令和7年度　インターネット上の偽・誤情報等への対策技術の開発・実証事業

偽・誤情報の拡散を抑制するためのSNSにおける シェア行動プロセス可視化と信頼性を評価する表示の検討

成果報告書 概要版

2026/3/19

研04_東京大学大学院工学系研究科

目次

目次

1-1. 研究・調査によりアプローチする課題・目指す姿

研究・調査によりアプローチする課題

- 既存の偽・誤情報対策（コンテンツモデレーション）の限界
  - プラットフォーム事業者による削除・凍結措置は、表現の自由の侵害リスクや判断基準の不透明性が常に問題視される。
  - ファクトチェックは質が高い一方で、検証に時間を要するため、拡散スピードの速い現代のSNS環境においては「事後的な対処」にとどまり、拡散防止効果が限定的である。
- 拡散行動の背後にある認知的メカニズムの軽視
  - ユーザーは情報の真偽よりも「感情的刺激」や「自身への肯定（確証バイアス）」を優先して拡散する傾向があるが、従来の対策はこうした心理的要因へのアプローチが不足していた。

上記課題を踏まえ目指す姿・ゴール

- 「情報の信頼性」の可視化による自律的な判断支援
  - 投稿そのものを削除するのではなく、その投稿を行ったアカウントの過去の実績（コミュニティノート被付与履歴など）を可視化することで、ユーザー自身が「拡散すべきか否か」を判断できる環境を整備する。
- ナッジ理論を応用した介入手法の確立
  - 強制力を持たず、しかし望ましい行動（誤情報の拡散抑制）へと自然に誘導する「ナッジ」の手法を用いることで、ユーザーの自律性を尊重しつつ、健全な情報空間を構築する新たなモデルを提案する。

1-2. 研究・調査により期待される偽・誤情報対策への効果

研究・調査により期待される偽・誤情報対策への効果

研究・調査により期待される偽・誤情報対策への効果

- 真偽判断コストの劇的な低減
  - 従来、ユーザーが情報の真偽を確かめるには複数のソースを確認するなどのコスト（Cognitive Budget）が必要であったが、本ツールにより「アイコン」のみで直感的にリスクを把握可能となる。

- 無意識的な拡散行動（反射的リポスト）の抑制
  - 「いいね」や「リポスト」といったボタンを押す直前に、視覚的な注意喚起（アラート）を表示することで、反射的な行動を中断し、一歩立ち止まって考える（熟慮的思考への切り替え）機会を創出する。

具体的な対策メカニズム

1. CN付与履歴の簡易表示:
   X（旧Twitter）のコミュニティノート（CN）データから、各ユーザーの過去のCNノート付与数を数値として表記。

2. CN付与履歴の確認:
   CN付与履歴の表記をクリックすることで過去にどのような投稿に対してCNが付与されたのかの履歴が確認可能

目次

2-1. 研究および有効性等に関する検証の全体像

研究および有効性等に関する検証の全体像

研究・調査のステップ

本研究では、以下の3つのフェーズを通じて、開発した技術の有効性を多角的に検証した。

1. システム開発（Chrome拡張機能の実装）
   - 対象プラットフォーム：X（旧Twitter）
   - 機能：タイムライン上の全ツイートに対し、発信者の過去のCN被付与履歴を視覚的に表示（アイコンの上に記号と数字による表記）。
   - データ基盤：X社が公開しているCNデータのファイルを元にアカウントのCNデータを取得するバックエンドシステムを構築。
2. 被験者実験（短期的な介入効果の検証）
   - 期間：2025年11月26日～12月10日
   - 対象：一般のXユーザー38名（介入群19名、対照群19名）
   - 手法：被験者のPCに拡張機能をインストールし、期間中のリポスト行動を全数記録。介入群には警告を表示し、対照群には表示しない（バックグラウンドログのみ取得）。
3. 追跡調査（長期的な行動変容と習慣化の検証）
   - 期間：実験終了後～2026年1月（約2ヶ月間）
   - 目的：Xの実際の環境での拡張機能に対するユーザーの行動変容を分析。また、警告表示に対するユーザーの「慣れ」の発生有無と、持続的な効果維持の可能性を検証。

2-2. 研究および有効性等に関する検証の個別詳細

実験の実施内容（システムとデザイン）

検証の枠組み

- 目的: 開発した信頼性可視化ツール（Chrome拡張）が、実際のユーザー行動に与える影響を定量・定性の両面から検証する。
- アプローチ: 一般ユーザーを対象とした比較実験（A/Bテスト）と、その後の実生活での追跡調査を組み合わせた複合的な検証を実施。

実験デザイン

- 参加者: X（旧Twitter）を日常的に利用する一般ユーザー 38名
  - 介入群 (n=19): 信頼性情報を表示するツールを使用
  - 対照群 (n=19): ツールを使用しない（通常表示）
- タスク: 指定された投稿セット（政治・災害・科学・エンタメなど多様なトピック）を閲覧し、「リポストするか」「信頼できるか」を判断。

図: 作成した実験ツール

2-2. 研究および有効性等に関する検証の個別詳細

定量的な検証結果（リポスト行動の変容）

リポスト数の比較（拡散抑制効果）

介入群は対照群に比べ、全体およびリスクの高い投稿（投稿者にCN付与）の両方でリポスト数が大幅に減少した。

指標	対照群 (Control)	介入群 (Intervention)	減少率
平均リポスト数（全体）	7.37回	3.00回	**59.3% 減少**
CN付与ユーザの投稿へのリポスト	2.74回	0.74回	**73.1% 減少**

結果の解釈

- 介入効果: 特に誤情報の疑いがある（CNが付与されている）投稿に対して、7割以上の抑制効果が見られたことは、ツールが「リスク回避」に寄与したことを強く示唆する。
- 統計的検定: サンプルサイズ（n=38）の制約により、有意水準5%での統計的有意差は検出されなかったが（$p > 0.05$）、効果量（Cohen's $d \approx 0.48$）は中程度の実用的な効果を示している。

2-2. 研究および有効性等に関する検証の個別詳細

定性的な検証結果（認知変容と受容性）

信頼性評価への影響（認知変容）

- 有意な差: CN付与ユーザの投稿に対する信頼性評価（5段階）において、介入群は対照群よりも有意に低いスコアを示した（$p = 0.0031$）。
- 意味: ツールは行動（リポスト）だけでなく、ユーザーの意識（この情報は怪しいという気付き）にも確実に働きかけていることが統計的に実証された。

群 × CN有無	平均評価
対照群・CN未付与	3.10
対照群・CN付与	3.12
介入群・CN未付与	3.06
介入群・CN付与	**3.00**

CN付与ユーザの投稿への信頼性評価（5段階）

ユーザー受容性 (SUSスコア)

- ユーザビリティ評価: システム使用性尺度（SUS）の平均点は 53.3点。
- ポジティブな意見:「判断に迷った時の指標になる」「バッジがあると思いとどまれる」
- ネガティブな意見:「画面が少しうるさい」「推しにバツがついているようで不快」
- 課題: 68点（平均基準）を下回っており、日常使いしたくなるUIへの改善が必要。

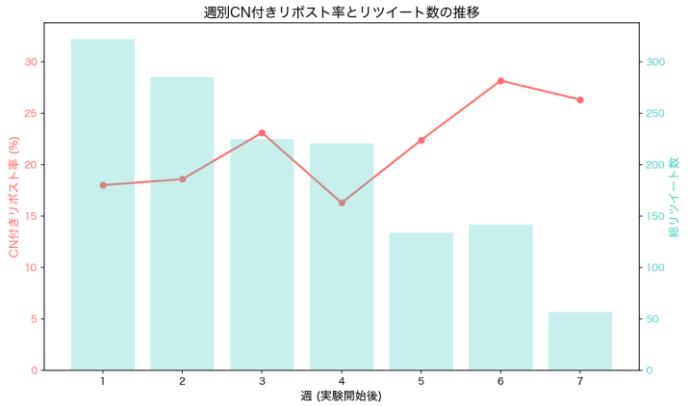
2-2. 研究および有効性等に関する検証の個別詳細

追跡調査（長期的な習慣化の検証）

実験後の長期利用データ分析

実験終了後もユーザー（34名）のログを分析し、効果の持続性を検証した。

- 初期の強力な抑制 (Week 1)
  - CN付き投稿のリポスト率は 18.01% と低水準。実験直後の「意識が高い」状態では、ツールが効果的に機能している。
- 習慣化による効果減衰 (Week 3以降)
  - 3週目以降、リポスト率は 23.11% ～ 28.17% へと上昇。
  - 考察：警告表示が日常の風景となり、注意を払わなくなる「馴化（Habituation）」が発生している。
- 結論：単一の警告デザインでは長期的な効果維持は困難であり、動的なUI変更等の対策が不可欠である。



図：週別CN付きリポスト率の推移（実験実施日基準）

目次

3-1. 研究・調査の総合的な考察

成果の総括：ナッジ介入の有効性

主要な成果

1. 行動変容の実証:
   - 強制的な削除や検閲を行わずとも、「情報の信頼性を可視化する」というナッジ（行動介入）の手法によって、利用者の拡散行動を抑制できることを実証した。
   - 特に、誤情報の疑いがある投稿（CN付与ユーザの投稿）へのリポストを7割以上削減できたことは、本手法の高いポテンシャルを示している。
2. 認知プロセスへの影響:
   - ツール使用者のCN付与投稿に対する信頼性評価が有意に低下したことは、ツールが単にリポストボタンを押すのを躊躇させるだけでなく、「情報の真偽を疑う」というクリティカルシンキングを促した証左である。
3. 技術的実現可能性:
   - 既存のプラットフォーム（X）の公開データのみを用いて、リアルタイムに信頼性を評価するシステムが構築可能であることを示した。

3-1. 研究・調査の総合的な考察

発見された課題と限界

1. サンプルサイズと統計的検出力

- 本実験（n=38）では、行動変容（リポスト数）において大きな効果量が観察されたものの、統計的有意差（$p<0.05$）の確定には至らなかった。

2. 習慣化（Habituation）とトピック依存性

- 習慣化の壁: 追跡調査により、警告表示の効果は2週間程度で減衰することが確認された。「慣れ」を防ぐためのUIの工夫（デザインの定期変更など）が必要である。
- トピック依存性: 政治やエンタメ（推し活）など、ユーザーの感情や信条が強く関わる領域では、理性の働き（システム2）が情動（システム1）に負け、警告を無視して拡散する傾向が見られた。

3-1. 研究・調査の総合的な考察

政策的示唆と社会的意義

プラットフォーマーへの提言

- 「発信者単位」の信頼性指標の導入：
  - 現在のXコミュニティノートは「投稿単位」の評価だが、本研究により「発信者（アカウント）単位」の信頼性情報も、ユーザーの判断に強い影響を与えることが示された。
  - この両輪を組み合わせることで、より強固な偽・誤情報対策が可能となる。

政策立案者への提言

- ユーザーエンパワーメントの推進：
  - プラットフォーム規制（削除・法規制）は表現の自由との衝突を生むが、本技術のような「ユーザーの判断を支援する技術」は、民主主義的な価値観と整合する。
  - こうした技術の研究開発および社会実装（リテラシー教育との連携など）を、政策的に支援していくことが、健全な情報空間の構築につながる。

3-2. 研究・調査にあたっての課題・展望

研究・調査にあたっての今後の課題

- サンプルサイズの拡大:
  本実験（n=38）では探索的な検証に留まったため、普遍的な結論を得るには数百人～千人規模の大規模実証が必要である。
- 多様な属性での検証:
  年代、リテラシーレベル、政治的志向などの属性による効果の違いを詳細に分析する必要がある。
- 多様なプラットフォームへの対応:
  現在はPC版Chrome拡張機能のみだが、SNSの利用中心であるスマートフォンアプリ（iOS/Android）への対応が、実際の効果を検証するためには必要である。

上記課題を踏まえた今後の展望

- 大規模実証（統計的確証の獲得）:
  今回の実験（n=38）では統計的有意差（$p<0.05$）が出ていない。そのため、まずはサンプルサイズを増やし効果の有無をさらに検証していくことが重要である。
- プラットフォームへの実装:
  本機能はブラウザ拡張機能として実装したが、将来的にはSNSプラットフォーム自体が標準機能としてこうした信頼性可視化システムを組み込むことが、情報空間全体の健全化には重要である。