あなたのスキルは社会に役立つ
エンジニアだからできる社会貢献

東日本大震災の発生直後に発足したHack For Japanや「市民が主体となって自分たちの街の課題を技術で解決するコミュニティ作り支援」を掲げるCode for Japanのメンバーを始めとして、日本各地で技術を活用した社会貢献活動が行われています。本連載では、防災や減災、地域の活性化や課題解決、そして人材育成など、「エンジニアだからできる社会貢献」の取り組みをお届けします。

第157回

日本初の偽情報対策ハッカソン「Hack the Disinfo」を開催！

●陣内 一樹（じんのうち かずき）

SNSの普及とAI技術の進化により、偽情報の脅威は日々増大しています。アメリカ大統領選挙にはじまり、衆議院選挙や兵庫県知事選挙でも話題になったこの新たな社会課題に対して、エンジニアにできることは何でしょうか。今回は、シビックテックによる偽情報対策の最前線をお伝えします。

本稿では、偽情報について説明した後、Code for Japanの取り組みを紹介します。Code for Japanが進めているコミュニティノートを活用したツール「BirdXplorer」の開発と、日本初の偽情報対策をテーマにしたハッカソンイベント「Hack the Disinfo 2024」は国内外から高い注目を集めています。

世界最大のリスク

急速な技術の革新、経済的な格差の拡大、地球温暖化、紛争や戦争。世界にはさまざまな問題やリスクがあります。その中で最大のリスクと言われて思い浮かべるのは何でしょうか？

世界経済フォーラムが2024年1月に発表した『グローバルリスク報告書2024』で短期の最大のリスクと評価されたのが、偽・誤情報（Misinformation and Disinformation）です[注1]。

注1）https://www.weforum.org/publications/global-risks-report-2024/

偽情報とは何か

偽情報は現代社会に大きな影響を与える問題として認識されていますが、その定義は1つに定まっていません。似たような言葉として「フェイクニュース」がよく使われますが、この用語は多義的であいまいであるため、専門家の間では「ディスインフォメーション（偽情報）」という言葉が好まれる傾向にあります。

日本でよく使われるのは、欧州連合（EU）の専門家会合が提唱した次の定義です。

- ミスインフォメーション（Misinformation）：害を与える意図はないが、誤って共有される虚偽の情報
- ディスインフォメーション（Disinformation）：意図的に害を与えるために共有される虚偽の情報
- マルインフォメーション（Malinformation）：害を与えるために共有される真実の情報。多くの場合、私的な情報を公の場に持ち出すことで行われる

偽情報は単なる間違いではなく、社会や公益に害を与えることを意図して作られ、拡散される情報という特徴を持ちます。

日本でも注目が集まる偽情報

日本国内でも偽情報の影響が顕著になっています。2023年8月に開始された福島第一原発の

処理水放出では、国内外でさまざまな偽情報が拡散されました。中国のSNSでは、日本人が魚を食べなくなったという虚偽の投稿が広まり、日本の水産物輸入禁止の一因になったと言われています。また、処理水の安全性を疑問視する根拠が不明確な主張もありました。これらの偽情報は、単なる環境問題だけでなく、外交や経済にも大きな影響を与えています。

2024年1月の能登半島地震後も、偽情報が問題となりました。「被災地で略奪が発生している」といったデマが拡散し、現地の混乱に拍車をかけました。さらに、AIを悪用した偽の被災地映像も出回り、真偽の判別が困難になっています。これらの偽情報は、被災者の不安をあおるだけでなく、適切な支援活動の妨げにもなりかねません。

これらの事例を通じて、日本国内でも偽情報対策の必要性に対する認識が急速に高まりました。

シビックテックによる偽情報対策

海外ではシビックテックによる偽情報対策が進められています。台湾の「Cofacts」は、台湾のシビックテックコミュニティg0v（ガブゼロ）のメンバーによって開発された、クラウドソーシングによるファクトチェックプラットフォームです。LINE上で動作するボットを通じて、市民が疑わしい情報を報告し、ボランティアのファクトチェッカーがその真偽を確認します。この取り組みは、市民参加型のファクトチェックモデルとして国際的に高く評価されています。また、Code for Africaではエンジニアとジャーナリストが偽情報を拡散するボットを検知するツールの開発を行っています。

コミュニティノートを活用したツール開発

2023年に韓国済州島で開催された日本・台湾・韓国の合同ハッカソン「Facing the Ocean」に参加した筆者たちCode for Japanメンバーは多くの刺激と知識を得ることができました。「Facing the Ocean」は開発を競う通常のハッカソンとは異なり、さまざまな国の人が集まり、知見を共有し、ネットワークを作る場です。これまで日本のシビックテックにおいて偽情報は主要なテーマではありませんでした。一方、台湾や韓国のシビックハッカーは約5年前から偽情報に取り組んでおり、多くの経験を積み重ねていました。

済州島での経験やリサーチをもとに筆者たちが注目したのがX（旧Twitter）のコミュニティノートです。コミュニティノートは、ユーザーが投稿に対して追加の文脈や情報を提供でき、それらの注釈がほかのユーザーによって評価され、十分な合意が得られた場合に投稿と共に表示されるしくみです。コミュニティノートの詳細なデータは公開され、毎日更新されています。このデータとTwitter APIのデータを組み合わせ、APIやダッシュボードを提供することで偽情報対策に役立つのではないかと考えました。同じコンセプトのツールは海外にもありません。また、現在は研究者向けのTwitter APIの無償提供が止まっているため、研究者からも関心を持ってもらっています。

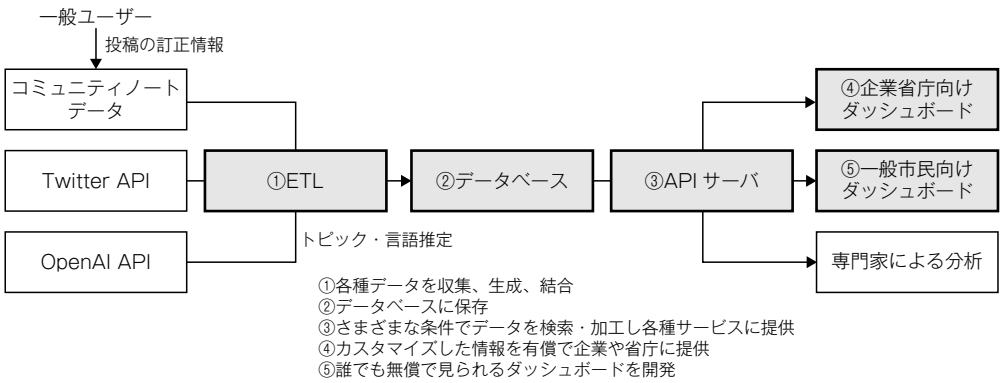
このアイデアをもとに、「BirdXplorer」と名付けられたプロダクト（**図1**）は2023年からオープンソースで開発が開始され、東京大学のシンポジウムでの発表やユーザーテストなどを通じて「キーワード検索をしたい」など具体的なフィードバックをもらい、改善を進めています[注2]。

日本初の偽情報対策ハッカソン

Code for Japanは、2024年11月3日～4日の2日間、日本初となる偽情報対策をテーマとしたハッカソン「Hack the Disinfo 2024」を開催しました（**写真1**）。日本の衆議院選挙やアメリカの大統領選挙と同時期に開催され、NHKや日本経済新聞でも取り上げられるなど高い注目を

注2）https://github.com/codeforjapan/BirdXplorer

◆図1　Code for Japanが開発中のBirdXplorer



集めました[注3]。ハッカソンではBirdXplorer APIを提供し、多くの参加者に使ってもらうことができました。そして、2日間という限られた時間でしたが、開発された作品は有識者からも高く評価されました。ここではとくに注目を集めた作品を紹介します。

茨城パンダ愛好会「パンダキャプチャー」

パンダを愛する4人組が作った「パンダキャプチャー」は、中国で主要なメディアであるショート動画を証拠保全し、内容に関するレポートをGoogle Driveに自動生成してくれるWebアプリです[注4]。中国では縦動画がはやっており、多くの偽情報も生まれています。パンダキャプチャーを使うことによって、動画が検閲で消されてしまうことを防ぎ、中国語で話されている内容もわかるようになります。中国語圏における日本の偽情報対策のツールとして非常にユニークで完成度の高い作品です。

注3) https://www3.nhk.or.jp/news/html/20241103/k10014628211000.html, https://www.nikkei.com/article/DGXZQOUA074YV0X01C24A0000000/

注4) https://github.com/ibaraki-panda-lovers/panda_capturer

◆写真1　ハッカソンの開発風景

aknて「disarm bot」

セキュリティエンジニアの若手チームの「aknて」は「disarm bot」を開発しました。「disarm bot」は偽情報への防御思考を高めるボットとして、偽情報対策の「DISARM Framework」に基づいた自律型マルチLLMエージェントです。ユーザーがDiscordでテーマを投げかけると、攻撃側と防御側そして中立の合計5つのエージェントが議論を行います。複数のエージェントが議論することによって多角的な視点をユーザーに提供できます。

偽情報研究所（giken）「AIと一緒に偽情報を見破ろう！」

チーム5人全員がOSINT（Open Source Intelligence）CTFプレイヤーである偽情報研究所は、マス層に届けるゲーム性のある「AIと一緒に偽情報を見破ろう！」を開発しました[注5]。

このアプリはChatGPTと協力して記事の真偽を見抜き、情報リテラシーを診断・採点する

注5) https://withai.disinfox.com/

ことができます。AIも偽情報を判断するには限界があるということを逆手にとって、間違えることもあるChatGPTと協力して偽情報を判断することで、偽情報に気づくポイントをトレーニングしていくアプリです。

InVID日本語化

日本を含む世界中のファクトチェッカーが使っているChrome拡張機能「Fake news debunker by InVID & WeVerify」（InVID）というオープンソースのツールがあります。発表者の富永さんはハッカソン前からフランスのAFPメディアラボに働きかけ、InVIDの日本語化を行いました。フランスが休日であったためにハッカソン期間中には本番環境への反映が終わりませんでしたが、11月13日に反映されました。ハッカソンの作品が実際に使われていくとても貴重な事例です。

すらいむちゃん

利用者自身がXにおけるフィルターバブルを認識するためのChrome拡張機能を開発しました。ポストにカーソルを当てると類似するジャンルのコミュニティノートを探し、投稿主の属性と類似のコミュニティノートに表示します。投稿主の属性は最近の投稿をOpenAIのAPIによってトピックに分類して表示します。Chrome拡張機能によってユーザーが容易に投稿の背景やバイアスを理解することで、フィルターバブルを認識できます。また、ハッカソン中にBirdXplorerに機能のリクエストをもらえたのはCode for Japanとしてとてもありがたかったです。

クロスキーパーズ「健全裁判（仮）」

ハッカソン期間中に結成されたクロスキーパーズは偽情報に対して後から対応するデバンキングではなく、いかに社会の分断を作らないかというプレバンキングに着目しました。

開発した健全裁判はゲーム形式で特定のポストやナラティブについて議論するWebアプリです。ゲームは、証拠集め→オンライン議論→評価の3フェーズに分かれています。議論のフェーズでは、True/Falseではなく、証拠の能力について互いの証拠の能力を補強するためについて話すことで健全な議論のベースを作ります。モデレーターが入ることで相手を否定しないようにしたり、良い証拠を提示したりするとポイントがもらえるというゲーム性を持たせるなどの工夫が考えられています。

チーム（仮名）

続いてもハッカソン期間中に結成されたチームです。このチームはCode for Japanがハッカソンに向けて開発したBirdXplorer APIをもとにツール開発とコミュニティノートの分析を行いました。

ツールはコミュニティノートを調べるためのツールで、APIとは異なり非エンジニアでも使うことが可能です。分析では人権・福祉・言論の自由といった関連のトピックにユーザーが反応していること、コミュニティノートの判定結果によって使用される単語に違いがあること、個別のノートの考察などが行われました。

おわりに

ハッカソンで発表された作品は、東京大学の小泉悠准教授や明治大学の齋藤孝道教授といった審査員や海外からのゲストから高い評価を受けました。ハッカソンにはエンジニアだけでなく、OSINTアナリストや記者などが参加したことも多様で実践的な作品が生まれる要因になりました。

現在、Code for Japanでは衆議院選挙とアメリカ大統領選挙、そして兵庫県知事選挙の偽情報についての分析をしています。偽情報の影響について、印象論ではなく、データやエビデンスに基づいて分析することは日本の偽情報対策を考えるうえで非常に重要です。BirdXplorer APIや今後リリースするダッシュボードに興味がある方はぜひCode for Japanまでお問い合わせください。SD