

Lead PI: Paul Resnick
wisedex.info@umich.edu

David Jurgens

Amy X. Zhang

Overview

Social media platforms have policies against harmful misinformation. Unfortunately, they enforce unevenly, especially for non-English content. WiseDex harnesses the wisdom of crowds and AI techniques to help identify more misinformation posts. The result is more comprehensive, equitable, and consistent enforcement, significantly reducing the spread of misinformation.

Description

Trust and safety teams at platforms like TikTok and Twitter have limited resources for developing custom software or machine learning models, despite sitting inside huge technology companies. They do, however, use keywords or regular expressions to search for posts purveying misinformation that is of special concern, with the results being queued for human review.

The core WiseDex innovation is to expand high-level, abstract policies into a large database of claims that violate the policies. Each claim profile will include keywords in multiple languages, making it easy for platform trust and safety teams to search for matching posts and flag them for human review. Claim profiles will also include instructive examples that help reviewers decide quickly and consistently.

A four-stage process leads to the addition of new claim profiles to the database. (1) Candidate content is collected from published fact-checks, reviewer escalations, and crowd workers searching for suspicious content. (2) Other workers articulate specific claims that cover some of the candidate content. (3) A subject matter expert assembles a casebook with arguments and evidence about whether a claim violates relevant policy criteria, and a representative lay jury considers that evidence and makes final judgments. Policy criteria are drawn from several platforms' policies, making it easy for a

platform to determine whether posts matching that claim violate its policies. (4) More crowd workers, assisted by AI, identify keyword search terms in multiple languages and assemble examples with explanations.

Building on the claim profiles database, we will develop a Claim-Matcher API to help platform trust and safety operations surface content for human review, and a Reviewer Assistant plug-in to improve the speed and consistency of that human review. We will also develop benchmark datasets of public posts, labeled by human reviewers. This will allow evaluation of the efficacy of keyword search terms and form the basis for public report cards showing the fraction of harmful misinformation that platforms have acted on.

Differentiators

The industry already relies on similar databases to aid in the detection of other kinds of harmful content. The Global Internet Forum to Counter Terrorism (GIFCT), an industry consortium, maintains a database of perceptual hashes of images and videos produced by terrorist entities, and the Internet Watch Foundation maintains one for child sexual abuse images and videos. Platforms can check whether images and videos posted to their platforms match any in the databases by comparing hashes. By analogy, a WiseDex misinformation claim profile functions like an image hash in those systems. However, identifying matching posts is more challenging in the misinformation space. Thus, WiseDex has developed its four-stage process for formulating claims that are both searchable by computers and recognizable to human reviewers.

Journalistic organizations around the world publish fact-checks that will often be excellent starting points for WiseDex claim profiles, jump-starting the

work of a subject matter expert assembling a casebook. The additional work of our final stage, to identify multi-lingual keyword search terms and gather posts with explanations, is still needed to enable consistent detection.

Several companies, such as Sentropy, Hive.ai, ActiveFence, and project partner Trust Lab, sell services to platforms to improve the coverage, speed, and consistency of enforcement against misinformation as well as other harmful content. WiseDex contributes several things to this ecosystem. First, because claims are judged with respect to a variety of policy criteria, platform trust and safety teams can choose only those claims that violate their own policies. Second, casebooks and judgments about claims are open to public scrutiny. Finally, WiseDex will produce benchmark datasets based on flagging and reviewing public posts. These will help vendors to improve their products, help platforms to compare the vendors' performance, and enable public accountability report cards.

Road Map

Year 1 will involve refining the four-stage claim intake process, culminating in the production of 400 claim profiles in English and Spanish. We will also develop commercial and research prototype software for claim matching and reviewing in Year 1. In Year 2, we will create 1,000 additional claim profiles, publish benchmark datasets and an accountability report card, and set post-grant licensing terms for the claims database. We will also settle on an organizational form, likely incorporating as a nonprofit organization or becoming a project of an existing industry consortium or nonprofit organization. In Years 3 through 5, following the end of the grant, we will continue to add claim profiles and expand to additional languages.

Partnerships

WiseDex contributors include: three universities, Michigan, Washington, and Massachusetts Institute of Technology; non-profit organization Meedan; and commercial vendor Trust Lab. Meedan will develop software that makes it easy for fact-checking organizations to share their user flags and fact-checks with WiseDex. The Election Integrity Partnership (Washington and Stanford) will provide a data stream of misleading content related to the 2022 U.S. elections that

will serve as one source of claims to profile.

We welcome inquiries from government agencies and other organizations that surface claims that should be profiled as part of the WiseDex database. We also welcome inquiries from advocacy organizations, especially those focused on misinformation in languages beyond English, about potential partnerships around the preparation of public accountability reports.

Intellectual Property

Task workflow software and the claim profile database will be available for licensing royalty-free for non-commercial use. Trust Lab will be granted an exclusive, royalty-free license for commercial use during the two-year grant period. We welcome licensing inquiries.