LLMの安全性シンポジウム 2025

フェイクニュースの社会課題を解決する偽情報対策技術

富士通株式会社

富士通研究所　データ＆セキュリティ研究所
ソーシャルトラストコアプロジェクト
シニアプロジェクトディレクター

山本　大

2025年7月23日

謝辞
本発表は、ＮＥＤＯ（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP22007）の結果得られたものです。

© 2025 Fujitsu Limited

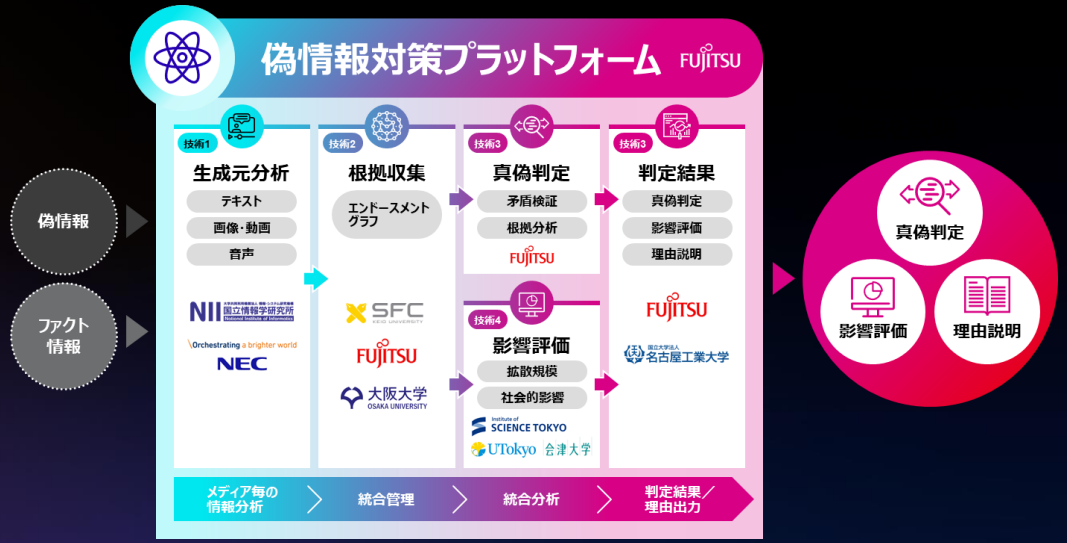
産学連携エコシステムで偽情報の社会課題解決に挑む

FUJITSU

「経済安全保障重要技術育成プログラム」をプライムで受託

オールジャパンの共創研究

偽情報対策技術・プラットフォームの開発



© 2025 Fujitsu Limited

生成AIの普及で、偽情報の問題は様々な分野へ

FUJITSU

生成AIやコミュニケーションツールの発展により、偽情報がより拡散しやすい環境

- 作成時間の短縮・品質向上
- 情報の偏り・拡散の巧妙化
  - フィルターバブル、エコーチェンバー ：情報の偏り
  - ソーシャルボットによる拡散の巧妙化

一度広まると

- **不安感からの情報の拡散**
- **不確かな情報による行動による混乱**

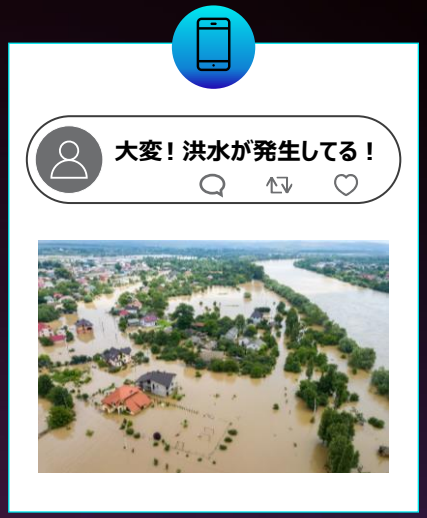
偽情報の拡大に対応するため、IT技術によりユーザー支援が必要

© 2025 Fujitsu Limited

偽情報の解決に向けたアプローチ

FUJITSU

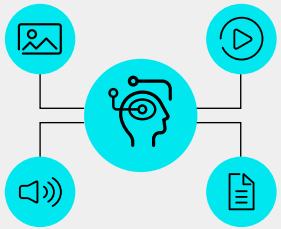
- インターネット情報に対し、第三者の情報/評価などを根拠として紐づけ、情報の真偽を分析する
- 国や自治体、カメラやセンサーなど信頼のおける第三者の情報が根拠になる世界



元の情報に根拠として紐づけ

根拠・エンドースメントの候補

メディア毎の情報分析
(画像 / 映像 / 音声)



第三者による情報

別カメラの画像

第三者による評価

根拠となる複数の情報の整合性や矛盾から、情報の真偽を分析する

© 2025 Fujitsu Limited

偽情報対策プラットフォーム

FUJITSU

様々な根拠や評価の関係性を「根拠/エンドースメントグラフ」で統合し、これらの整合性や矛盾を分析することによって、デジタル上で流通する情報の真偽を判定可能にする。
さらに、偽情報の社会への影響を予測する、新たな「偽情報対策プラットフォーム」を実現

© 2025 Fujitsu Limited

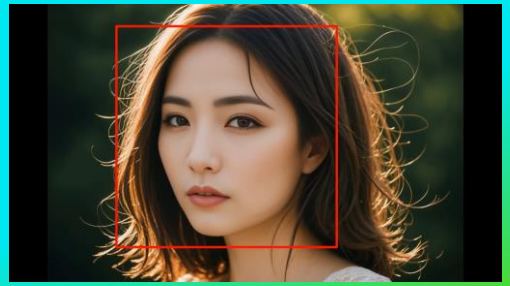
研究開発技術（1/2）

FUJITSU

ディープフェイク検知

真偽判定の対象となるSNS投稿などの情報を構成するメディアデータ（画像、映像、音声）に対し、ディープフェイクを用いた意図的な偽情報を検知する技術に加え、改ざん箇所や生成手法を、判定の確信度を含める形で推定し、これらを根拠情報として出力する研究開発を実施

顔映像の真贋判定

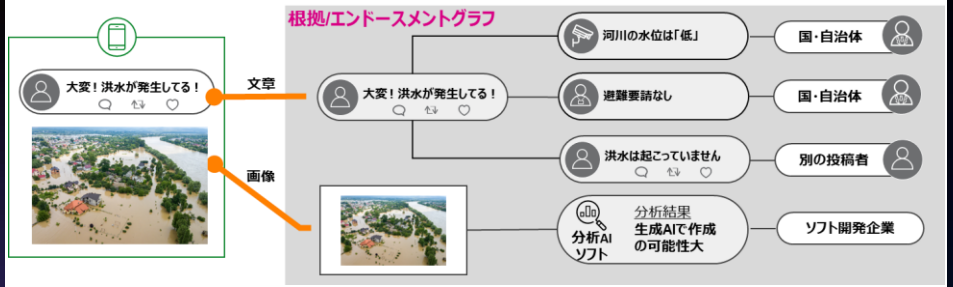


緑枠:真　赤枠:偽

根拠/エンドースメント管理

インターネット上から収集した様々な根拠情報の関係性を「根拠/エンドースメントグラフ」として統合し、構造化した上で蓄積し、総合的な真偽判定支援や影響度評価において活用可能にする技術を開発

グラフとして接続された根拠に矛盾がないことが情報の正しさの証明に



© 2025 Fujitsu Limited

研究開発技術（2/2）

FUJITSU

根拠・エンドースメント分析（総合真偽判定）

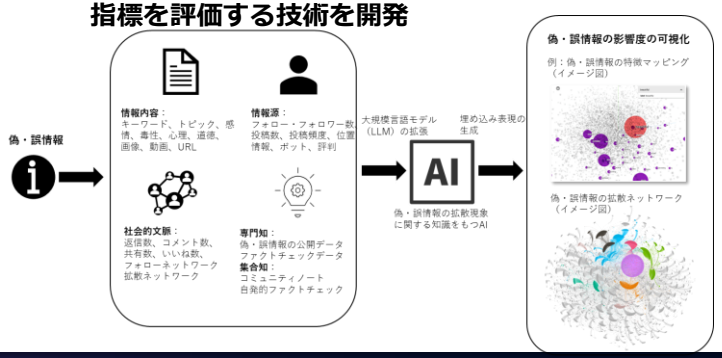
「根拠・エンドースメントグラフ」から、大規模言語モデル(LLM)によって真偽の根拠を分析し、判定結果とともに自然言語(文章)としてユーザに根拠を説明

紐づけられた根拠の矛盾を分析し、情報の真偽判定の根拠を説明

偽情報の影響度評価

偽情報特有の特徴量との類似性を評価し、拡散規模や影響度を評価

- SNS等での拡散による社会的影響の評価は対処の検討にとって重要
- SNSデータから**偽情報評価用AIモデル**を構築
  - 過去の偽情報との類似度や拡散速度等の偽情報の特徴を分析し、**拡散規模や社会的な影響度**などの指標を評価する技術を開発



© 2025 Fujitsu Limited

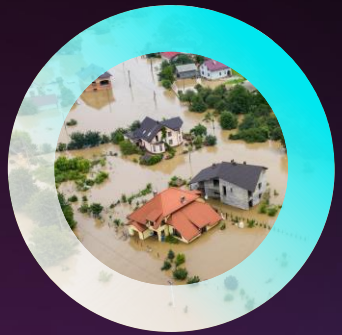
新たなユースケースの創出

FUJITSU

公的機関 (自治体や省庁等)

ファクトチェックの自動化

洪水、土砂災害、地震などの自然災害などの偽情報や経済安全保障に関わる情報攻撃に対するファクトチェックに利用



民間企業 (メディアや金融等)

エビデンス収集・確認の自動化

様々な業務における人手によるファクトチェックやエビデンス確認の負荷や期間を短縮

段階的に拡大

様々な利用シーン

認知セキュリティなど偽情報対策のユーザ影響を考慮した仕組み作りも併せて、様々なユーザ価値に仕立て展開

© 2025 Fujitsu Limited

今後について

- 2025年度末までに、4つの技術を統合した偽情報対策プラットフォームを構築
- 2025年第2四半期よりトライアル・実証実験を実施予定
- これにより9者はNEDOを通じて、本事業の成果を新産業創出および国際競争力強化につなげ、日本の重要性、優位性の向上に貢献

- 富士通は、世界の繁栄と安定が両立する、信頼性のあるデジタル社会を共に創るという当社のマテリアリティの1つである「デジタル社会の発展」に貢献する価値をお客様や社会に提供

© 2025 Fujitsu Limited

Thank you!

© 2025 Fujitsu Limited