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

画像・動画を中心としたSNS上の投稿の真偽判定システムの開発・実証
成果報告書

2026/3/19

技11_Sakana AI株式会社

目次

目次

1-1. 開発・実証のサマリ

アプローチする課題・目指す姿	• Xを始めSNSは偽・誤情報の主戦場となっており、社会に大きな影響を与えている。この対策として情報の真偽検証が挙げられるが、現状は多角的な視点を要するマニュアル作業であり、膨大な工数と高度な専門性が求められるほか、昨今の生成AI技術の進歩によりその判定はより難しくなっている。膨大な偽・誤情報の中から、どの情報を対象に、いかに効率よく、標準的に、高度に検知するかが、偽・誤情報対策の鍵となっている。 • 本実証事業では、SNS空間で拡散される①偽・誤情報・論調（ナラティブ）を可視化し、真偽判別すべき情報の見極めと、②その判別の実施、及び③対策案の立案を支援するシステムを開発する。これにより、プラットフォーム事業者や一般ユーザが早期対応・判断できる健全な情報流通環境の構築を目指す。		
技術区分	コンテンツの真偽判別支援技術、改ざん検知技術	実施体制 （下線：技術開発主体）	Sakana AI株式会社
対象とするモーダル種	文章、画像、音声、動画		

技術開発の取組・成果

1．X上のナラティブ（論調）とそれを構成するXの投稿及びその偽・誤情報スコアの可視化
- ナラティブを一覧化し、対策を打つべき偽・誤情報の優先度付けを可能とした（効率化）

2．偽・誤情報の真偽判定システムの構築
- 動画と画像を含む実用的な真偽判定システムの構築と、それを評価するために実際のXの投稿から収集したベンチマークを整備し、平均84%の検知精度を実現（効率化、高度化）

3．偽・誤情報対策実施者が処置するべき対応策の立案
- AIによる次のアクションの提案ならびにカウンター発信の効果検証を可能とするシミュレーション技術を確立（効率化、標準化）

社会実装に係る取組・成果

- 最前線の実務者（メディア、ファクトチェック有識者団体）との連携により、現場ワークフローに即した真に実効性のあるシステム要件を確立
- 国家安全保障に係るインテリジェンスの専門家である**中曽根平和研究所 情報空間のリスク研究会**様によるレビューを経て、その有用性を確認
- 普及活動を通じた導入見込み先の拡大
- 市場ニーズの多様性を踏まえ、「共通基盤」と「カスタマイズ」を組み合わせたハイブリッドな提供モデルを策定し、ビジネスモデルの具現化を実現

技術開発及び社会実装にあたっての課題・展望

【X以外のプラットフォームへの展開】
- 本システムのコア機能は、X以外のプラットフォームへも横展開が可能なアーキテクチャとなっている。実装にあたっては、プラットフォームごとのAPI仕様やデータ制約に起因する機能差分を考慮し、導入効果とコストのバランスを精査した上で、顧客ニーズに合わせた最適な適用範囲を定義していく。

【導入先ごとのカスタマイズ】
- 事業開始当初の想定を超え、幅広い事業者から関心が寄せられた結果、適用領域が拡大した。今後は、多様なニーズを効率的に満たすための、「共通基盤」と「カスタマイズ」の境界を明確に定義・標準化した上で、スケーラブルなプロダクトとしての製品化・商用化を推進する。

代表者コメント

Sakana AI
事業開発本部長
谷口博基

生成AIにより偽・誤情報の脅威が深刻化する中、本事業では幅広い有識者と密に連携し、可視化・判定・対策を一気通貫で支援する技術を開発し、情報分析の専門家より高い評価を得ることができました。今後は、実証で得られた知見を基に「社会を守るインフラ」としての製品化を推進し、信頼できるデジタル空間の実現に貢献してまいります。

目次

2-1. 開発技術によりアプローチする課題

開発技術によりアプローチする課題

- Xを始めSNSは偽・誤情報の主戦場となっており、社会に大きな影響を与えている。この対策として情報の真偽検証が挙げられるが、現状は多角的な視点を要するマニュアル作業であり、膨大な工数と高度な専門性が求められるほか、昨今の生成AI技術の進歩によりその判定はより難しくなっている。膨大な偽・誤情報の中から、どの情報を対象に、いかに効率よく、標準的に、高度に検知するかが、偽・誤情報対策の鍵となっている。

SNS上の膨大な情報	偽・誤情報の複雑、巧妙化	対策
限られたリソースでどの情報をチェックすればよいのか？	膨大なチェック工数人間の目で見極めることができない	有効な打ち手がわからない

2-2. 開発技術により目指す姿・ゴール

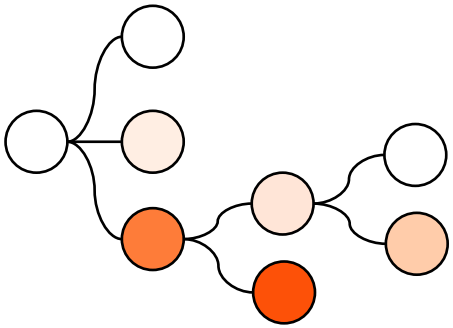
開発技術を通して目指す姿・ゴール

- 本実証事業では、SNS空間で拡散される①偽・誤情報・論調（ナラティブ）を可視化し、真偽判別すべき情報の見極めと、②その判別の実施、及び③対策案の立案を支援するシステムを開発する。これにより、プラットフォーム事業者や一般ユーザが早期対応・判断できる健全な情報流通環境の構築を目指す。

SNS上の膨大な情報	偽・誤情報の複雑、巧妙化	打ち手
限られたリソースで どの情報をチェックすればよいのか？	膨大なチェック工数 人間の目で見極めることができない	有効な打ち手がわからない

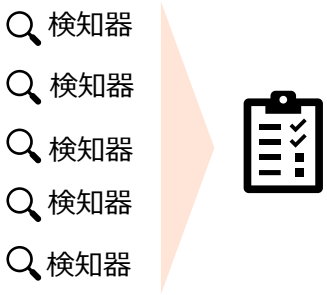
SNS空間の可視化

階層ナラティブツリーによる分析



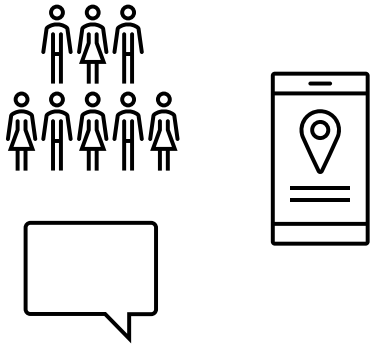
総合的な偽・誤情報判定

複数検知器の組み合わせ



対策案の立案

SNS空間のシミュレーション



2-3. 開発技術により対処可能なユースケース

開発技術により対処可能なユースケース

- ユースケース１：災害時における被害情報・安否・救助要請の真偽確認

膨大な投稿が短時間に集中する災害時において、「ナラティブの可視化」機能により、拡散しているトピックとその真偽を早期に特定する。人手による全量チェックが不可能な状況下で、優先して対処すべき情報を見極め、迅速な人命救助や混乱防止を支援する。

2026年1月6日の島根県で発生した地震に関する投稿の可視化

①発生直後の衝撃や安否確認に関する投稿が多く発生していることがわかる

②そのうち複数の投稿について、AI生成、改ざん、情報操作が疑われることを示唆。

→ **消防・救援のための迅速・正確な情報把握に役立つ**

2-3. 開発技術により対処可能なユースケース

開発技術により対処可能なユースケース

- ユースケース２：専門性・工数を要する真偽判別の効率化

専門家が数日かけて裏取りを行うようなファクトチェックにおいて、本実証システムを用いることで、人間の調査プロセス（仮説検証の繰り返し）を模倣して多角的な検証を行い、判定に至る論理構成やレポートの下書きまでの素材、証跡提供の支援を行う。

2-3. 開発技術により対処可能なユースケース

開発技術により対処可能なユースケース

- ユースケース３：誤った認識を打ち消すための「カウンター発信」

偽・誤情報への対抗や正しい情報の周知において、開発したSNSシミュレーションを用い、カウンター発信（誤情報の拡散を抑制するために行う、事実に基づく訂正や反論の投稿）による世論への波及効果を事前に検証することができる。
ペルソナ群に対して、複数の発信案に対する反応や感情の推移を比較することで、炎上等のリスクを最小化しつつ、最も有効なメッセージの立案を可能にする。これにより、現状の分析にとどまらず、具体的な対策実行までを支援する。

カウンター発信案　　SNSシミュレーション　　それぞれの発信への反応を比較

目次

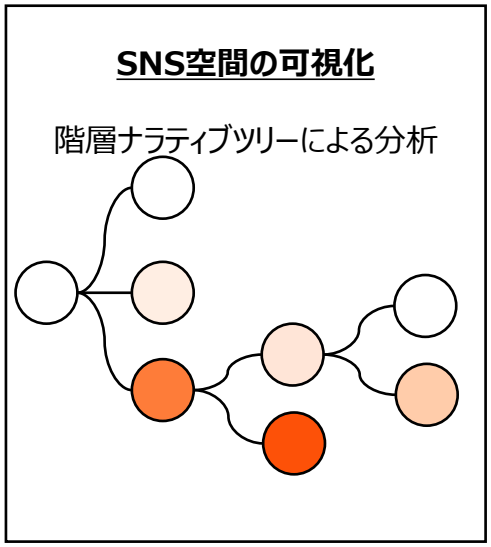
3-1. 技術開発の全体像

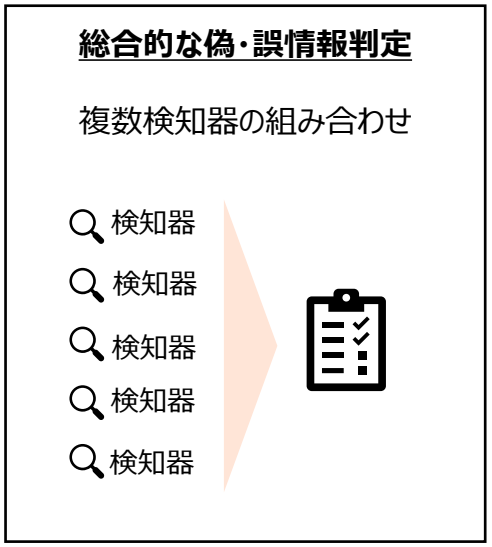
技術開発に係る取組・成果の全体像

①SNS空間の可視化
- SNS空間の可視化方法として、階層ナラティブツリーマップを提案・開発
- 各ナラティブを簡潔かつ高解像度で表現するためのノベルティサーチ技術の開発

②総合的な偽・誤情報判定
- 画像、動画の生成、加工の跡を検知するためのAI機能の開発
- 自動真偽判別機能とリファレンス検索（インターネットを使ったキーワード検索、画像のリバースイメージ検索）
- ユーザー分析/反応分析

③対策案の立案
- AIによる次のアクションの提案
- カウンター発信の立案とその有効性を評価するためのABMシミュレーションの開発

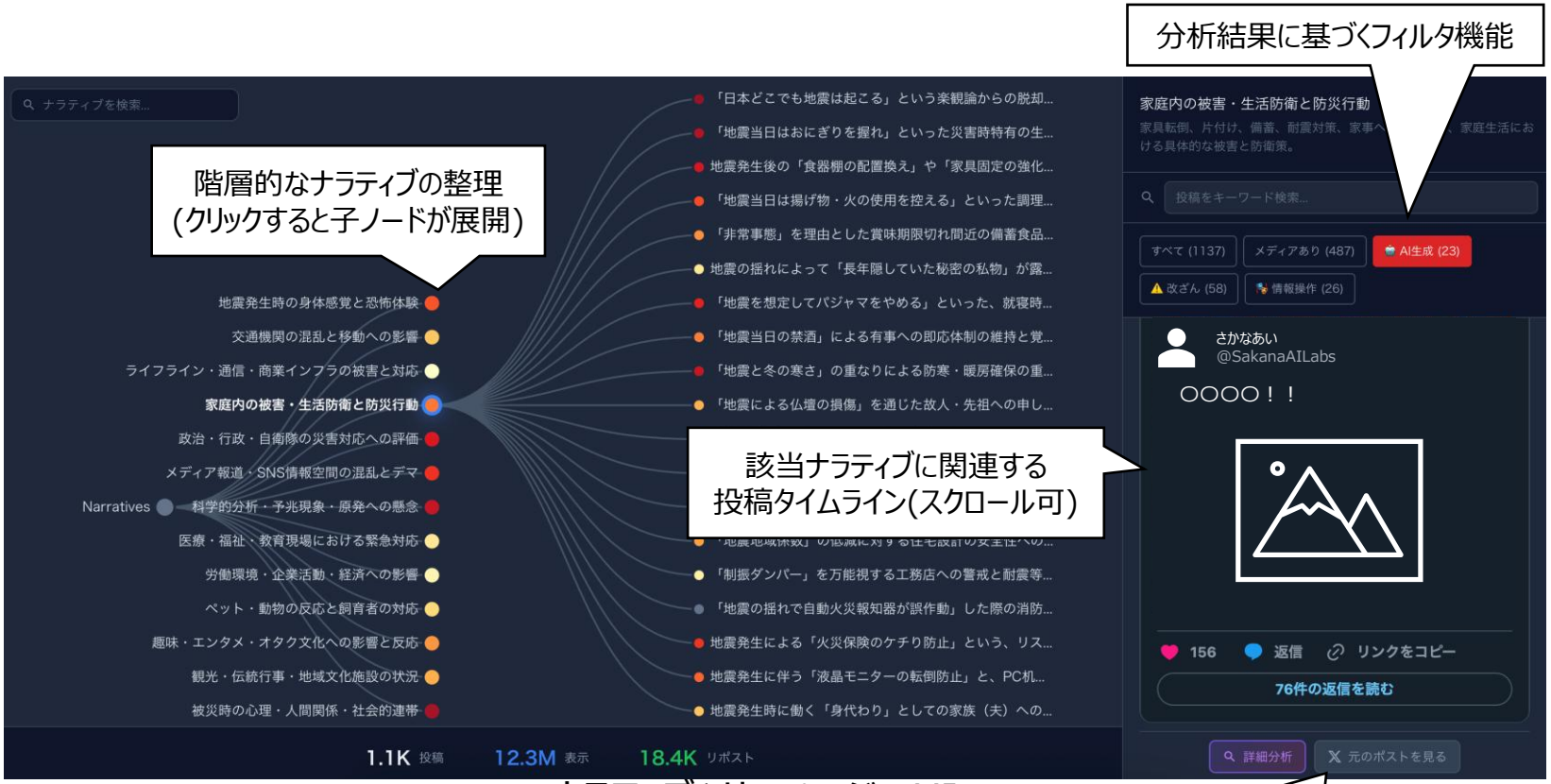




3-2. 技術開発の個別詳細

①SNS空間の可視化 ： 階層ナラティブツリーによる情報整理

- 言論空間のナラティブ（論調）を、AIにより階層構造として可視化。ツリー上のノードを選択することで、関連投稿をタイムライン形式で即座に確認できる。
- また、AI判定（生成画像・改竄検知等）に基づくフィルタリング機能を実装し、検証が必要な投稿を効率的に絞り込み、詳細分析へとシームレスに接続する導線を確立した。（後述）

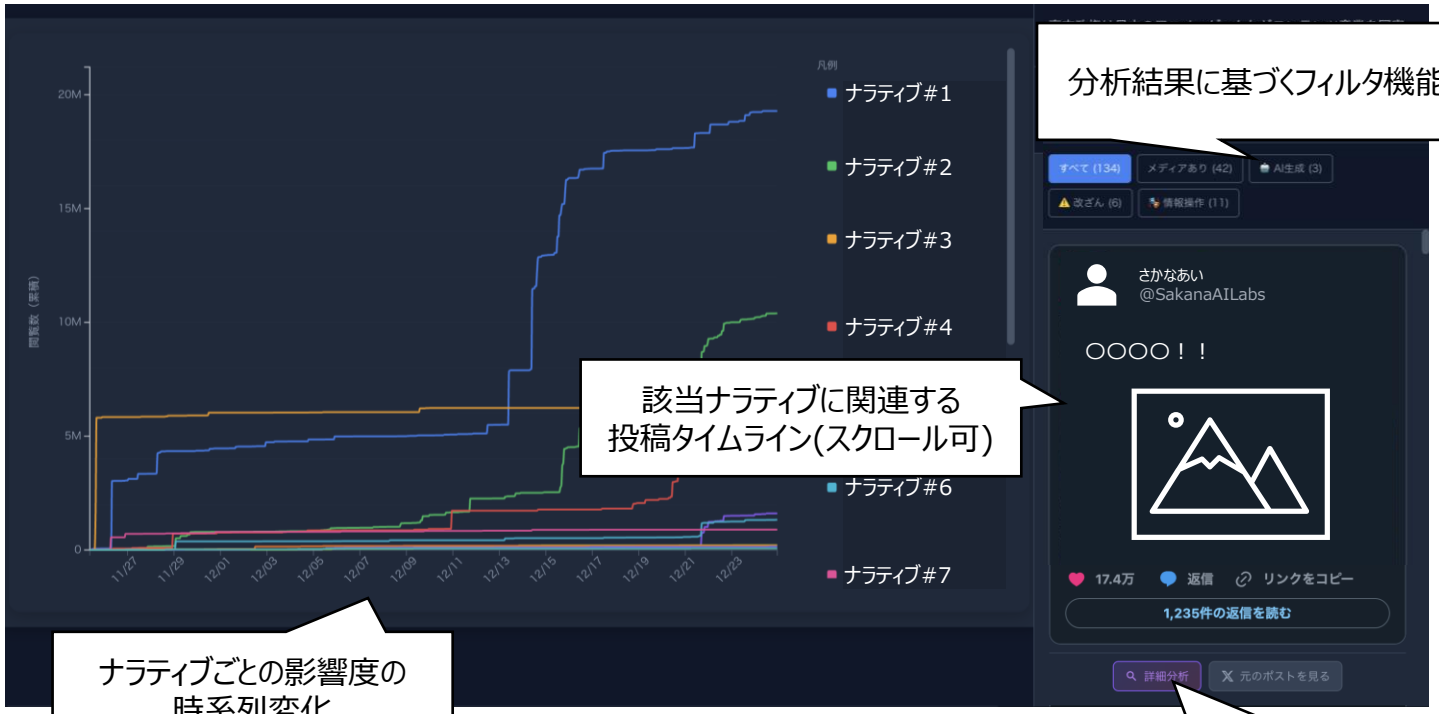


ナラティブツリーページのUI
（Xにおける災害時の投稿を対象にした例）

3-2. 技術開発の個別詳細

①SNS空間の可視化 ： 時系列変化分析

- ナラティブのノードを選択することで、その下層（子ノード）に紐づく情報の時系列変化を詳細に分析可能な機能を実装した。
- これにより、論調の急激な拡大やその要因となる投稿の特定が可能となり、真偽判定を行うべき情報の緊急度を見極めるための判断材料を提供する。

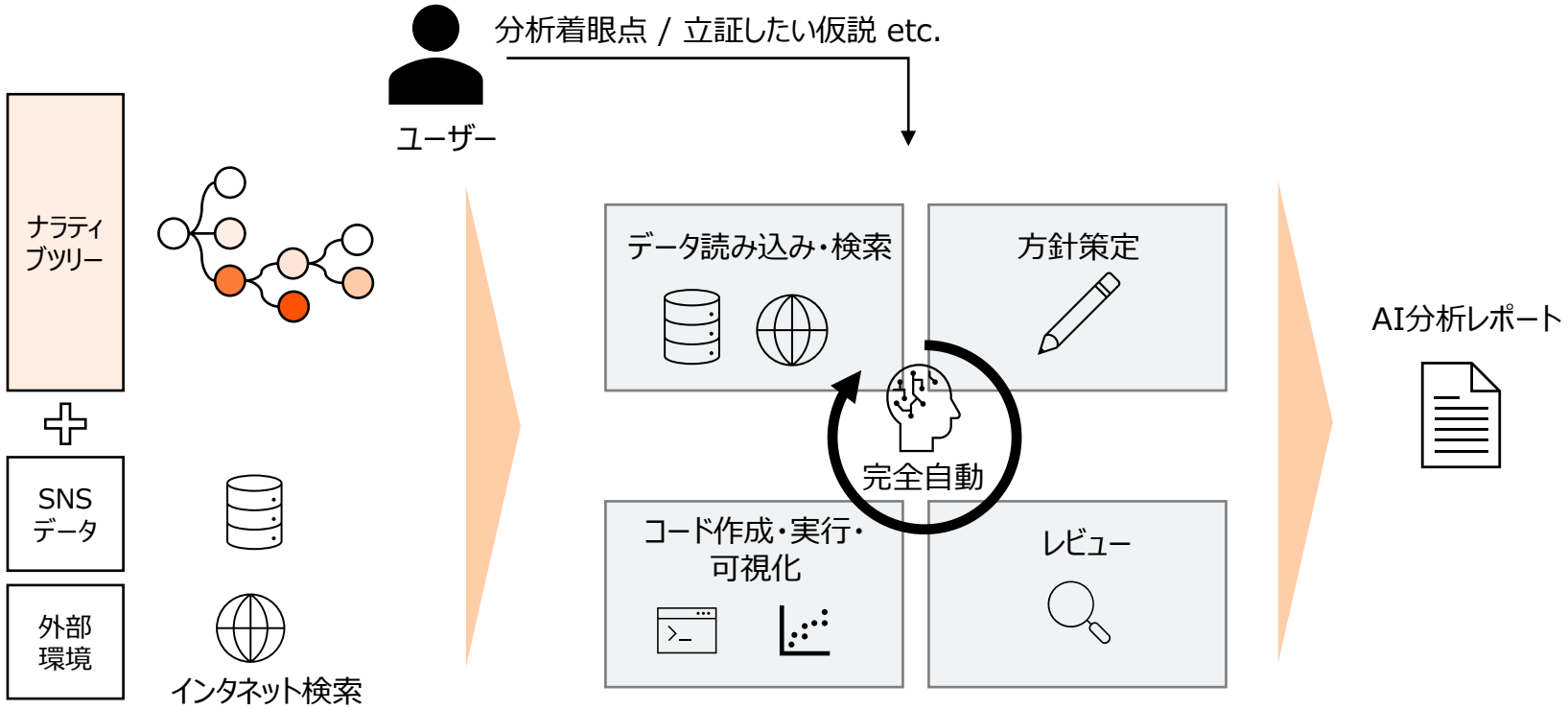


時系列分析ページのUI

3-2. 技術開発の個別詳細

①SNS空間の可視化 ： 自動データ分析および分析レポートの作成

- ナラティブツリーを構築することで、SNS空間を広範に、解像度高く言語化することができた。このデータから余すことなくインサイトを発掘するために、AIが自律的にデータを分析する機能を実装した。ナラティブデータならびに、付随するSNSの時系列変化情報やインターネット上の情報などをAIが自動抽出・分析し、ユーザーの指示に従った包括的なレポートを作成する。
- 本AI分析機能により、人間が思いつかなかったような多様な着眼点を得ること、または、人間が立案した仮説をAIに検証させるといったことが可能となる。

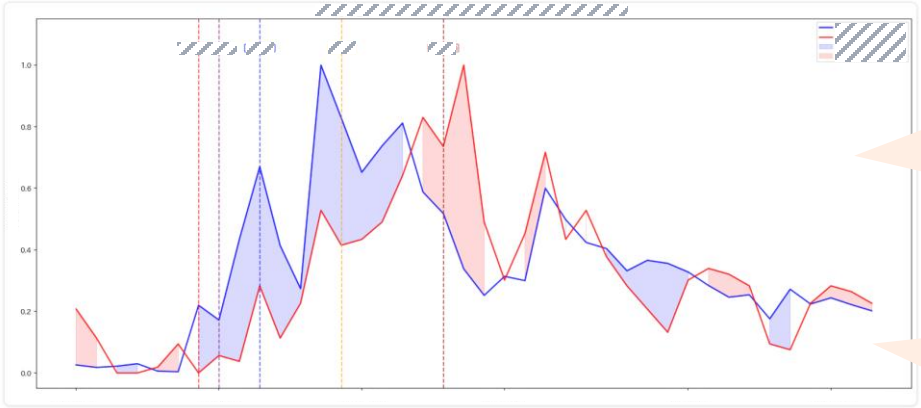


3-2. 技術開発の個別詳細

①SNS空間の可視化 ： 自動データ分析および分析レポートの作成（サンプル）

- 多角的にデータを分析し、多くの仮説を生成し、その立証、反証を行う。

生成されたレポート



分析観点に従い、２つの時系列データのギャップ分析の必要性をAIが自動で発案し、グラフ化までを実施。

関連するイベントを自律的にインターネット検索し、データとの関連性を図示。

観測結果：
〇月〇日から、△△のタイミングで、□□が先行するが、☆☆のタイミングで、××が逆転する。

作成したグラフを確認し、観測結果を言語化

仮説１ ：xxx

■仮説の主張：xxx

■根拠データ：xxx

■反証：xxx

■仮説の確度：xxx

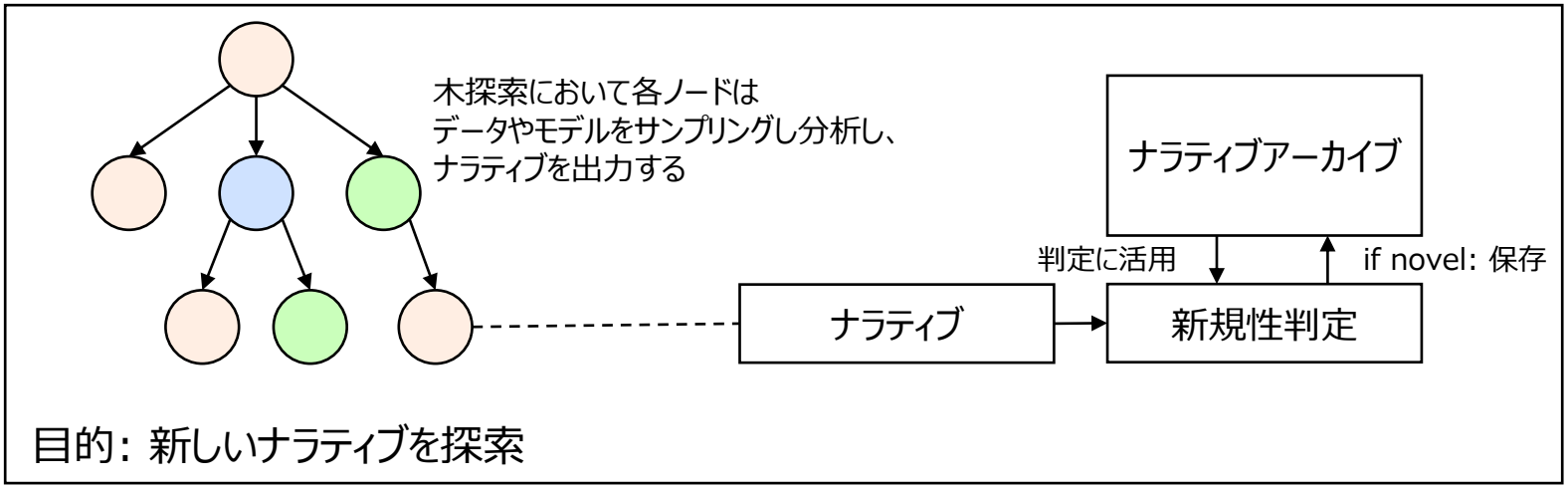
仮説の立案を行い、その仮説の根拠、立証、反証、確度を定量的に考察。

3-2. 技術開発の個別詳細

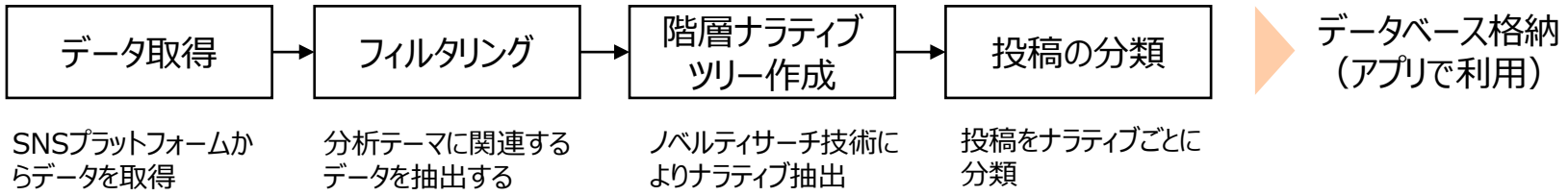
①SNS空間の可視化 ： 多様なナラティブの抽出アルゴリズムの開発

- 階層ナラティブツリーにおいては、投稿の表層的なタグ付けやカテゴリ分類に留まらず、社会的波及力を有する具体的な論調（ナラティブ）を対象とし、その深層構造まで詳細な分析を行う。
- 本機能の実現にあたり、AIが自律的に重要情報を探索する「ノベルティサーチ」技術を独自開発した。これにより、データ入力上の制約を克服し、膨大な投稿群の中から真に重要な文脈を精緻に抽出する手法を確立している。

ノベルティサーチ技術の模式図
（Xの投稿データを再帰的にサンプリングし、新規性の高いナラティブを効率的に探索・抽出する）



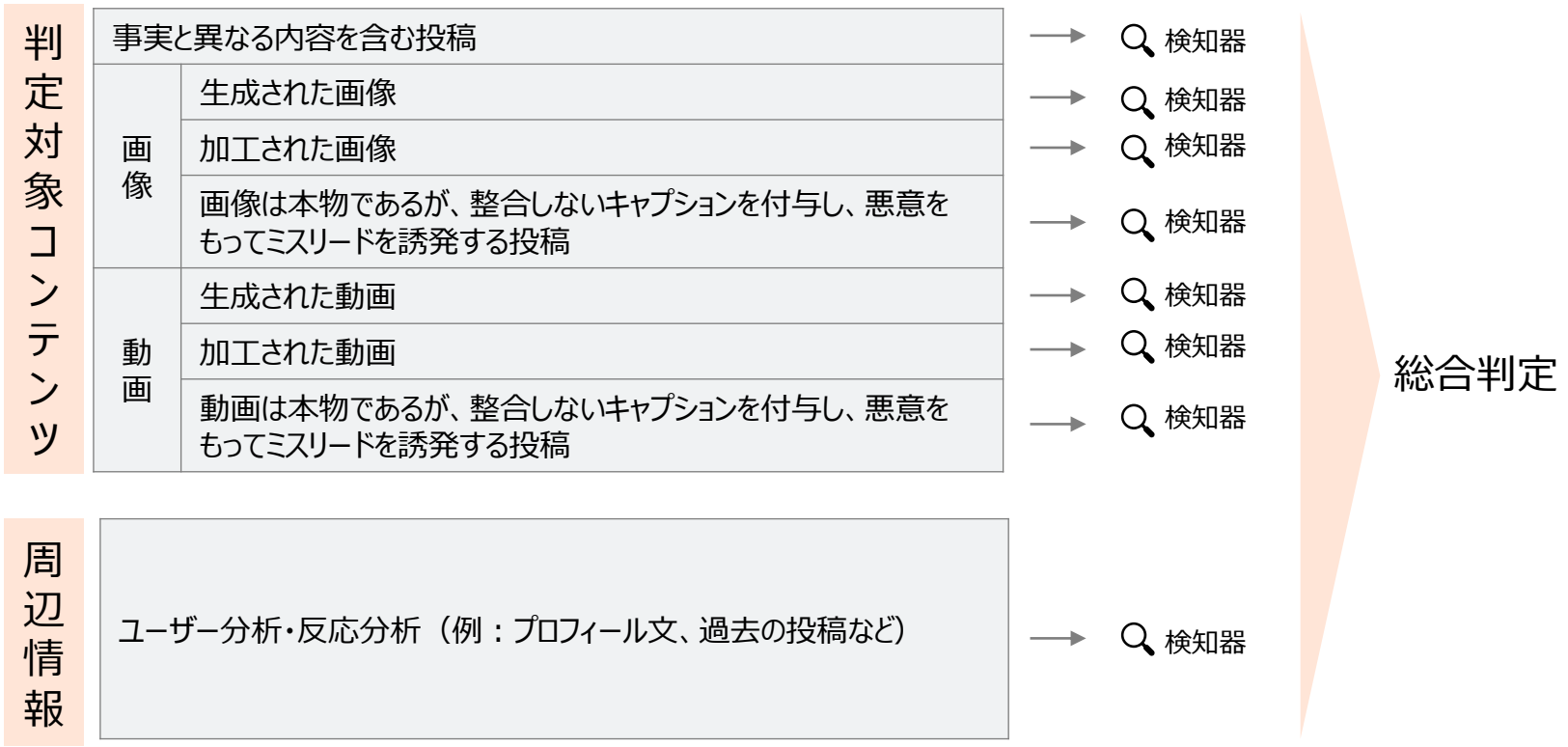
データ取得からアプリで利用するデータ形式までの一連の変換プロセスも自動化



3-2. 技術開発の個別詳細

②総合的な偽・誤情報判定 ： アプローチ概要

- 偽・誤情報の類型ごとに有効な検知器を開発し、それらを統合するアプローチを採用した。さらに、コンテンツ分析に加えユーザープロファイリングも実施し、実務者を多面的に支援するシステムを構築した。
- この多面的・多段的なアプローチは、単一のAIスコアを提示することに比べて、各検知器のスコアとその判定根拠を知ることができるため、AIの判断プロセスを十分に説明することができ、とくに情報の信憑性を厳格に取扱うユーザーにとって有益なアプローチである。



3-2. 技術開発の個別詳細

②総合的な偽・誤情報判定： アプローチ概要（UIイメージ）

- 後述する多角的な観点に基づく分析結果は、本アプリケーション上で統合レポートとして閲覧・確認が可能である。個別の検知結果をシステム上で一元的に可視化することで、総合的な判定状況を俯瞰できる仕様とした。

詳細分析ページのUI (主張分析の例)

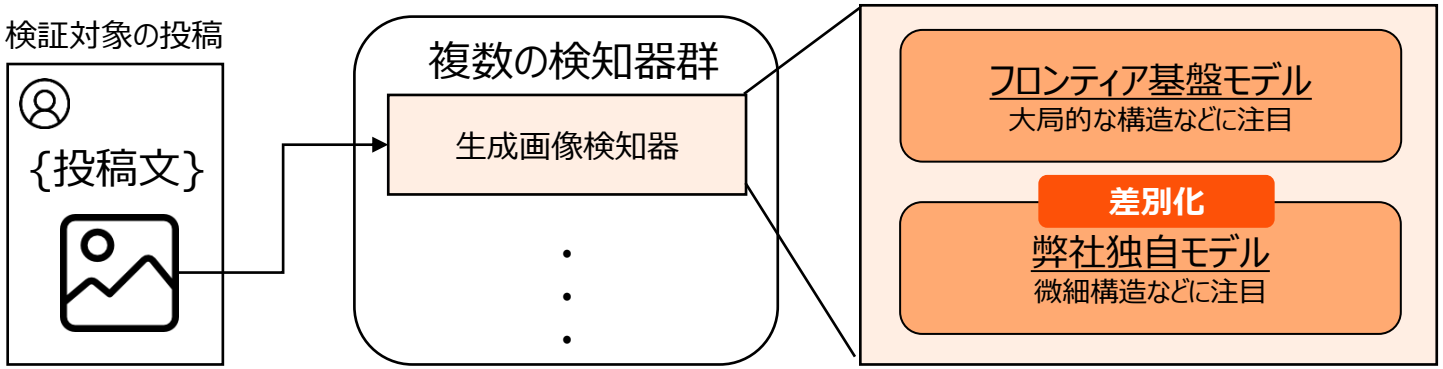
3-2. 技術開発の個別詳細

②総合的な偽・誤情報判定： 画像、動画の生成跡、加工跡の検知

- 画像・動画の生成・加工痕検知において、「フロンティア基盤モデル※」と「弊社独自モデル」の併用アプローチを採用した。生成元モデルごとの検知精度の偏りを、両者の相互補完（詳細はP.48参照）によって解消し、あらゆる生成物に対し安定した検知を実現している。

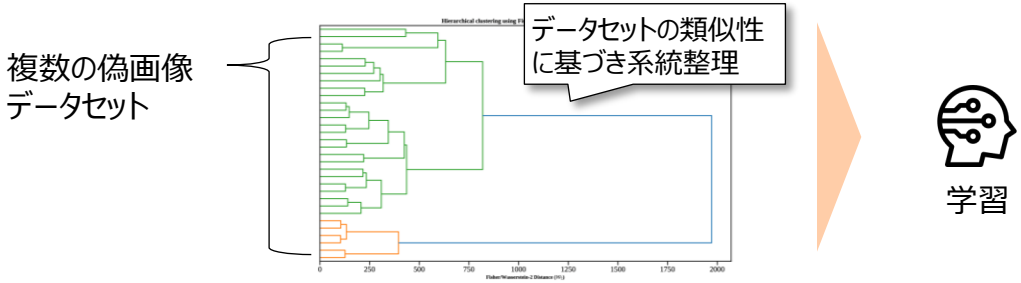
※ フロンティア基盤モデルとは、GPT-5やGemini-3など、世界最高水準の性能を有する最先端の汎用大規模言語モデルのこと

フロンティア基盤モデルと弊社独自の生成画像判定モデルの併用イメージ



弊社独自モデルの特徴

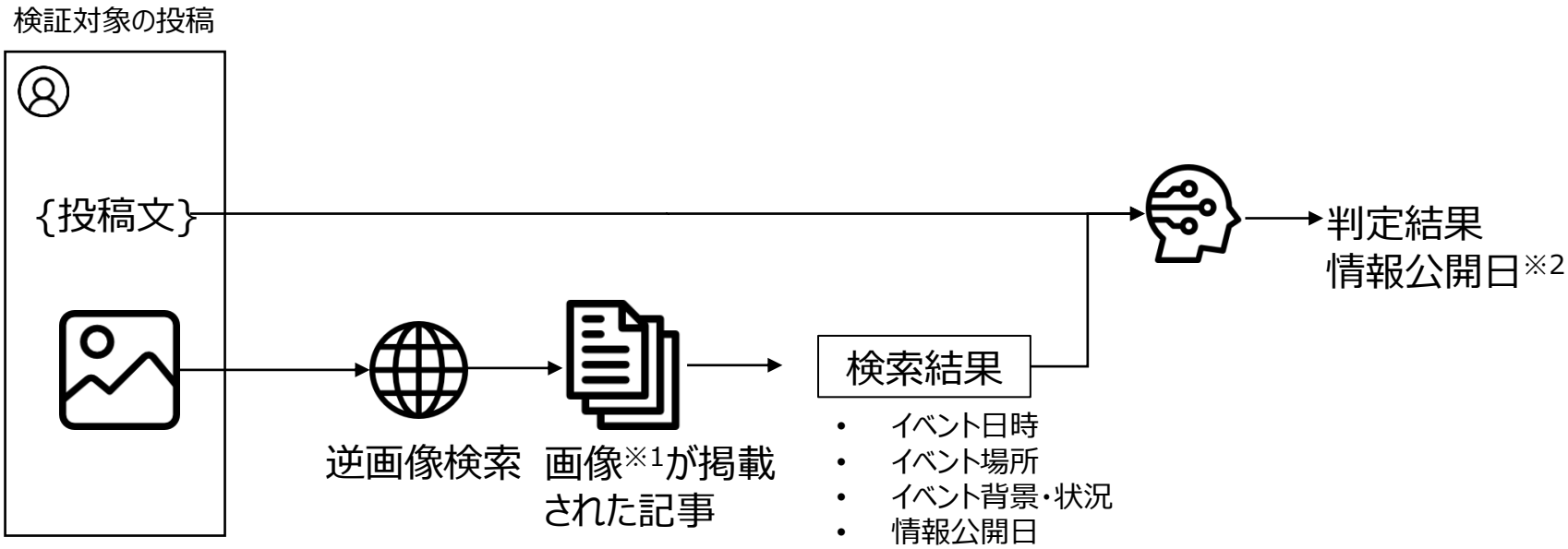
様々な偽画像に対応するために、複数のデータセットを用意し、それらを類似度に基づき体系的に整理したうえで学習している



3-2. 技術開発の個別詳細

②総合的な偽・誤情報判定： 画像、動画のすりかえ検知

- 画像、動画を逆画像検索して（動画の場合はフレームごとに）、その画像、動画の時間、場所、背景をインターネットで調査し、投稿文と照合を行う。
- なお、可能なものに関しては逆画像検索結果が表示されているページの公開日時を取得し、投稿時点より前からあるかどうかを判定する

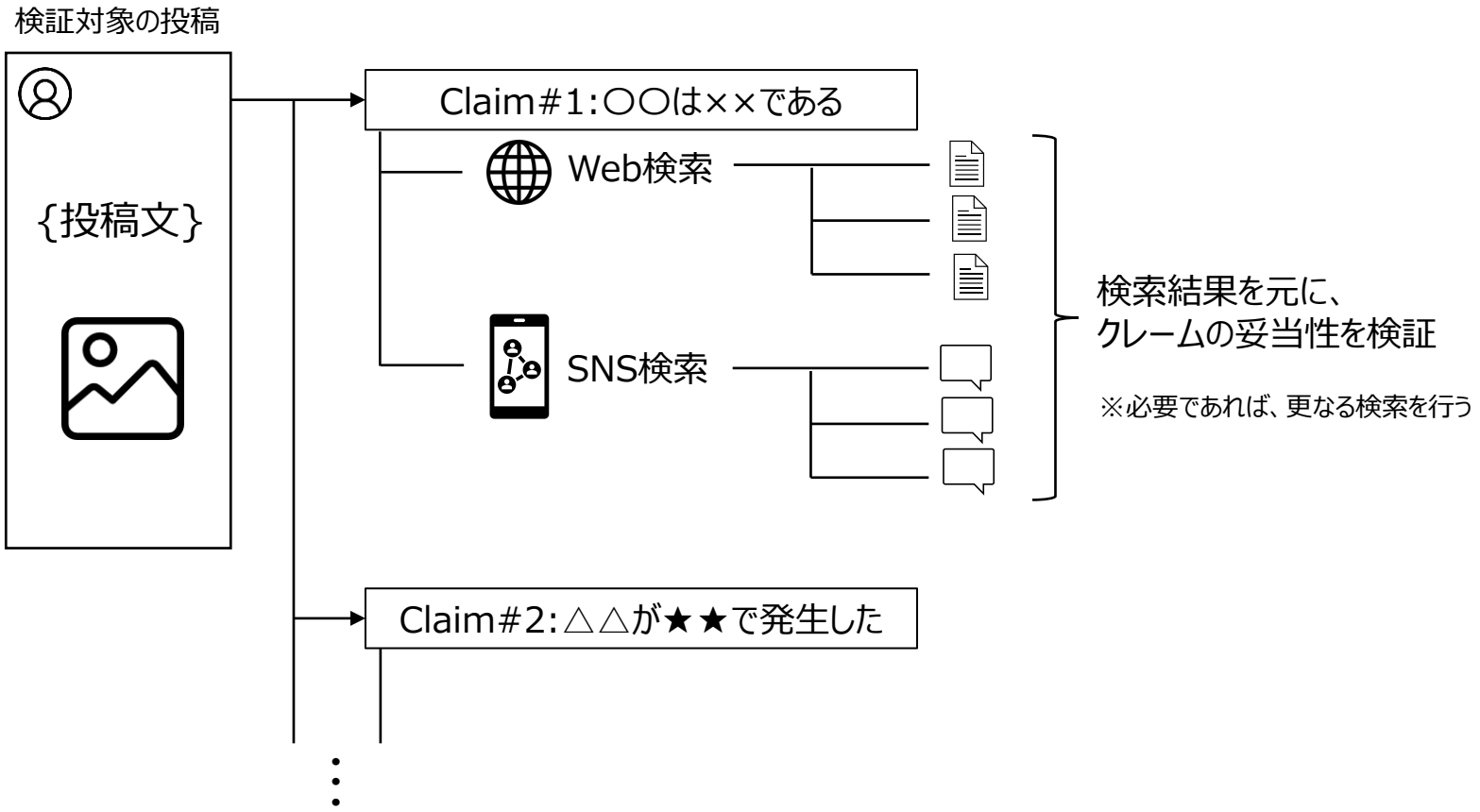


※1 完全一致、部分一致、類似一致画像を対象とする。
※2 データ取得出来た場合

3-2. 技術開発の個別詳細

②総合的な偽・誤情報判定：真偽判別

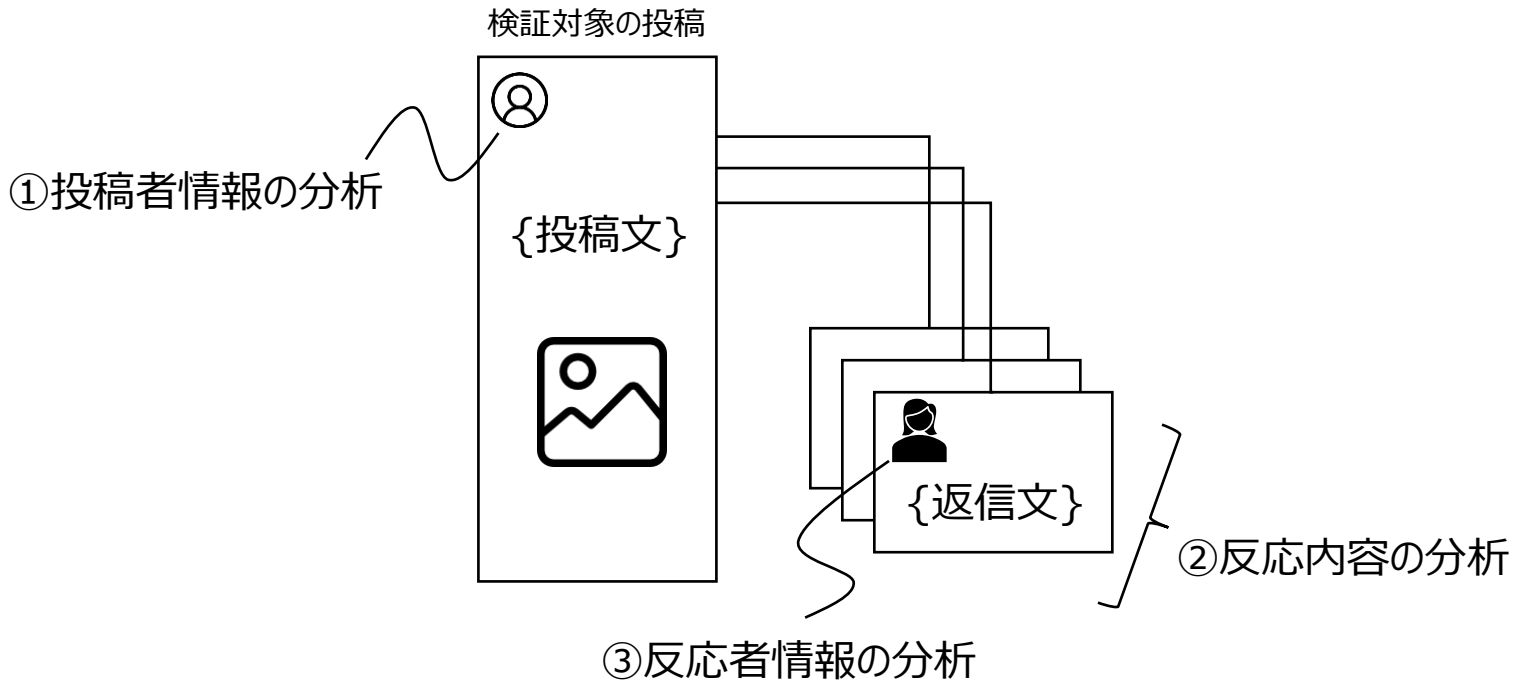
- 画像・動画および投稿文を分析し、その中から検証可能なクレーム(主張)を抽出し、それらに対する検証を行うための情報をWeb検索とSNS検索を組み合わせ検索する。これにより、災害など速報性の高い内容に関してもアプローチを行うことが可能となる。
- クレームの抽出、検索クエリの作成、検索結果の分析などは全てAIエージェントが自律的に行う。



3-2. 技術開発の個別詳細

②総合的な偽･誤情報判定： ユーザー分析･反応分析 (周辺情報)

- 検証対象の投稿に関する周辺情報をAIを用いて分析整理することで、補助情報を提供する。
- 具体的な周辺情報としては、①投稿者の過去投稿やプロフィールなどの要約・整理、②反応内容の要約・整理、③反応者の挙動確認(機械的な行動等)といった機能を開発実装した。



3-2. 技術開発の個別詳細

③対策案の立案：SNS空間のシミュレーション

- Agent Based Modeling (ABM) とは、人間の振る舞いや判断を再現したエージェント同士を相互作用させ、社会現象を模擬する技術である。
- 本実証事業では、このABMをAIエージェントと組み合わせて使うための弊社独自フレームワークである「Shachi※」を活用しSNS空間を再現することで、偽・誤情報に対するカウンター発信の有効性をシミュレーションの枠組みを構築した。

※ https://github.com/SakanaAI/shachi

ABMを使ったSNS空間のシミュレーションイメージ

3-2. 技術開発の個別詳細

③対策案の立案：ペルソナの設計

- ABM（エージェントベースドモデリング）を用いたSNSシミュレーションにおいて、実在のアカウントを模した精緻なペルソナの構築は極めて重要である。
- 本実証事業では、シミュレーション対象に関連する投稿データを収集し、階層ナラティブツリーと同様のノベルティサーチ技術を応用することで、網羅的かつ高解像度なペルソナ生成を実現した。
- これらペルソナを実装したAIエージェントのシミュレーション空間内での挙動を分析することにより、偽・誤情報に対するカウンター発信（対抗言説）の有効性を定量的に検証することが可能となる。

階層的に様々なペルソナを幅広く、解像度高く作成

SNSデータ
シミュレーション対象のデータを取得

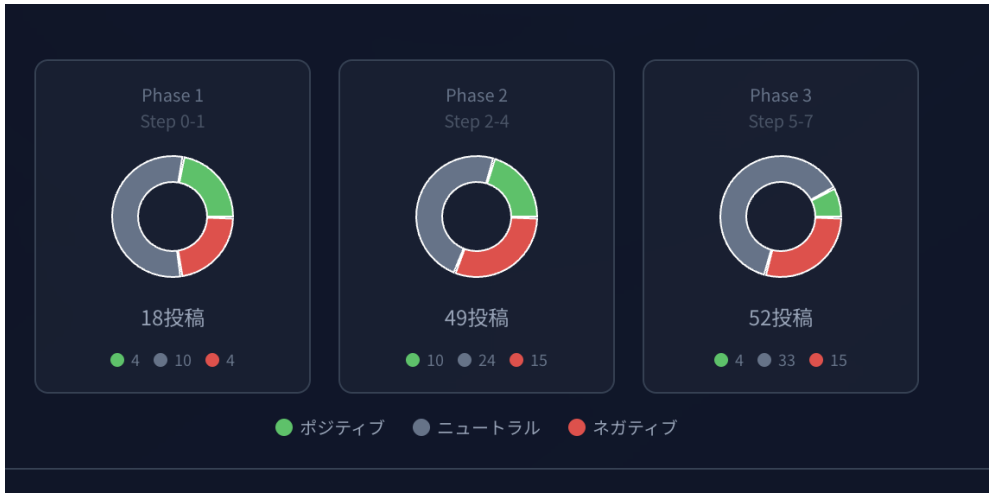
3-2. 技術開発の個別詳細

③対策案の立案：シミュレーション結果の分析

- カウンター発信が「誰に」「どのように」影響を与えたかという微視的（ミクロ）な分析が可能。これにより、特定のナラティブが有効に機能した「要因」と「背景」を特定し、次なるアクションの精度を高めることができる。

シミュレーションのスナップショット
（一部記載表現を変更）

複数のカウンター発信の効果を比較（例）

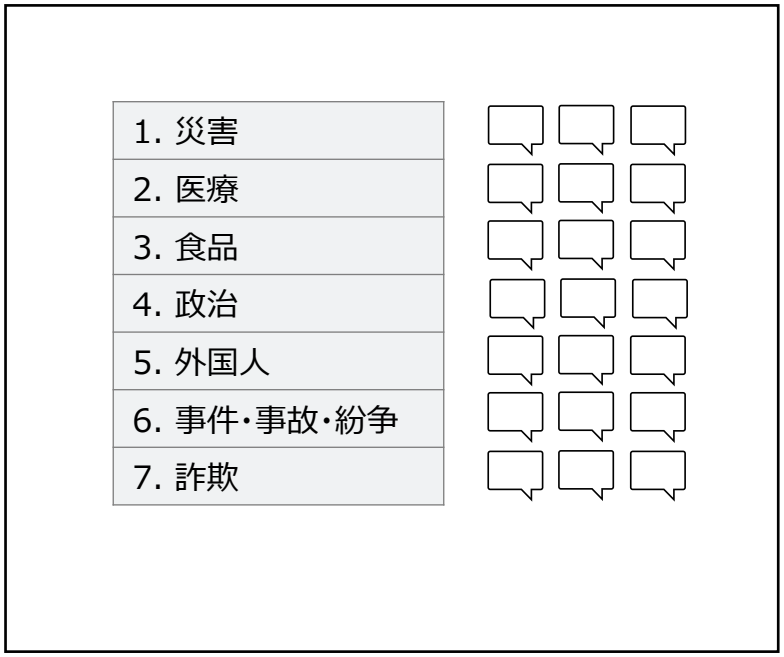


目次

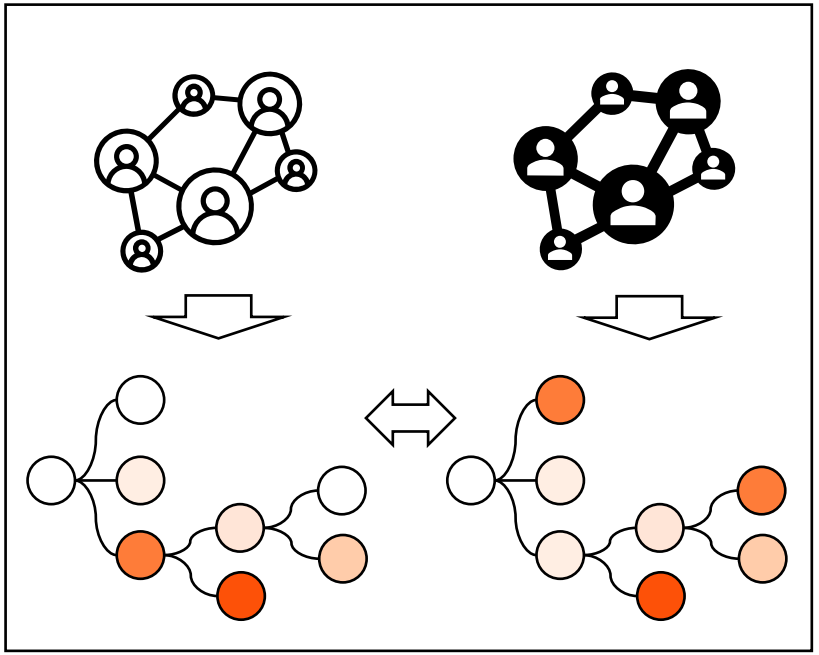
4-1. 検証及び調査の全体像

検証及び調査に係る取組・成果の全体像

- 開発された対策技術それぞれに対して、有効性を検証し更に高めていくための取り組みを行なった。
  - 「総合的な偽・誤情報判定」に関しては、作成されたシステムの性能を評価するためのベンチマーク作成と評価を行なった。
  - 「対策の立案」に関しては、シミュレーションの現実性の評価を行なった。



コミュニティノートを活用した評価ベンチマーク作成と評価



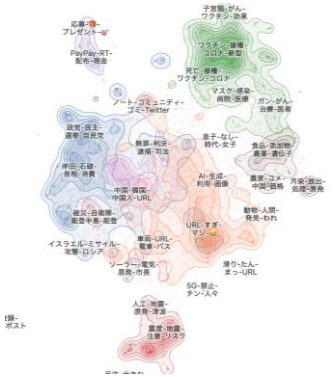
SNSシミュレーションの評価

4-2. 検証及び調査の個別詳細

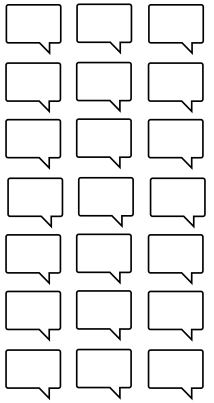
ベンチマークの作成方法

- 開発した偽・誤情報判定モデルの精度評価を行うため、評価用ベンチマークの策定を実施した。
- まず、評価対象となるトピックを選定するにあたり、Xのコミュニティノートを分析し、頻出テーマを定量的に確認し、カテゴリを選定した。策定したカテゴリを有識者にヒアリングし、実務で重視されているものがカバーされていることや乖離が無いことを確認頂いた。（コミュニティノートの分析の詳細はP45-46に記載）
- データセットの構築においては、コミュニティノートが付与されたXの投稿群から、偽・誤情報の類型および決定したテーマに合致するものを収集・抽出した。これにより、実際のプラットフォーム上で流布されている投稿で構成された、実効性の高いベンチマークデータセットを作成した。

コミュニティノート分析　　有識者ヒアリング　　社会的影響度が高いテーマに関するベンチマークデータセット



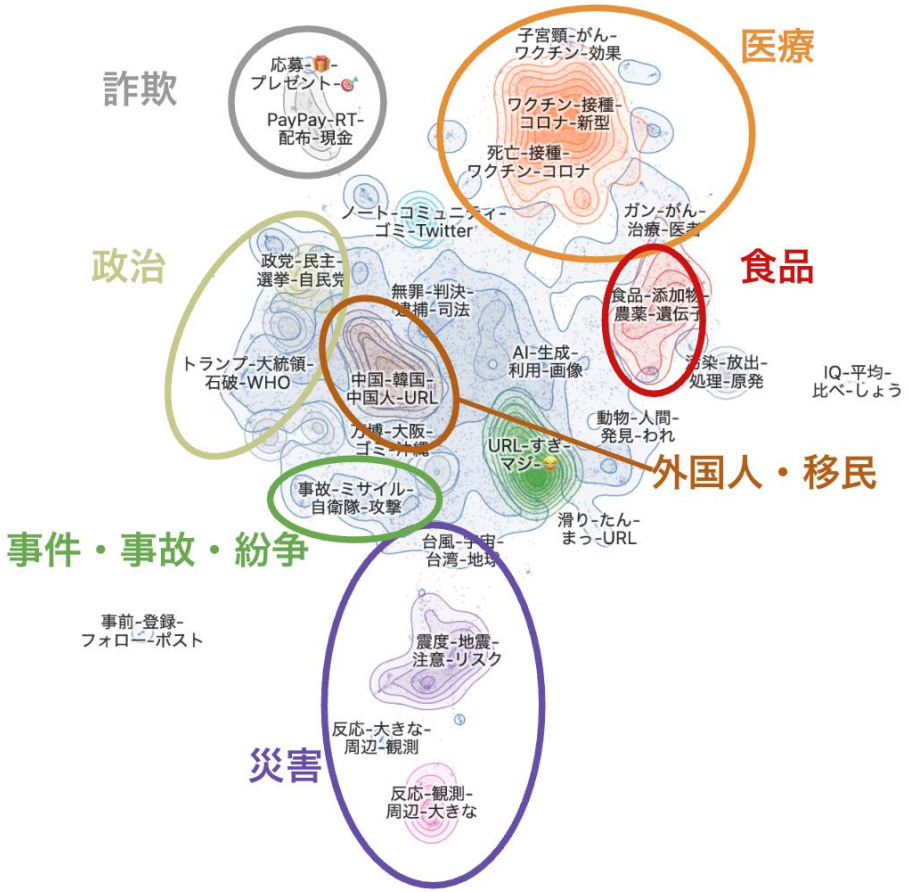
1. 災害
2. 医療
3. 食品
4. 政治
5. 外国人
6. 事件・事故・紛争
7. 詐欺



4-2. 検証及び調査の個別詳細

ベンチマークの作成方法（カテゴリわけの詳細）

- コミュニティノートが付与された投稿を、AI技術を用いて内容の類似度に基づきマッピング・可視化した。
- この分布状況に基づき、偽・誤情報の「深刻度」や「流通量」を総合的に評価し、優先的にベンチマーク対象とすべきカテゴリ案として、「災害」、「医療」、「食品」、「政治」、「外国人」、「事件・事故・紛争」、「詐欺」の７つを選定した。



4-2. 検証及び調査の個別詳細

ベンチマークの作成方法 (Xコミュニティノートの詳細)

- Xのコミュニティノートには様々なタグが付与されている。それらの情報を活用することで、どのような種類の偽・誤情報かを判断し、大量のコミュニティノートの中から多様な評価用ベンチマークデータセットを作成するシステムを構築した（作成したベンチマークはすべて人手でチェックしている）

コミュニティノートに付与されている情報例

noteId	ノートの一意ID
noteAuthorParticipantId	ノート作成者の匿名ID
createdAtMillis	作成時刻（エポックミリ秒）
tweetId	対象ツイートのID
classification	判定カテゴリ（誤情報等）
believable	信じやすさ指標
harmful	有害性指標
validationDifficulty	検証の難易度指標
misleadingOther	誤解理由=その他
misleadingFactualError	事実誤り
misleadingManipulatedMedia	改ざんメディア
misleadingOutdatedInformation	古い情報
misleadingMissingImportantContext	文脈不足
misleadingUnverifiedClaimAsFact	未検証主張を事実扱い
notMisleadingOther	誤解なし=その他
notMisleadingFactuallyCorrect	事実に基づく
notMisleadingOutdatedButNotWhenWritten	投稿当時は正確
notMisleadingClearlySatire	明確な風刺
notMisleadingPersonalOpinion	個人的意見
trustworthySources	信頼できる情報源有無
summary	ノート本文の要約
isMediaNote	メディア専用ノートか

4-2. 検証及び調査の個別詳細

作成したベンチマークデータセット

- それぞれの偽・誤情報の類型と投稿テーマについてポジティブ（偽・誤情報）とネガティブ（本物の情報）を収集し、以下の通りベンチマークデータセットが作成された。

上：ポジティブ件数（偽・誤情報） 下：ネガティブ件数（本物）			カテゴリ							合計
			1.災害	2.医療	3.食品	4.政治	5.外国人	6.事件・事故・紛争	7.詐欺	
偽・誤情報の類型	ファクトチェック		15	8	7	7	6	0	6	49
			15	8	8	8	7	0	5	51
	画像	生成された画像の検知	4	0	0	12	2	5	0	23
			11	13	10	14	5	8	4	65
		加工された画像の検知	6	14	6	33	1	11	9	80
			11	13	10	14	5	8	4	65
		文脈操作	1	1	0	13	0	0	3	18
			11	13	10	14	5	8	4	65
	動画	生成された動画の検知	4	0	0	5	4	5	0	18
			12	0	0	3	0	0	0	15
		加工された動画の検知	0	0	0	6	0	4	2	12
			12	0	0	3	0	0	0	15
		文脈操作	2	0	1	9	2	9	0	23
			12	0	0	3	0	0	0	15

4-2. 検証及び調査の個別詳細

性能評価結果のまとめ

- 動画の加工および文脈操作に関する精度は70%程度となったが、その他の真偽判定、画像（生成・加工・文脈操作）、動画（生成）に関しては、85%～90%という高い精度での偽・誤情報識別を実現した。※1
- また、処理速度についても当初計画の「1件あたり5分」を下回る時間を達成している。
- 本報告書の第5章では、これらの性能特性を持つ検知器を組み合わせた「統合的検知モデル」の有用性について、有識者レビューの結果を報告する。

			ベンチマーク件数		性能	速度
			ポジティブ	ネガティブ	Balanced Accuracy	
偽・誤情報の類型と各検知器	真偽判定		49	51	0.88	約2分
	画像	生成された画像の検知	23	65	0.89※2	約10秒
		加工された画像の検知	80	65	0.85	約10秒
		画像は本物であるが、整合しないキャプションを付与し、悪意をもってミスリードを誘発する投稿	18	65	0.91	約2分
	動画	生成された動画の検知	18	15	0.86※2	約30秒
		加工された動画の検知	12	15	0.72	約30秒
		動画は本物であるが、整合しないキャプションを付与し、悪意をもってミスリードを誘発する投稿	23	15	0.77	約2分

※1 真偽判定する際はインターネット検索等を利用するが、その際に、当該ベンチマークで使用したコミュニティノートやSNSデータは参照しない。（データリーク防止）
※2 本紙で示す生成画像・動画の検知精度は、フロンティアVLMを用いて計測した。フロンティアモデルと弊社独自モデルの使い分けに関する見解は次頁を参照

4-2. 検証及び調査の個別詳細

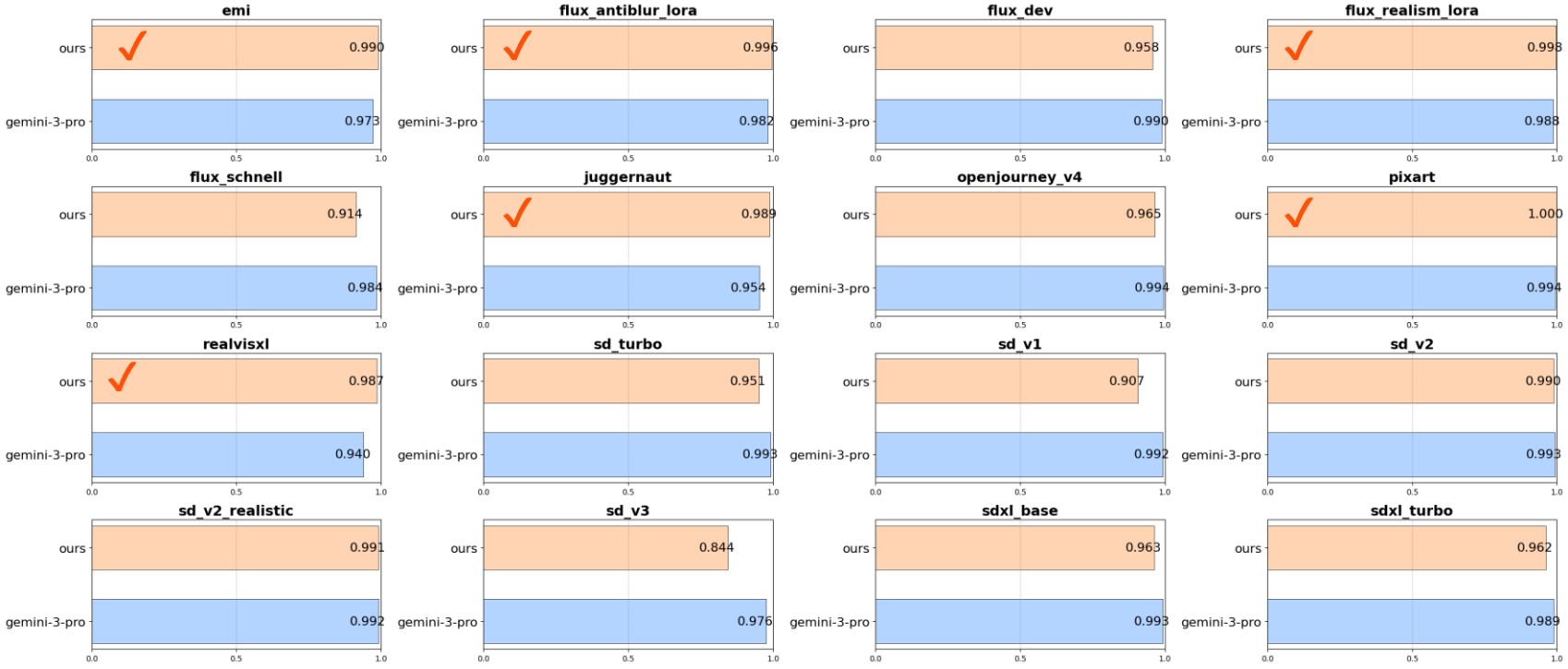
生成画像検知のための特化型モデルの有用性と適用の方向性

- 前頁のベンチマーク評価で使用したフロンティアVLMは強力な検知能力を有する一方で、特定の生成モデルに対しては「死角」も存在する。弊社独自モデルは、この「死角」となる領域において有効なスコアを示しており、VLMの検知漏れをカバーすることができる。また、VLMと比較して圧倒的に軽量であるため、低コストかつ短時間での大量処理が可能。この特性を活かし、全量データを高速に処理する「一次スクリーニング」として機能させつつ、両者を組み合わせることで、高効率で見落としリスクを抑えた検知システムが構築可能である。

弊社独自モデルとフロンティアVLM（gemini-3-pro）の比較

COCO Datasetを用いたベンチマーク評価（Real / Fake 各5,000件 × 16種類の画像生成モデルで網羅的に検証）

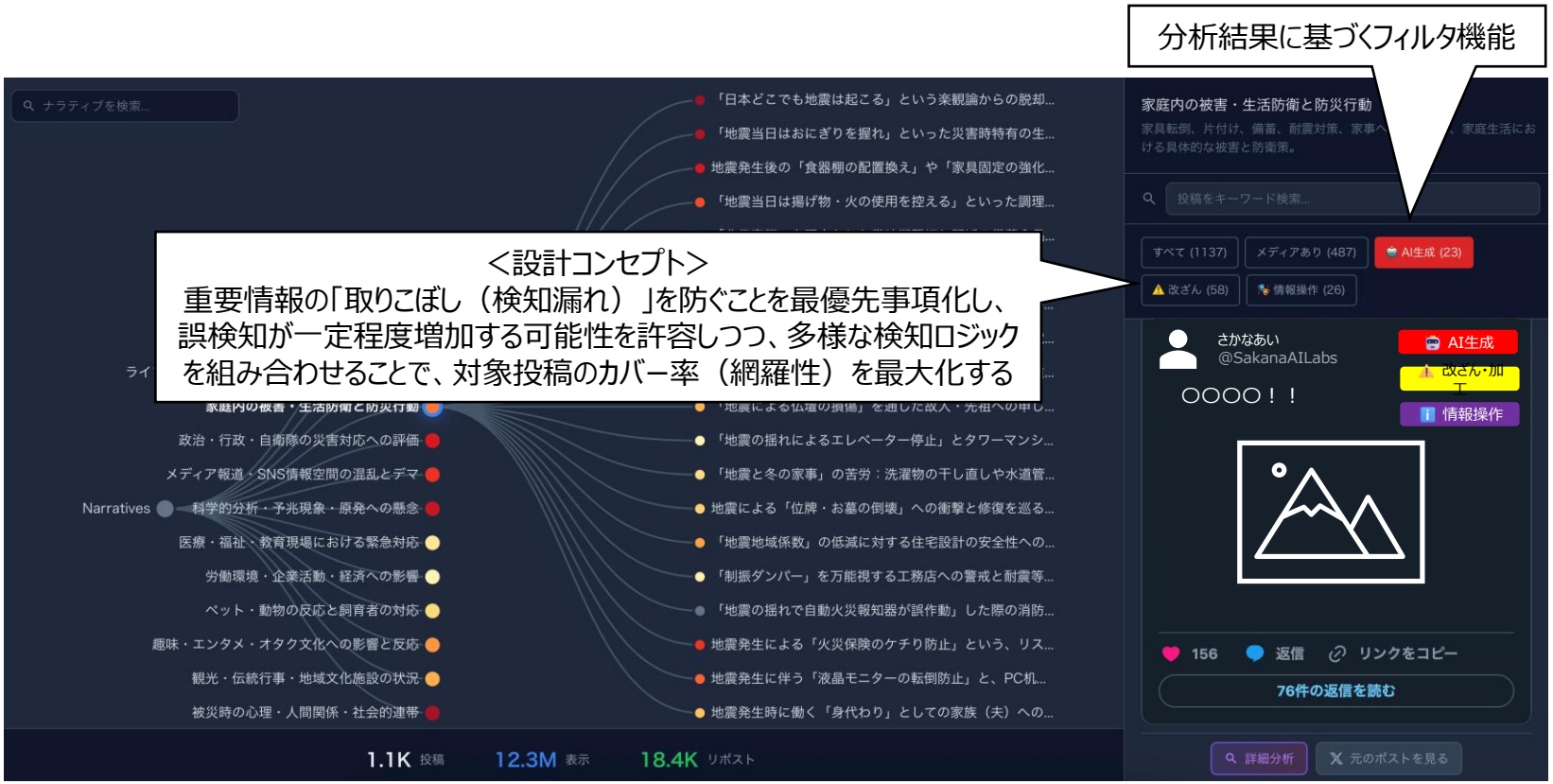
✓：独自モデル > gemini-3-pro



4-2. 検証及び調査の個別詳細

プロダクトへの反映

- 偽・誤情報判定の現場目線を踏まえ、ナラティブツリーのUIにおいて偽・誤情報の見落としをできる限り低減することを狙い、フロンティア基盤モデルと弊社独自モデルを併用する仕組みを採用することとした。



ナラティブツリーページのUI

4-2. 検証及び調査の個別詳細

費用見積もり

- P47に示す性能評価した詳細分析機能について、それら機能の処理コストは分析1件あたり約100円未満で実現可能（計画値は300円/件）。また、本実証システム一式のサーバー費用は約15万円程度から導入可能と試算（導入先企業の規模に応じて変動）

合計 （含） ファクトチェック 画像 生成検知 画像 加工検知 画像 文脈操作 動画 生成検知 動画 加工検知 動画 文脈操作	**<100円 / 件（※1）**

※1　本試算にはデータ取得費用は含めていない。

費用項目		月額（概算※2）
サーバー費用	アプリケーション、モデルサーバーを含む	15万円

※2　ユーザー数2~3人規模のファクトチェックチームを有する企業様を想定しているが、ユーザー規模（リクエスト負荷）に応じて変動する点に留意。

4-2. 検証及び調査の個別詳細

SNSシミュレーションの実験（実験条件）

実験目的

- 構築したSNSシミュレーションの社会反応の再現性および妥当性の検証
- 独自のペルソナ開発手法を用いた、カウンター発信の微細な表現差（トーン・文言）に対する感度の評価

実験方法

地震発生時のデマ拡散とそれに対するカウンター発信をペルソナに投下し反応を調査

デマ情報

ダムで微量の漏水が確認されているらしい。老朽化が進んでいるのでは。

カウンター発信

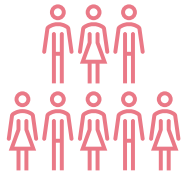
A　ダムの微量な漏水に関しまして、**専門家による調査**を実施しましたが、異常は確認されませんでした。

B　ダムの微量な漏水に関しまして、**外観の部分調査**を実施しましたが、異常は確認されませんでした。

ペルソナ１
NVIDIA Nemotron※

※ https://huggingface.co/datasets/nvidia/Nemotron-Personas-Japan

ペルソナ２
本手法



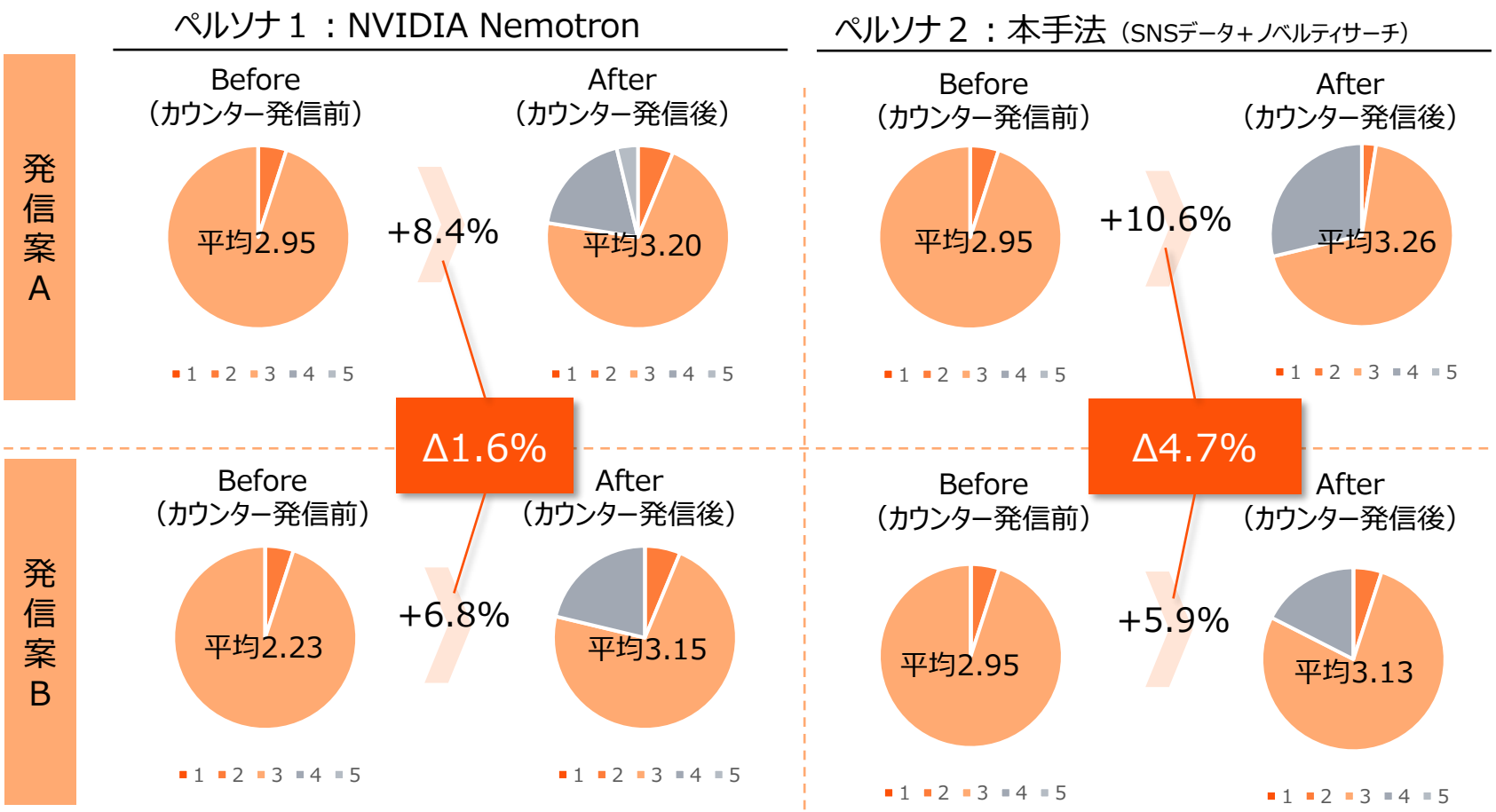
1部文言が異なるカウンター発信A/Bに対してそれぞれのペルソナの反応の感度を測定
（実務で求められる解像度で検出できるか）

4-2. 検証及び調査の個別詳細

SNSシミュレーションの実験（実験結果1/2）

- カウンター発信の前後で、デマ情報に係る信頼度をアンケート聴取し、その差分を評価。
- 発信案AとBの微妙なニュアンスの違いに対して、ペルソナB（本手法）ではより大きな差分を確認出来た。

ダムで漏水が発生していることを
5：信用していない
4：どちらかというと信用していない
3：どちらともいえない
2：どちらかというと信用している
1：信用している



4-2. 検証及び調査の個別詳細

SNSシミュレーションの実験（実験結果2/2）

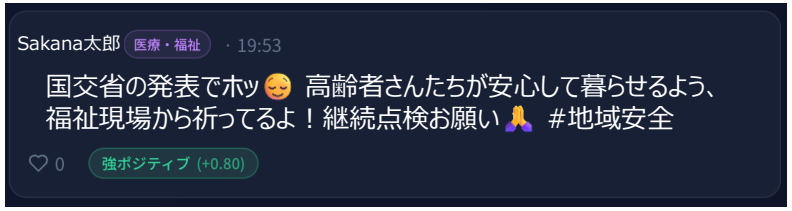
- ペルソナ１では“素直”で”画一的”な反応が多く、発信案AとBの違いがもたらすリスクを補足できていない。
- 一方で、ペルソナ２（SNSデータ＋ノベルティサーチを使った本手法）はわずかな表現の違いにも感度高く反応するため、言葉の選び方レベルでの緻密なリスク検証・事前チェックを実現できる。

※以下は架空のシナリオのシミュレーション結果です。

ペルソナ１：NVIDIA Nemotron

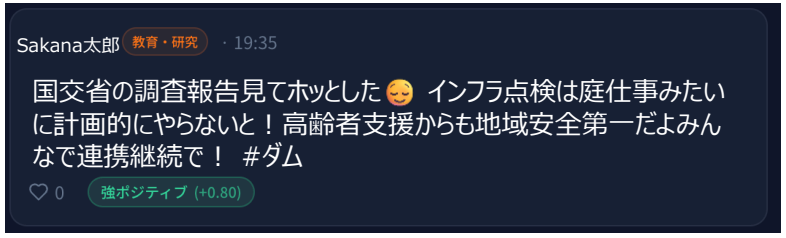
発信案A

カウンター発信を受けた投稿の一例



発信案B

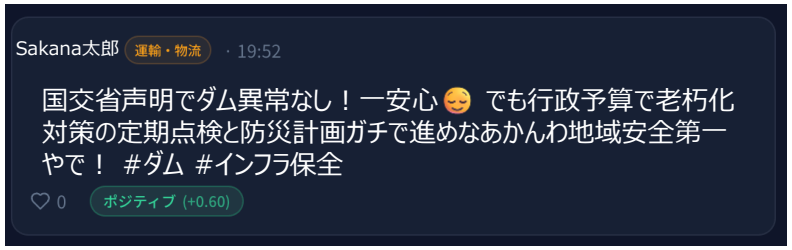
カウンター発信を受けた投稿の一例



ペルソナ２：本手法（SNSデータ＋ノベルティサーチ）

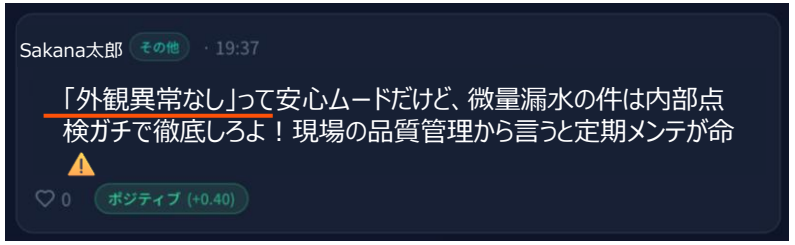
発信案A

カウンター発信を受けた投稿の一例



発信案B

カウンター発信を受けた投稿の一例



目次

5-1. 社会実装に向けた取組の全体像

社会実装に係る取組・成果の全体像

- 本実証では、有識者との連携によるシステム構築・検証に加え、安全保障/インテリジェンス分野への展開や導入先の課題分析を通じ、社会実装に向けたビジネスモデルを精緻化した。

（1）座組構築

最前線で偽・誤情報対応にあたる有識者・実務者にヒアリング・レビューを依頼

（2）ヒアリング・レビュー

段階的なヒアリング・レビューによる要件定義と有効性検証

事前・中間ヒアリング	偽・誤情報に係る業務課題を聴取し、開発指針を策定
レビュー	開発したシステムの有効性レビュー

（3）広範な活用可能性の調査

安全保障・インテリジェンス分野への展開に向け、関連研究会やシンポジウムへ参画し、潜在需要および導入領域の拡大可能性を調査

（4）導入見込み先の検討

想定される導入先の業務特性を踏まえ、社会実装時に想定される課題（ペイン）と、それに対する解決の方向性を検討

（5）ビジネスモデル

(2)～(4)の成果を踏まえ、持続可能な社会実装に向けたビジネスモデルを検討

5-2. 社会実装に向けた取組の個別詳細

(1) 偽・誤情報への取組の座組構築

- 本実証では、真に有効な偽・誤情報対策システムを構築するため、日常的に偽・誤情報への対応や課題に直面している有識者に対し、幅広く段階的にヒアリングならびにレビューを実施することとした。具体的には、以下の考え方に基づき、メディア事業者、ファクトチェック有識者団体、シンクタンクと連携を行った。

事前・中間ヒアリング（2025年9月~12月）

偽・誤情報に係る課題を聴取し、業務要件と開発方向性を決めることを目的に以下の協力団体様に対して幅広くヒアリングを実施

- **LINEヤフー株式会社様**
- **ファクトチェック有識者団体**
- **(株)Classroom Adventures**

聴取した内容

- 偽・誤情報に係る業務課題
- 偽・誤情報対策に期待するもの
- 開発途中のシステムやベンチマークに関する意見

レビュー（2026年1月~2月）

本実証システムは情報の正確性が極めて重要となるユーザーを対象としているため、評価者にはシンクタンクや学識者といった専門家を選定した。具体的には、国家安全保障に係る調査、分析、政策提言を行う中曽根平和研究所様の複数の研究員に依頼し、有識者レビューを実施。

公益財団法人
中曽根平和研究所 情報空間のリスク研究会 様

提供価値（仮説）

- **効率化**：大量の情報を真贋判定
- **標準化**：シニア、ジュニアでも同じレベルの作業ができる
- **高度化**：今までになかった分析とインサイトの抽出ができる

聴取した内容

- 本実証システムの各機能における有用性
- 業務適用に向けた要望、改善案

5-2. 社会実装に向けた取組の個別詳細

(2) 事前・中間ヒアリング結果

- 国内主要メディア事業者であるLINEヤフー株式会社様に対し、SNS投稿データの活用実態および運用課題についてヒアリングを実施した。

LINEヤフー株式会社様

- Yahoo!リアルタイム検索では、Xの投稿をリアルタイムで検索できる機能の他、トレンド、人気ポスト、SNSバズまとめを提供している
- サービスの核心である**リアルタイム性**を維持しつつ、並行して適切な**フィルタリング**を行う高度な運用が求められている。

5-2. 社会実装に向けた取組の個別詳細

(2) 事前・中間ヒアリング結果

- ファクトチェック有識者として(株)Classroom Adventure様およびファクトチェック有識者団体からご意見を伺った。

Classroom Adventure様

- **どの投稿をファクトチェックすべきかを選別するのが難しい。**偽・誤情報の数が多い一方で、ファクトチェック調査をする人員が少ないため、厳選する必要がある。**拡散度、深刻度、身近さ**という3つの観点から判断することが多い。
- ファクトチェックでは、**投稿者が誰か、という情報も有益**である。そのため、過去にファクトチェックしたアカウントかどうか、あるいは、そのようなアカウントと強い関係があるアカウントかどうかがわかると良い。
- まったく別のアプローチとして、**精度よりも速報性を重視して、AIが自動でファクトチェックを行い、その結果を公開していく**、という方向もあるかもしれない。

ファクトチェック有識者団体

- Xに加えて**YouTube ShortsやTikTok、Instagram**に言論空間が移ってきている。テキストや画像よりも**動画の重要性が高まっている。**
- Xは検索機能があるために分析しやすいが、YouTube ShortsやTikTok、Instagramは検索機能が貧弱であり、分析しにくい、という技術的課題がある。
- ファクトチェックの精度をどの程度求めるのか、という課題がある。画像や動画が改変されているかどうかはすぐに確認できるが、**事実関係の確認をするためには、周辺情報を調査する**必要があり、日数がかかる。正確性と速報性のバランスを考慮する必要がある。
- 時期を逃しても、質の高い検証記事は「将来の類似デマ」への耐性をつけるために重要であり、**高品質な記事をアセットとして蓄積することは重要。**
- ボットによる投稿かどうかがわかるとよい。

5-2. 社会実装に向けた取組の個別詳細

(2) 構築したソリューションのレビュー結果

- 中曽根平和研究所様に、開発中のアプリケーションを評価いただいた（1/3）。

			コメント	社会実装に向けた方向性
① SNS空間の可視化	評価された強み・有用性	情報空間の体系的な把握	• 従来は把握が困難であったSNS上の情報空間を、体系的に理解できるようになった。 • 膨大なSNSデータに対し、人間の力では困難だった情報の抽出・整理が容易になった。	
		実務への適用可能性（ユーザビリティ）	• ツールの使い勝手が良く、操作性が高い。 • 偽・誤情報の判定業務において、判断の支援・効率化に寄与する。	
	要望	ナラティブ分析の高度化とリスク分類	• ナラティブの分類に加えて、「感情を刺激するもの」「暴動につながる危険度が高いもの」など、リスクレベルに応じた分類・色分けが欲しい。 • 特定の対象国に見られる独特なナラティブ（歴史問題や軍国主義との紐づけなど）ができるようにしてほしい。	• ユーザーごとに特定のナラティブ分類軸があることがわかった。ユーザーごとにカスタマイズできるような導入アプローチを整備していく。 個別カスタマイズ
		拡散・増幅構造の可視化と特定	• 要因分析：ツイート急増の原因について、イベントごとの切り分けや要因分析ができるようにしてほしい。 • 安全保障上での活用を考えると、情報操作の中でも、「外国からの情報操作」であるかどうかを見極める必要がある。（例えば、Impression稼ぎのような情報操作とは切り分けたい。 • 増幅主体の特定：外国からの影響工作において、日本人インフルエンサーや生成AIアカウントによる「意図的な増幅（コメント無しリポストの連打など）」の特徴を検知したい。 • 可視化手法：リポストの中心人物や拡散のハブとなっているアカウントを特定したく、ネットワーク図を用いた拡散状況の可視化が望ましい。	• アカウントのプロファイルや挙動に着目した要因分析が期待されていることがわかった。本実証システムのユーザプロファイル分析機能を応用して、利用主体ごとにカスタマイズできるような導入アプローチを整備していく。 個別カスタマイズ • ネットワークについてはまずはデータとして取得可能性から調査する必要がある。 フィージビリティを要検討

5-2. 社会実装に向けた取組の個別詳細

(2) 構築したソリューションのレビュー結果

- 中曽根平和研究所様に、開発中のアプリケーションを評価いただいた（2/3）。

			コメント	社会実装に向けた方向性
②偽・誤情報判定	評価された強み・有用性	安全保障分野での活用への期待	• 単なるSNS分析ツールとしてではなく、「安全保障面での利用」として期待できる。	
		実務への適用可能性（ユーザビリティ）	• ツールの使い勝手が良く、操作性が高い。 • 偽・誤情報の判定業務において、判断の支援・効率化に寄与する。	
		マルチモーダルな解析能力	• テキストだけでなく、画像や動画の判定までできる点が優れている。	
	要望	扇動目的の分析（どの感情を狙っているか）	• 情報操作の狙いとして、ターゲットの「どの感情（怒り、不安など）」を扇動しようとしているのかを分析したい。	• ユーザーごとに分析切り口のニーズがあることがわかった。ユーザーごとに分析切り口をカスタマイズできるような導入アプローチを整備していく。 個別カスタマイズ

5-2. 社会実装に向けた取組の個別詳細

(2) 構築したソリューションのレビュー結果

- 中曽根平和研究所様に、開発中のアプリケーションを評価いただいた（3/3）。

			コメント	社会実装に向けた方向性
③対策案の立案	評価された強み・有用性	多様な活用シーンとインパクト	• プレバンキング（偽・誤情報の拡散防止措置）として有用であり期待できる。 • SNS空間の汚染、生成AIペルソナの実空間利用など、防御だけでなく攻撃や悪用の可能性も含めて有益である。	

5-2. 社会実装に向けた取組の個別詳細

(2) ヒアリングおよびレビューを踏まえた本実証システムの要件と今後の開発方針

事前中間ヒアリングに基づき開発した本実証システムは、有識者レビューにて高く評価された一方で、実務における分析視点の個別性も明らかとなった。この結果を受け、共通基盤をベースとしつつ、個別のニーズに柔軟に対応できるハイブリッドなシステム設計を推進する。

事前・中間ヒアリング結果

ヒアリング結果	本実証システムにおける要件
テーマごとの重要性	重要テーマをベンチマークに組み込む 済
品質の高いファクトチェック（アセットとして蓄積）	徹底的な事実確認フローと、根拠となるエビデンスの明示機能を搭載 済
Xに加えて、他のSNSプラットフォームでも偽・誤情報対策の重要性が高まる。中でも動画が重要。	まずはAPIが整備されたXで実施し、その中で動画も含めた偽・誤情報検知を行う。 済
リアルタイム性と即応的な簡易ファクトチェックの可能性	リアルタイム性が求められる事業者のサービスが存在するものの、まずは高いニーズが見込まれる“品質の高いファクトチェック”のアプローチをとった。 済

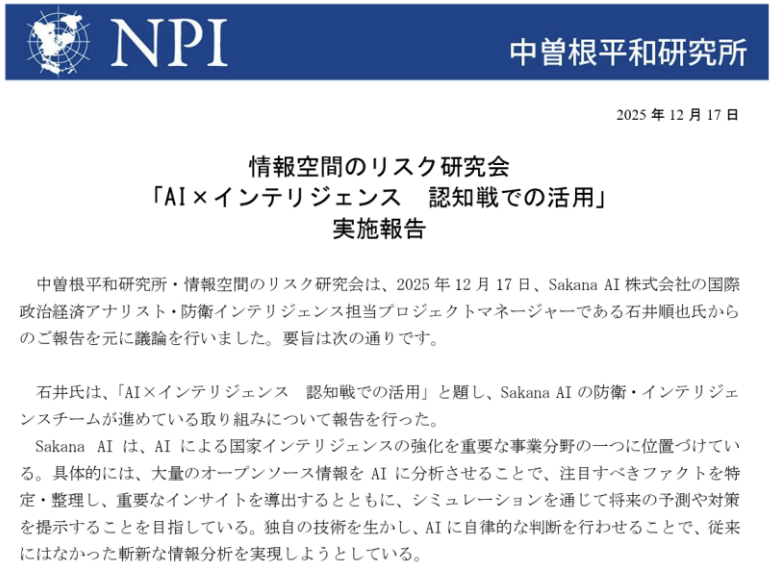
有識者レビュー結果

レビュー結果	今後の開発方針
扇動する感情ごとにナラティブを分類したい。外国からの影響工作なのか見極めたい。	ユーザーごとに分析視点が異なることが判明したため、共通基盤上で柔軟にカスタマイズ可能な設計とする。 TODO
拡散状況の可視化	データ取得性を調査し、フィージビリティスタディを行う。 TODO

5-2. 社会実装に向けた取組の個別詳細

(3) 広範な活用可能性の調査

- 安全保障分野の研究者、政策担当者、実務者の方々から強い関心を示していただいた。
- 例えば、安全保障関連のシンクタンクである中曽根平和研究所様の研究会で、Xの偽・誤情報検知AIを含めたソーシャルメディア分析手法とその分析事例について話す機会をいただいた。出席していた安全保障分野の研究者からは、モダリティ、言語、投稿の影響評価、分析技術の販売方法など、質問やコメントをいただいた。
- また、防衛装備庁技術シンポジウム2025（2025年11月）での講演や陸上自衛隊主催のLandpower Forum in Japan（2025年12月）への出展において、偽・誤情報検知AIを含めたソーシャルメディア分析の取組について発表したところ、防衛省含む中央官庁の職員、自衛隊隊員、安全保障分野の研究者、情報分析アプリケーションを開発している企業など、幅広い組織の方から関心を示していただいた。その後、いくつかの組織とは、個別に打ち合わせを行い、機能、技術、販売方法や時期など、広範な質問やコメントをいただいた。



NPI 中曽根平和研究所

2025 年 12 月 17 日

情報空間のリスク研究会
「AI×インテリジェンス　認知戦での活用」
実施報告

中曽根平和研究所・情報空間のリスク研究会は、2025 年 12 月 17 日、Sakana AI 株式会社の国際政治経済アナリスト・防衛インテリジェンス担当プロジェクトマネージャーである石井順也氏からのご報告を元に議論を行いました。要旨は次の通りです。

石井氏は、「AI×インテリジェンス　認知戦での活用」と題し、Sakana AI の防衛・インテリジェンスチームが進めている取り組みについて報告を行った。

Sakana AI は、AI による国家インテリジェンスの強化を重要な事業分野の一つに位置づけている。具体的には、大量のオープンソース情報を AI に分析させることで、注目すべきファクトを特定・整理し、重要なインサイトを導出するとともに、シミュレーションを通じて将来の予測や対策を提示することを目指している。独自の技術を生かし、AI に自律的な判断を行わせることで、従来にはなかった斬新な情報分析を実現しようとしている。

出典：中曽根平和研 情報空間のリスク研究会
(https://www.npi.or.jp/research/2026/01/06125815.html)

オーラルセッション　場所:「瑠璃の間」 11日（火）　11：00～16：05		
時　間	演　題	発　表　者　（敬称略）
15:25	防衛分野における最先端ＡＩの活用	Sakana　AI株式会社　佐藤　元紀

出典：防衛装備庁 技術シンポジウム2025
(https://www.mod.go.jp/atla/research/ats2025/index.html)

5-2. 社会実装に向けた取組の個別詳細

(4) 導入見込み先の検討

導入見込み先として、主に「メディア企業」「ファクトチェック団体」「安全保障関連機関」の3つのセグメントを定義。想定されるペインポイントとその対応策を検討した。

①メディア企業

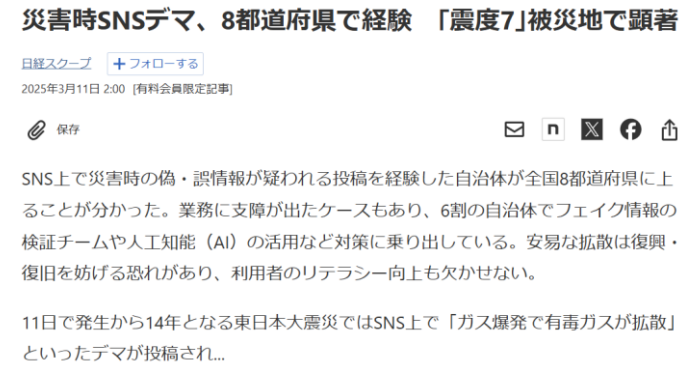
ペインポイント

1) 事実確認（裏取り）には、多角的な証拠収集と緻密な検証が不可欠であり、その調査プロセスに膨大な人的リソースを要している。
2) 即応性を維持しつつ、巧妙化する偽・誤情報を流布しないような仕組みが必要。特に災害時など。

想定される対応策

1) 関連情報の自動収集および根拠（エビデンス）の提示機能により、詳細な検証業務を強力に支援し、真偽判定の「品質」と「効率」を両立させる。
2) AIによる一次スクリーニングで"明らかな偽・誤情報"等をフィルターアウトし、人が判断すべき情報を絞り込むことで、有事の対応速度を向上させる。

災害時の対応については、メディア企業のみならず自治体においても重要な課題として認識されている。すでに具体的な対策の検討を開始している自治体も見受けられ、将来的には有力な導入先候補になり得ると推察される。（記事参照）



災害時SNSデマ、8都道府県で経験　「震度7」被災地で顕著

日経スクープ　＋フォローする

2025年3月11日 2:00 [有料会員限定記事]

保存

SNS上で災害時の偽・誤情報が疑われる投稿を経験した自治体が全国8都道府県に上ることが分かった。業務に支障が出たケースもあり、6割の自治体でフェイク情報の検証チームや人工知能（AI）の活用など対策に乗り出している。安易な拡散は復興・復旧を妨げる恐れがあり、利用者のリテラシー向上も欠かせない。

11日で発生から14年となる東日本大震災ではSNS上で「ガス爆発で有毒ガスが拡散」といったデマが投稿され...

日本経済新聞2025年3月11日記事

出所：
https://www.nikkei.com/article/DGXZQOUE06BHY0W5A300C2000000/

5-2. 社会実装に向けた取組の個別詳細

(4) 導入見込み先の検討

②ファクトチェック団体

ペインポイント	「嘘をつくのは一瞬だが、検証には数日かかる」という非対称性の中で、拡散力が高く社会的に有害な偽・誤情報を優先的に特定（トリアージ）することが難しく、経験による差が出やすい。
想定される対応策	大量の投稿をナラティブ別に分類したもの、あるいは1つ1つの投稿の深刻度と拡散度を評価し、どのナラティブあるいは投稿が真偽判定するべきか提案する機能

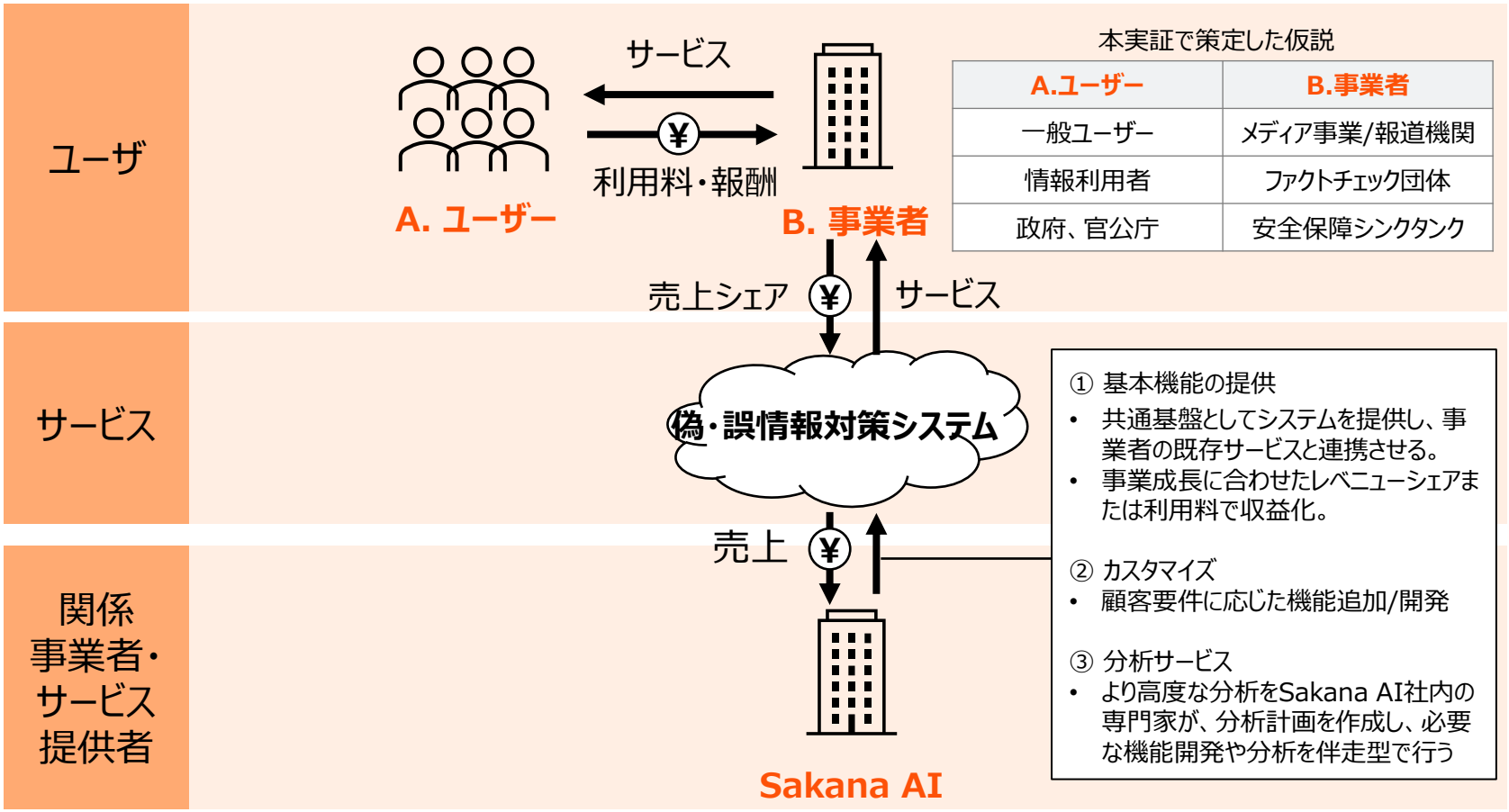
③安全保障関連の組織・研究者

ペインポイント	画像・動画を含んだ情報の分析は多くの工数を要する。 また、情報の出所や拡散経路の特定に加え、大衆の感情をどう操作しようとしているかといった様々な分析視点が求められる。
想定される対応策	本実証システムのような画像、動画を一気通貫で分析できる機能を「共通基盤」として、そのうえにユースケースに応じた機能を実装可能な「カスタマイズ領域」を構築。 さらに、「分析サービス」を提供することで、高度な個別ニーズに柔軟に対応する。

5-2. 社会実装に向けた取組の個別詳細

(5) ビジネスモデル

- 本システムは、既存のメディアや調査機関（事業者B）のサービスに機能を組み込むB2B2Cモデルを採用する。収益源は、プラットフォーム利用による**レベニューシェア**および、高度な個別分析ニーズに応える**カスタマイズ**と**分析サービス**がある。



目次

6-1. 普及啓発活動の全体像

普及啓発活動に係る取組・成果の全体像

- 安全保障およびインテリジェンス関連のイベントや研究会へ参画し、各界・諸団体との連携強化を図った。
- 当初、本システムの主要ユーザーはメディア事業者やファクトチェック団体を想定していたが、一連の活動を通じ、安全保障分野の研究者、政策立案者、実務者からも高い関心が示され、導入候補先の外延を広げることに成功した。
- また、ユーザー属性ごとに「関心事」や「分析の切り口」が異なる実態も明らかとなり、今後のシステム導入アプローチをより精緻に具体化することに繋がった。

①防衛装備庁技術シンポジウム2025

出典：防衛装備庁 技術シンポジウム2025
(https://www.mod.go.jp/atla/research/ats2025/index.html)

②Landpower Forum in Japan 2025

イベント名称	Landpower Forum in Japan（ＬＦＪ）
開催日時	令和７年１２月１７日（水）０９：３０～１７：００ 令和７年１２月１８日（木）０９：３０～１５：１５
場所	東京ドームシティ・プリズムホール （東京都文京区後楽１丁目３-61）
主催	陸上自衛隊（陸上幕僚監部、教育訓練研究本部）
協力団体	公益財団法人陸修偕行社、 一般社団法人日本防衛装備工業会、 一般社団法人日本ＵＡＳ産業振興協議会
展示規模	７６ブース（８８社）
入場条件	完全事前登録制（入場審査通過者のみ） ※ＷＥＢ入場申請の内容によって、入場不可の場合があります ※入場料無料

出典：Landpower Forum in Japan 2025
(https://k3rws.stage.ac/LFJ2025/)

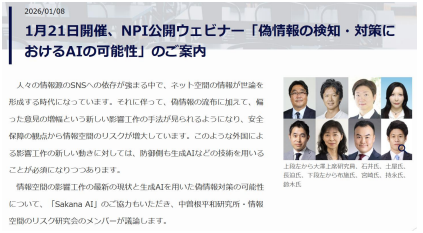
③令和7年度防衛産業参入促進展

出典：令和7年度防衛産業参入促進展
(https://www.atla-event.com/2025su/)

④中曽根平和研究所 情報空間のリスク研究会

出典：中曽根平和研 情報空間のリスク研究会
(https://www.npi.or.jp/research/2026/01/06125815.html)

⑤中曽根平和研究所 公開ウェビナー



出典：中曽根平和研 公開ウェビナー
(https://npi.or.jp/event/2026/01/29135036.html)

6-2. 普及啓発活動の個別詳細

普及啓発活動の実績

- 政府主催のイベントでの登壇、政府主催の展示会への出展など、Sakana AIの開発内容を広く周知する機会があった。
- これらのイベントは、一般市民、メディアの方も参加可能なイベントであったため、幅広い方々に関心を持っていただくことができた。
- 2件の記事にしていただいたことで、さらに多くの方に弊社の取組を周知することができた。

①防衛装備庁技術シンポジウム2025での講演

- 2025年11月11日に防衛装備庁が主催する防衛装備庁技術シンポジウム2025で講演する機会をいただき、Sakana AIが認知戦などのインテリジェンス分析の開発を進めていることを発表した。
- このイベントには、防衛省・自衛隊の方々だけでなく、官民学からインテリジェンス分析に関心のある方々が参加しており、Sakana AIの分析技術について、幅広く議論することができた。
- また、メディアの方も参加しており、後日、右図のように、日経Xtechにて記事として掲載された。

沸騰！防衛テック　フォローする

注目AIベンチャーSakana AIが防衛分野参入、無人機の自律制御に挑む

内田 泰 日経クロステック／日経エレクトロニクス編集委員
2025.11.21

電機　4 min read　有料会員限定記事

シンプル表示　印刷　保存

PR 【Automotive World 2026】最新車載向け半導体を一挙公開
PR IT／製造／建設分野の製品・サービス選択支援情報サイト：日経クロステックActive

米Google（グーグル）などで活躍した一流の研究者が集まって日本で創業した、注目のベンチャー企業Sakana AI（サカナAI、東京・港）が、防衛分野に参入する。同社のApplied Teamに所属する佐藤元紀氏は、防衛装備庁が2025年11月11～12日に開催した「防衛装備庁技術シンポジウム2025」において「防衛分野における最先端AIの活用」というテーマで講演した。同社によれば、まだ具体的なプロジェクトを進める段階に至っていないが、防衛省などと話を進めているという。

Sakana AIは、このAI Scientistを防衛分野に適用できると考えている。具体的には、無人機の自律制御と、偽情報対策や認知戦などのインテリジェンス分析への適用を目指す。

出所：https://xtech.nikkei.com/atcl/nxt/column/18/02438/111700044/

6-2. 普及啓発活動の個別詳細

普及啓発活動の実績

②防衛装備庁令和7年度防衛産業促進展への出展

- 2025年11月11日に防衛装備庁が主催する防衛装備庁技術シンポジウム2025で講演する機会をいただき、Sakana AIが認知戦などのインテリジェンス分析の開発を進めていることを発表した。
- このイベントには、防衛省・自衛隊の方々だけでなく、官民学からインテリジェンス分析に関心のある方々が参加しており、Sakana AIの分析技術について、幅広く議論することができた。

出所：https://www.atla-event.com/2025su/

6-2. 普及啓発活動の個別詳細

普及啓発活動の実績

③陸上自衛隊主催のLandpower Forum in Japan への出展

- 2025年12月17, 18日に陸上自衛隊が主催するLandpower Forum in Japanにブース出展し、Sakana AIがインテリジェンス分析の開発を進めていることをご来場の皆様にご説明した。
- このイベントには、防衛省・自衛隊の方々だけでなく、官民学からインテリジェンス分析に関心のある方々が参加しており、Sakana AIの分析技術について、幅広く議論することができた。

イベント名称	Landpower Forum in Japan（ＬＦＪ）
開催日時	令和７年１２月１７日（水）０９：３０～１７：００ 令和７年１２月１８日（木）０９：３０～１５：１５
場所	東京ドームシティ・プリズムホール （東京都文京区後楽１丁目３-61）
主催	陸上自衛隊（陸上幕僚監部、教育訓練研究本部）
協力団体	公益財団法人陸修偕行社、 一般社団法人日本防衛装備工業会、 一般社団法人日本ＵＡＳ産業振興協議会
展示規模	７６ブース（８８社）
入場条件	完全事前登録制（入場審査通過者のみ） ※ＷＥＢ入場申請の内容によって、入場不可の場合があります ※入場料無料

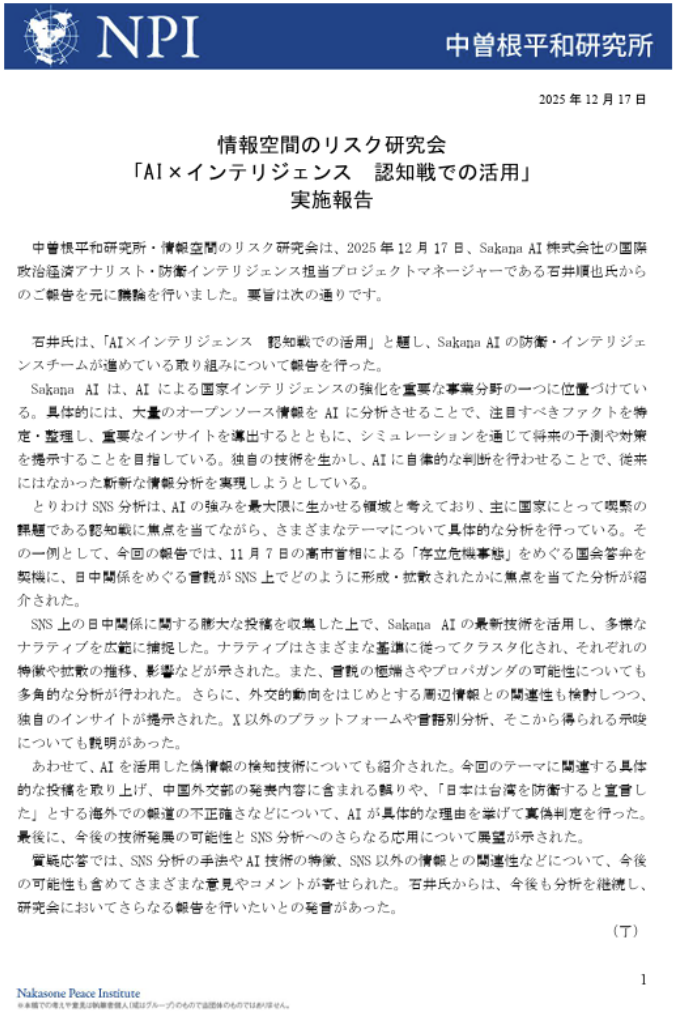
出所：https://k3rws.stage.ac/LFJ2025/

6-2. 普及啓発活動の個別詳細

普及啓発活動の実績

④中曽根平和研究所様によるプレスリリース

- 2025年12月に中曽根平和研究所・情報空間のリスク研究会において議論した内容について、まとめていただき、要旨として公表された。



NPI 中曽根平和研究所

2025 年 12 月 17 日

情報空間のリスク研究会
「AI×インテリジェンス　認知戦での活用」
実施報告

中曽根平和研究所・情報空間のリスク研究会は、2025 年 12 月 17 日、Sakana AI 株式会社の国際政治経済アナリスト・防衛インテリジェンス担当プロジェクトマネージャーである石井順也氏からのご報告を元に議論を行いました。要旨は次の通りです。

石井氏は、「AI×インテリジェンス　認知戦での活用」と題し、Sakana AI の防衛・インテリジェンスチームが進めている取り組みについて報告を行った。

Sakana AI は、AI による国家インテリジェンスの強化を重要な事業分野の一つに位置づけている。具体的には、大量のオープンソース情報を AI に分析させることで、注目すべきファクトを特定・整理し、重要なインサイトを導出するとともに、シミュレーションを通じて将来の予測や対策を提示することを目指している。独自の技術を生かし、AI に自律的な判断を行わせることで、従来にはなかった斬新な情報分析を実現しようとしている。

とりわけ SNS 分析は、AI の強みを最大限に生かせる領域と考えており、主に国家にとって喫緊の課題である認知戦に焦点を当てながら、さまざまなテーマについて具体的な分析を行っている。その一例として、今回の報告では、11 月 7 日の高市首相による「存立危機事態」をめぐる国会答弁を契機に、日中関係をめぐる言説が SNS 上でどのように形成・拡散されたかに焦点を当てた分析が紹介された。

SNS 上の日中関係に関する膨大な投稿を収集した上で、Sakana AI の最新技術を活用し、多様なナラティブを広範に捕捉した。ナラティブはさまざまな基準に従ってクラスタ化され、それぞれの特徴や拡散の推移、影響などが示された。また、言説の極端さやプロパガンダの可能性についても多角的な分析が行われた。さらに、外交的動向をはじめとする周辺情報との関連性も検討しつつ、独自のインサイトが提示された。X 以外のプラットフォームや言語別分析、そこから得られる示唆についても説明があった。

あわせて、AI を活用した偽情報の検知技術についても紹介された。今回のテーマに関連する具体的な投稿を取り上げ、中国外交部の発表内容に含まれる誤りや、「日本は台湾を防衛すると宣言した」とする海外での報道の不正確さなどについて、AI が具体的な理由を挙げて真偽判定を行った。最後に、今後の技術発展の可能性と SNS 分析へのさらなる応用について展望が示された。

質疑応答では、SNS 分析の手法や AI 技術の特徴、SNS 以外の情報との関連性などについて、今後の可能性も含めてさまざまな意見やコメントが寄せられた。石井氏からは、今後も分析を継続し、研究会においてさらなる報告を行いたいとの発言があった。

（T）

1

Nakasone Peace Institute

出所：https://npi.or.jp/research/2026/01/06125815.html

6-2. 普及啓発活動の個別詳細

普及啓発活動の実績

⑤中曽根平和研究所様による公開イベントへの参加

- 中曽根平和研究所様が開催する公開ウェビナーに、Sakana AIの石井順也が登壇し、生成AIをつかった偽・誤情報の流布などの情報分析手法について議論した。

2026/01/08

1月21日開催、NPI公開ウェビナー「偽情報の検知・対策におけるAIの可能性」のご案内

人々の情報源のSNSへの依存が強まる中で、ネット空間の情報が世論を形成する時代になっています。それに伴って、偽情報の流布に加えて、偏った意見の増幅という新しい影響工作の手法が見られるようになり、安全保障の観点から情報空間のリスクが増大しています。このような外国による影響工作の新しい動きに対しては、防御側も生成AIなどの技術を用いることが必須になりつつあります。

情報空間の影響工作の最新の現状と生成AIを用いた偽情報対策の可能性について、「Sakana AI」のご協力もいただき、中曽根平和研究所・情報空間のリスク研究会のメンバーが議論します。

上段左から大澤上席研究員、石井氏、土屋氏、長迫氏、下段左から布施氏、宮崎氏、持永氏、鈴木氏

出所：https://www.npi.or.jp/event/2026/01/08170000.html

目次

7-1. 技術開発及び社会実装における課題・展望

技術開発及び社会実装にあたっての今後の課題およびそれらを踏まえた今後の展望

(1) X以外のプラットフォームを対象とした偽・誤情報・ナラティブ分析機能の開発

本実証事業におけるヒアリングを通じ、言論空間の主戦場がX（旧Twitter）からTikTokやYouTube Shortsをはじめとする動画プラットフォームへと移行しつつある実態が明らかとなった。

今回はXを対象とした開発を行ったが、今後は対象プラットフォームを拡大する方針である。その推進にあたっては、以下の観点を検討する必要がある。

技術的制約

- X以外のプラットフォーム、特に動画メディア等においては、取得可能なデータ項目やアクセス量にシステム上の制約が存在する場合がある。
- そのため、将来にわたり安定的かつ網羅的なデータ収集を実現する手段として、各プラットフォームの仕様に準拠したデータ取得スキームを精査し、サービスの持続可能性を担保できる最適な体制を検討していく。

プラットフォームの選択

- 拡張対象については、ファクトチェック有識者団体や安全保障関連組織など、属性の異なる潜在顧客との協議を継続し、ニーズを精査した上で決定する。各顧客層の要望と導入効果、開発コストのバランスを考慮し、最適な適用範囲を定義していく。

プラットフォームの特徴に応じた新たな機能開発

- 各プラットフォームで異なるモダリティ（テキスト、動画、画像）、ユーザー層、利用目的を踏まえ、X向けとは異なる分析手法の構築が求められる。
- また、単一のメディアに限らず、複数のプラットフォームを横断的に分析（クロスプラットフォーム分析）することで、従来の手法では得られなかった新たな知見の創出を目指す。

7-1. 技術開発及び社会実装における課題・展望

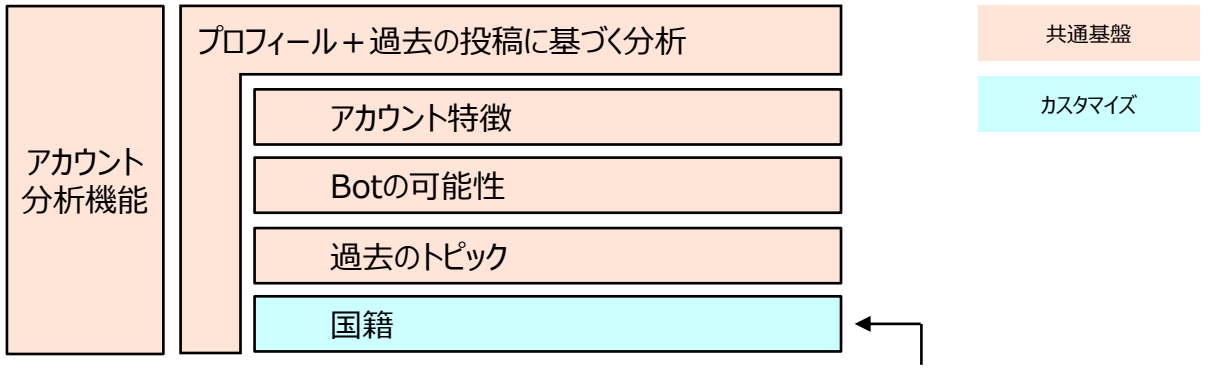
技術開発及び社会実装にあたっての今後の課題およびそれらを踏まえた今後の展望

(2) 導入見込み先拡大と導入アプローチの具体化

- 協力団体との連携、普及活動を通して、安全保障関連の組織・研究者を始め多くの事業者から関心を獲得し、導入見込み先は本実証事業計画時の想定（メディア事業者、ファクチェック団体）よりも拡大した。一方で、事業者ごとに分析の観点や切り口が多岐にわたることも明らかとなった。

- ユーザーごとの分析視点は異なるものの、根幹機能には高い共通性を確認できた。したがって、今後は多様なニーズを効率的に満たすための、「共通基盤」と「カスタマイズ領域」の境界を明確に定義・標準化する方針で、引き続き導入見込み先と協議を進め、スケーラブルなプロダクトとしての製品化・商用化を推進していく。

「共通基盤」と「カスタマイズ」のイメージ
（アカウント分析機能の例）



安全保障の観点では、情報操作の「発信源（国・組織）」の特定が重要視されるため、カスタマイズ機能として追加するイメージ

7-1. 技術開発及び社会実装における課題・展望

技術開発及び社会実装にあたっての今後の課題およびそれらを踏まえた今後の展望

(3) 速報性を重視した「リアルタイムファクトチェック」のニーズ調査

本実証事業における偽・誤情報検知は、高精度かつ深層的な分析を基本方針としている。一方で、協力団体へのヒアリングを通じ、報道機関においては「情報の即時性（リアルタイム性）」がサービスの核心であり、特に災害時等においてその需要が顕著であるとの指摘を受けた。また、ファクトチェック有識者からも、簡易的かつ即応性のある自動ファクトチェックの有用性が示唆された。

これらを踏まえ、本事業で得られた知見を活かしつつ、将来的な発展性として「速報性」に特化した機能提供を検討する。これにより、より広範な社会実装の可能性を模索していく。具体的には、以下の項目について関係各所と協議を進める方針である。

ニーズ調査:

- メディア各社および地方自治体を対象に、災害時の情報課題や「自動簡易判定」への具体的需要を調査する。

事業モデルの検討:

- 迅速な導入を可能にする安価なSaaS型モデルの構築や、外部資金活用の可能性を検討する。

技術開発:

- 速報性を担保するための、軽量かつ高速なAI推論モデル（簡易エンジン）の実装可能性を検証する。

リスク評価とUI設計:

- AIによる自動判定に伴う法的リスクを精査する。特に簡易判定においては、判定根拠の透明性を確保し、「AIによる推定である」旨を明示するとともに、最終的な人間による確認（Human-in-the-loop）を促すUI/UX設計を検討する。

目次

8-1. 実施体制及び役割分担

本事業の実施体制図

各団体の役割・業務範囲

有識者ヒアリング先

- LINEヤフー株式会社様
- (株)Classroom Adventures 様
- ファクトチェック有識者団体

有識者レビュー

- 公益財団法人　中曽根平和研究所
  情報空間のリスク研究会 様

8-2. 全体スケジュール

主な実施事項		令和7年						令和8年	
		8月	9月	10月	11月	12月	1月	2月	3月
対策技術の開発	(可視化) 階層ナラティブツリーナラティブ				→	→	→		
	(検知) ディープフェイク検出アルゴリズムの開発	→	→	→					
	(検知) ファクトチェック			→	→				
	(検知) 画像・動画の生成・加工の検知		→	→					
	(検知) 画像・動画の文脈操作の検知				→	→			
	(検知) 統合的検知アルゴリズム					→	→		
	(対策) SNSシミュレーション					→	→		
	アプリケーション化					→	→	→	
対策技術の有効性等に関する検証及び調査	偽・誤情報の収集及びベンチマークの作成		→	→	→	→			
	ベンチマークを用いた精度検証の実施				→	→			
対策技術の社会実装に向けた取組	偽・誤情報への取組の座組構築	→	→	→	→	→			
	事前・中間ヒアリング		→	→		→			
	レビュー						→	→	
	広範な活用の可能性に関する調査			→	→	→			
普及啓発活動への協力				→	→	→			