記者説明会

富士通と産学組織が9者で共創し、世界初の偽情報対策プラットフォームの構築を開始

2024年10月16日

富士通株式会社
情報・システム研究機構国立情報学研究所
日本電気株式会社
慶應義塾大学SFC研究所
国立大学法人東京科学大学
国立大学法人東京大学生産技術研究所
公立大学法人会津大学
国立大学法人名古屋工業大学
国立大学法人大阪大学

式次第

FUJITSU

1 登壇者による説明

- **「本プロジェクトの全体像」**
  富士通株式会社 データ＆セキュリティ研究所　リサーチディレクター　山本 大
- **「技術1：ディープフェイク検知技術の研究開発」**
  大学共同利用機関法人情報・システム研究機構国立情報学研究所 コンテンツ科学研究系　教授　山岸 順一
- **「技術2：メディア分析基盤」**
  慶應義塾大学 大学院政策・メディア研究科　特任教授　鈴木 茂哉
- **「技術4：偽情報評価技術（影響度評価技術の研究開発）」**
  国立大学法人東京科学大学 環境・社会理工学院　教授　笹原 和俊
- **「技術3：総合真偽判定、全体まとめ」**
  富士通株式会社 データ＆セキュリティ研究所　リサーチディレクター　山本 大

2 質疑応答

3 フォトセッション

本プロジェクトの全体像

発表者　富士通株式会社 データ＆セキュリティ研究所
リサーチディレクター　山本 大

本日のキーメッセージ

FUJITSU

国内屈指のアカデミアや企業によるオールジャパン体制で、偽情報の検知から根拠収集、分析、評価までを統合的に行う点で世界初となる偽情報対策プラットフォームの構築を開始

背景：偽情報が大きな社会問題に

FUJITSU

様々な分野で偽情報が問題化

自然災害

医療

政治

経済

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

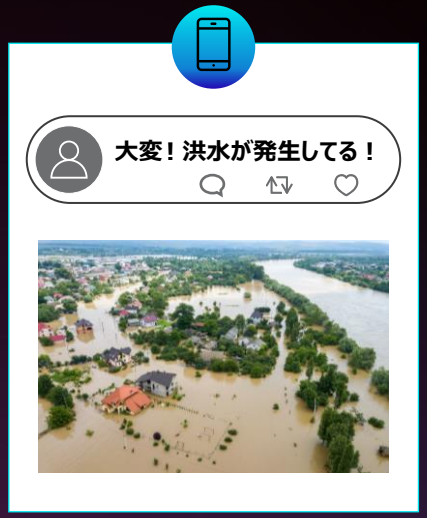
一度広まると

- **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**

解決に向けたアプローチ

FUJITSU

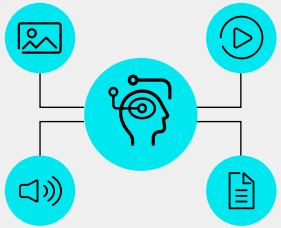
- インターネット情報に対し、第三者の情報/評価などを根拠として紐づけ、情報の真偽を分析する
- 国や自治体、カメラやセンサーなど信頼のおける第三者の情報が根拠になる世界



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

本プロジェクトの全体像（概要）

FUJITSU

様々な根拠の関係性を「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
分析することで真偽を判定し、社会への影響度を評価する偽情報対策プラットフォームを構築

本プロジェクトの全体像（詳細）

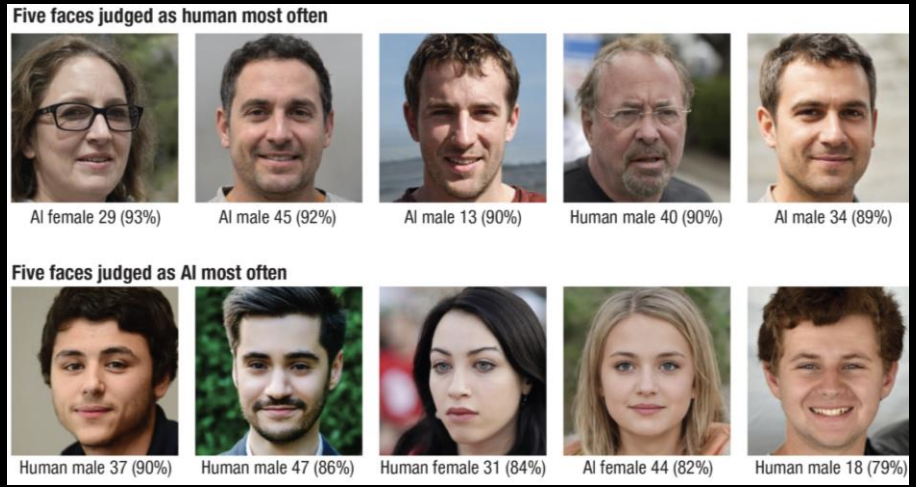
FUJITSU

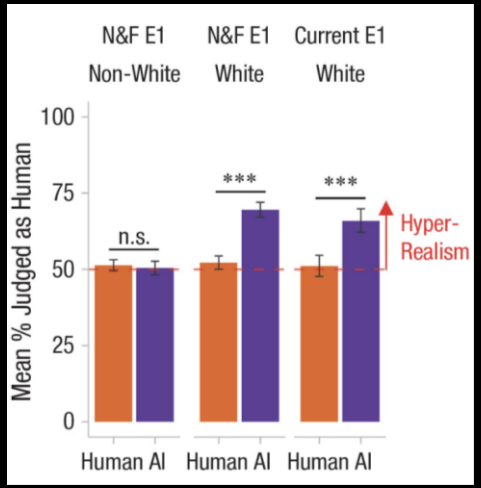
技術1：ディープフェイク検知技術の研究開発

発表者　大学共同利用機関法人情報・システム研究機構国立情報学研究所
コンテンツ科学研究系
教授　山岸 順一

ディープフェイクは人間には見抜けない状況になりつつある

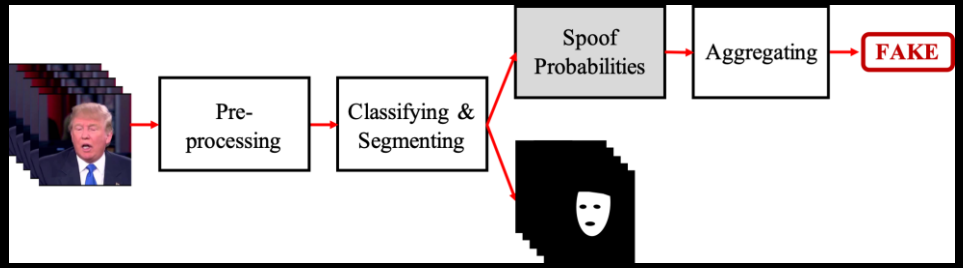
- E. J Millerらによる最新研究によると、最新生成AIによる顔画像の一部は本当の人間の顔画像よりも、より人間らしいと被験者に誤判断される割合が高いと報告
- 多くの被験者は、本当は誤っているにも関わらず、自分の判断結果に自信があると報告





Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. (2023). AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. Psychological Science, 34(12), 1390-1403. https://doi.org/10.1177/09567976231207095

顔のディープフェイク検出モデルの学習と推論



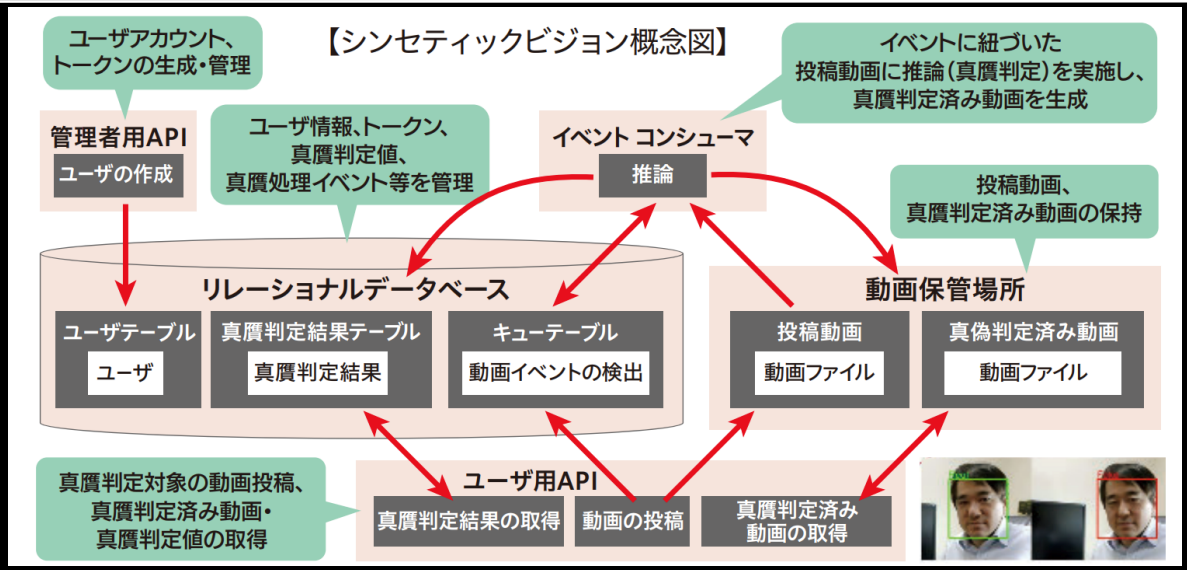
D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, pp. 1-7, 2018 (被引用数 1500回)

H. H. Nguyen, J. Yamagishi, & I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019, pp. 2307-2311, 2019 (被引用数 685回)

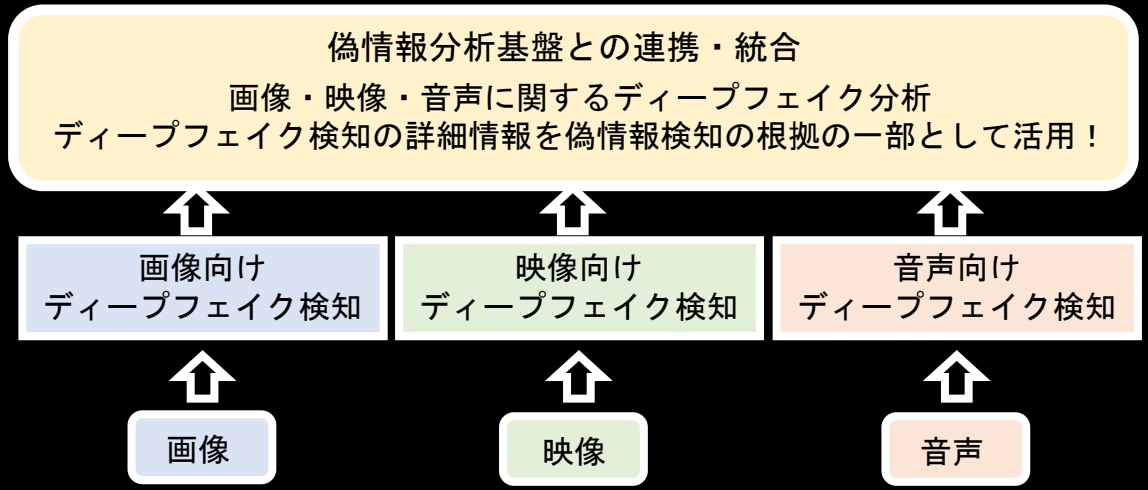
H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, "Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos," The Tenth IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2019), 2019, (被引用数 517回)

シンセティックビジョン

Check the full story!
NII Today No. 100



偽情報分析のためのメディア解析技術

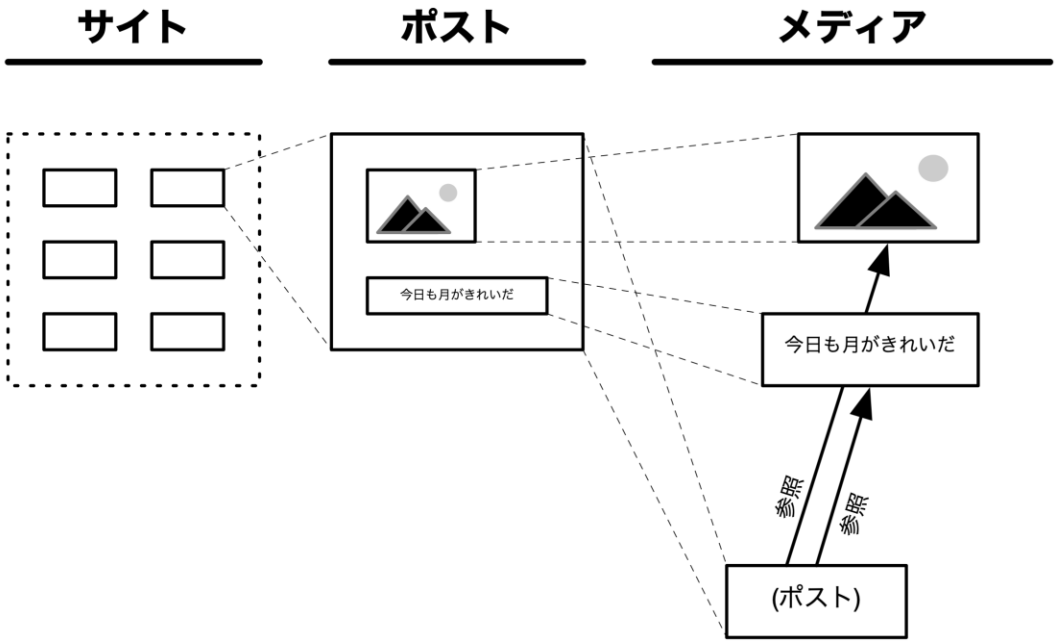


技術2：メディア分析基盤の研究開発

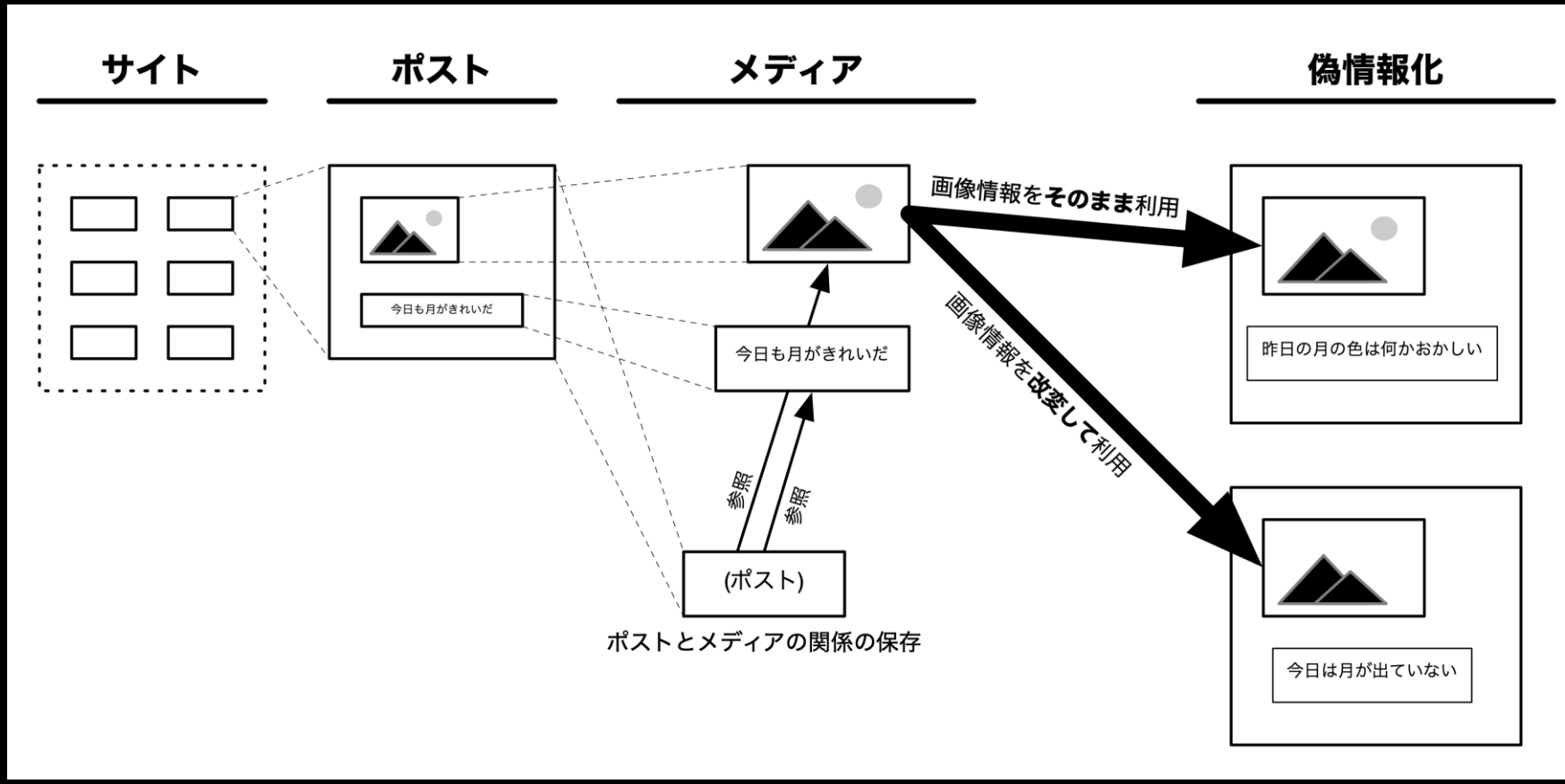
発表者　慶應義塾大学 大学院政策・メディア研究科
特任教授　鈴木 茂哉

メディア分析基盤

偽情報や誤情報分析の対象情報を
分析に適切な大きさの情報ごとに解析結果を付与整理し
様々な分析結果を統合し俯瞰的な分析を可能とする基盤



ポストとメディアの関係の保存



ポストとメディアの関係の保存

情報の断片とそれらの関係 + 個別分析結果ラベル
→ 情報の断片間の関連づけによる整理

- 情報の断片: 取り扱う単位
  - 発信者
  - ポストに含まれる情報（テキスト、画像、時刻、位置情報…）

    …

- 情報の断片相互の関係:
  - 発信者とポストの関係
  - ポストとポストに含まれる情報の関係
  - ポストとポストの関係

    …

- 情報の断片 と 情報の断片の関係 について、分析した結果をラベルとして付与

- 情報の断片間の関係はデータ構造の一種であるグラフによって表現　→ 根拠、エンドースメントグラフ

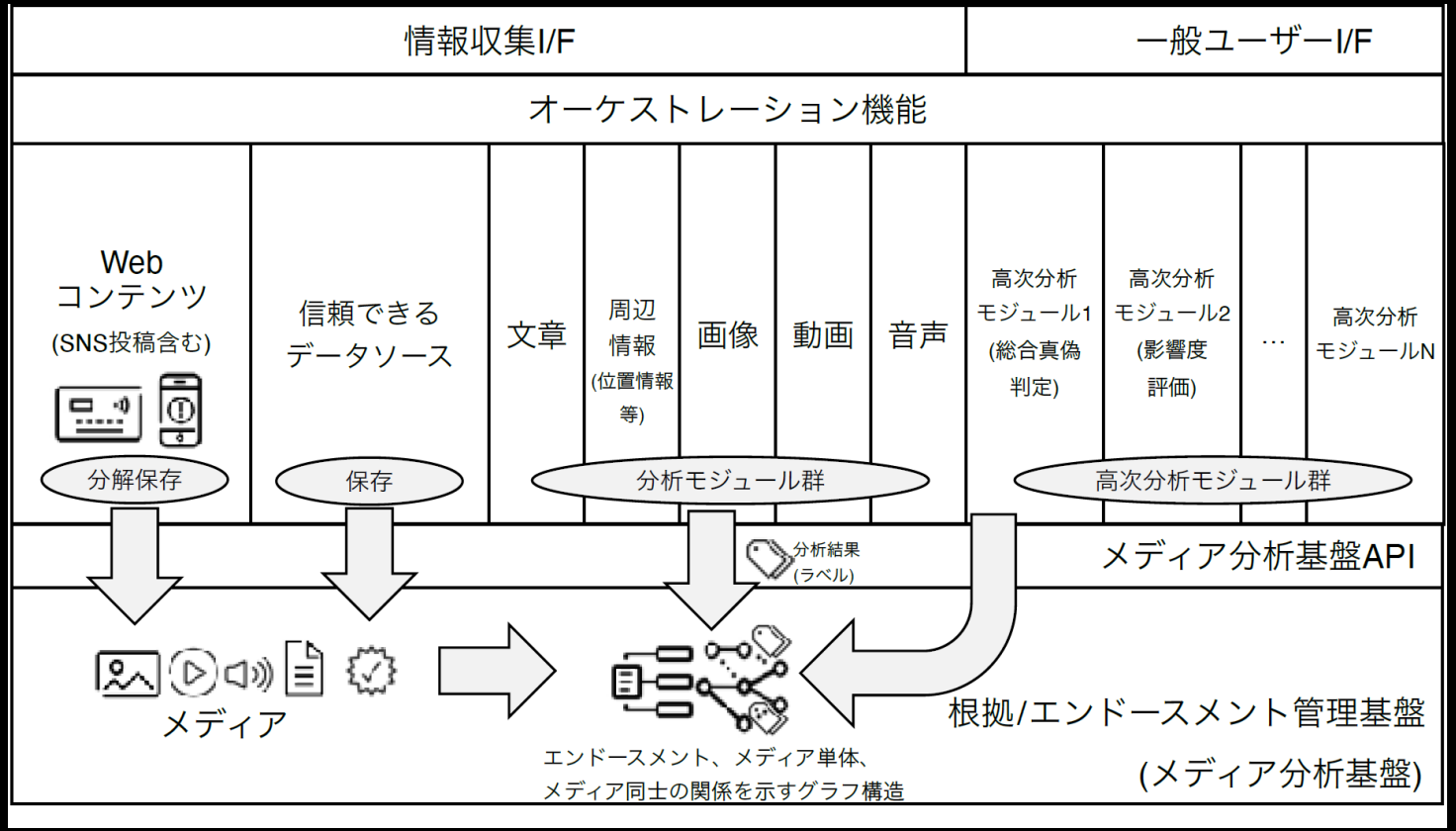
分析のステップ

- 情報の断片に分解し、グラフ表現にする

- ラベリング（メディア分析）
  - 情報の断片や、情報同士の関係を分析する。結果としてラベルを貼る

- 統合的俯瞰的分析（高次分析）
  - 情報の断片の関係、分析によって付与されたラベルを活用しつつ、高次分析アルゴリズムを適用し、総合的俯瞰的分析結果を得る

メディア分析基盤



技術4：偽情報評価技術
（影響度評価技術の研究開発）

発表者　国立大学法人東京科学大学 環境・社会理工学院
教授　笹原 和俊

偽情報の影響度評価の従来技術と課題

従来技術

- SNSの投稿内容から典型的な特徴量を測定し、教師あり学習で分類器を作成
- 偽情報の評価は主に、ファクトチェックやニュース組織、SNS業者内部の検証チームによって行われた評価のデータベースとの照合

課題

- 過去の限定されたデータセットに依存しており、新しい偽情報に迅速に対応できない
- ファクトチェック組織等の人手による方法では、情報拡散の速度に追いつかず、スケールしない

偽情報評価技術の概略図

偽・誤情報

情報内容：
キーワード、トピック、感情、毒性、心理、道徳、画像、動画、URL

情報源：
フォロー・フォロワー数、投稿数、投稿頻度、位置情報、ボット、評判

社会的文脈：
返信数、コメント数、共有数、いいね数、フォローネットワーク
拡散ネットワーク

専門知：
偽・誤情報の公開データ
ファクトチェックデータ
集合知：
コミュニティノート
自発的ファクトチェック

大規模言語モデル（LLM）の拡張

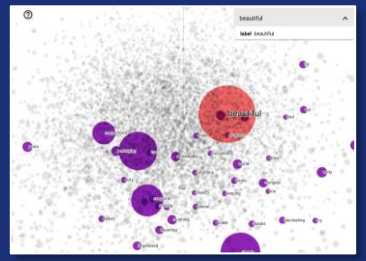
埋め込み表現の生成

AI

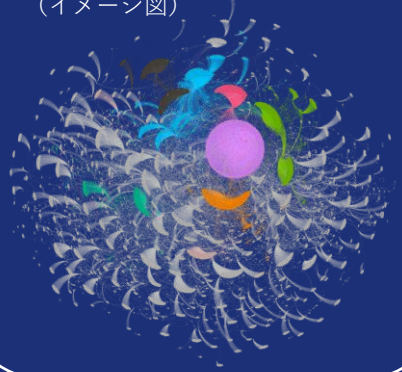
偽・誤情報の拡散現象に関する知識をもつAI

偽・誤情報の影響度の可視化

例：偽・誤情報の特徴マッピング（イメージ図）



偽・誤情報の拡散ネットワーク（イメージ図）



研究開発の内容と効果

研究開発内容

- 多様な特徴量抽出とAIモデルによる偽情報拡散度推定
- 埋め込み表現生成と情報拡散特徴の可視化（ブラウザ対応）
- 連携：東大（データ収集・処理・分析）、東京科学大（AIモデル）、会津大（可視化技術）

期待される効果

- 偽情報の社会的影響の評価支援
- 偽情報の識別と拡散メカニズムの理解促進
- 研究者・ファクトチェッカーによる実践的活用

準備状況

X（旧Twitter）上で拡散する偽情報の評価を早期に行うため、Xの日本語全量規模のストリームから評価に必要となる多様な情報をリアルタイムに抽出し、共有する基盤技術に着手

X

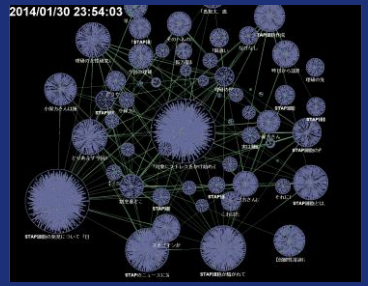
日本語の
全量投稿

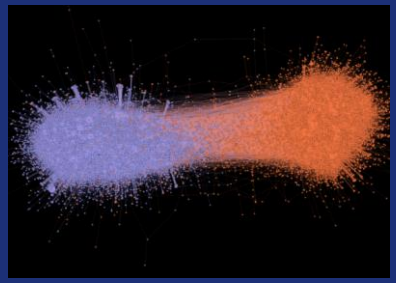
情報源

情報内容

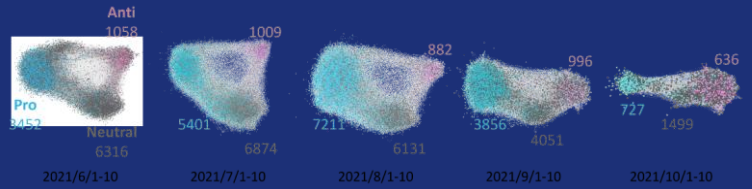
社会的文脈

偽情報評価のための多様な分析





トピック毎の情報拡散　エコーチェンバー



議論の多極化の推移

FUJITSU

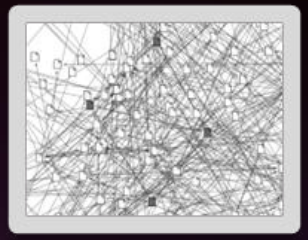
技術3：総合真偽判定支援、全体まとめ

発表者　富士通株式会社 データ＆セキュリティ研究所
リサーチディレクター　山本 大

技術3：根拠・エンドースメント分析（総合真偽判定）

FUJITSU

「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



根拠・エンドースメントグラフ
（イメージ）

入力

総合真偽判定

グラフから根拠となる情報を抽出

抽出した情報間の矛盾を分析

抽出した根拠から説明文を生成

出力

判定結果と根拠を説明

判定結果

理由説明

偽情報対策特化の日本語LLM

- 特定領域における言語能力の向上
- 推論の高速化
- ハルシネーションの抑制

「Fugaku-LLM」「Takane」の開発で培われた技術を活かす

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

本プロジェクトにおける取り組み

- 名古屋工業大学 ： 認知科学に基づくUI、情報提供技術を開発し、適切なユーザー行動を促す
- 大阪大学 ： 根拠情報の一つとなるIoTセンサーデータの収集技術を開発
- NEC ： 画像、映像、音声に含まれる内容をテキストとして抽出するメディア理解技術を開発

想定ユーザー

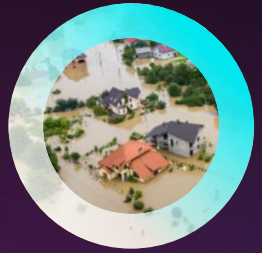
FUJITSU

比較的リテラシーの高い、組織ユースによる試行を通じて本プラットフォームを改善した後、事前にリテラシー教育を受講した一般ユーザーへ段階的に範囲を拡大

公的機関 (自治体等)

災害時のファクトチェック

洪水、土砂災害、地震などの
自然災害時における
偽情報に対する
ファクトチェックに利用



民間企業 (ファクトチェック機関等)

ファクトチェック自動化

記事作成前に行われる
手作業によるファクトチェックを
本プラットフォームで実施し、
負荷や期間を短縮

段階的に拡大

一般ユーザー

各種ファクトチェック

偽情報対策プラットフォームが
ユーザーに良くない影響を
逆に与えてしまわないように
注意深く進める

今後について

FUJITSU

- 2024年度は民間企業・公的機関向けユースケースの分析と機能要件の抽出を行うとともに、各技術の研究開発を実施
- 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

Thank you!

FUJITSU