0.005	-0.007	-0.007	-0.014	-0.027***
	(0.012)	(0.008)	(0.012)	(0.009)	(0.016)	(0.010)
社会規範	0.003	0.004	0.007	0.008	0.003	-0.009
	(0.011)	(0.008)	(0.012)	(0.009)	(0.015)	(0.010)
Constant	-0.001	-0.014	-0.001	-0.018	-0.006	0.029
	(0.023)	(0.017)	(0.023)	(0.017)	(0.031)	(0.019)
Observations	2,385	2,385	2,385	2,385	2,385	2,385
R^2	0.064	0.059	0.070	0.068	0.050	0.062
Adjusted R^2	0.055	0.050	0.061	0.059	0.041	0.053
Residual Std. Error (df = 2361)	0.198	0.144	0.200	0.151	0.266	0.165
F Statistic (df = 23; 2361)	7.017***	6.432***	7.772***	7.453***	5.446***	6.773***
Note:	*p<0.1; **p<0.05; ***p<0.01					

5. 考察

社会規範メッセージのみが全体の推定結果において有意な効果を示さなかったことについて，メッセージの情報と表現という２つの原因が考えられる．メッセージの情報については，誹謗中傷に同調する多くの人々には自身が少数派であるという事実は既知のものであり，社会規範を意識させ行動変容を促すのに十分でなかったということが推測される．表現については，「炎上に加担」という文言が誹謗中傷的返信にいいねや再投稿を選択する行為であると意識させられなかったということが推測される．

転売投稿への「現地」に対する同調が利己性メッセージと利他性メッセージによって，それぞれ抑制されたことが明らかとなった．「現地」は，誹謗中傷にあたらない返信の中で唯一元の投稿に対して反対の立場を取るため，ナッジメッセージが誹謗中傷というよりも意見の対立自体を抑制しているということが推測される．また，デート投稿への「経済力」への同調が利他性メッセージによって抑制されている．誹謗中傷だと捉えた回答者が一定数存在し，利他性の働きによってナッジメッセージの抑制効果を受けたと推測される．

なお，本研究の限界として，本研究で明らかとなったナッジメッセージの効果を SNS 上の様々な投稿に対する効果として一般化できないという点があげられる．本研究ではナッジメッセージが異なる話題において必ずしも同一の効果をもたらすとは言えないことが示唆されている．そのため，ナッジメッセージによる SNS 上の誹謗中傷一般の抑制には，さらなる検証の積み重ねが必要であると考えられる．

6. 結論

本研究では，ナッジメッセージによる SNS 上の誹謗中傷への同調の抑制を試みた．利己性，利他性，社会規範の３つのメッセージを作成して検証した結果，利己性メッセージと利他性メッセージが有意な抑制効果を示した一方で社会規範メッセージは効果を示さなかった．ただし，利己性メッセージと利他性メッセージは誹謗中傷にあたる内容に限らず，単なる意見の対立自体を抑制している可能性がある．また，社会規範メッセージは多くの回答者に対して社会規範を意識させるのに十分でなかったために，回答者全体に対する有意な抑制効果が見られなかったことが考えられる．なお， SNS 上の誹謗中傷全般に対する一般的な効果を明らかにするには，さらなる検証が必要である．

引用文献

小向太郎， 2021．ネットの誹謗中傷問題は解消するのか？～プロバイダ責任制限法改正と今後の課題～．情報処理 Vol.62 No.11，57.

吉田 基信，松本 和幸，吉田 稔，北 研二，2022. BERT を用いた SNS 上における攻撃的文章訂正システム．情報処理学会第 84 回全国大会講演論文集 2022 巻 1 号.