Project Summary

NSF Convergence Accelerator Track F: WiseDex: Enabling Fast, Comprehensive Misinformation Enforcement

PI Paul Resnick
Co-PIs: Adam Berinsky
Senior Personnel: Scott Hale, David Jurgens, Benji Loney, Rada Mihalcea, David Rand, Veronica Perez-Rosas, Amy Zhang

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

The core innovation of the WiseDex project is to expand high-level, abstract policies about categories of content which are hard to systematically detect into a large database of profiles of particular claims. 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, again in multiple languages, that help reviewers decide whether specific posts match or not.

A four-stage process leads to the addition of new claim profiles to the database. First, candidate content is collected from published fact-checks, reviewer escalations, and crowd workers searching for suspicious content that does not match existing claim profiles. Next, crowd workers, assisted by AI, articulate specific claims that cover some of the candidate content. Third, a subject matter expert assembles a casebook containing arguments and evidence about whether a claim violates relevant policy criteria and a jury deliberates about that evidence and makes final judgments about the claim topic. The policy criteria are drawn from several platforms' policies, making it easy for each platform to determine whether posts matching a claim violate its policies. Finally, more crowd workers, again assisted by AI, identify keyword search terms and assemble both matching and non-matching posts with explanations to help reviewers make consistent decisions.

Intellectual Merit Research challenges in organizing crowdsourcing workflows and in natural language processing will be addressed as part of this project. We will conduct iterative design, deployment, and evaluation of software to support the crowdsourcing tasks necessary to identify and profile new claims. Natural language processing techniques will be developed to assist crowd workers in interactive clustering to identify new claims, to extract search keywords, and to match posts with claims.

Broader Impacts Widely circulating misinformation can lead people to make decisions that harm their own or others' health and safety, and erode trust in institutions. WiseDex will directly help to reduce the circulation of misinformation, by providing software services that integrate with platform trust and safety operations to surface content for human review and improve the speed and consistency of that human review. We will also create benchmark datasets of public posts from social media platforms, annotated with the claims that they contain. These datasets will be useful for training and evaluation of detection systems and allow effectiveness comparisons among services provided by different researchers and vendors. The benchmark datasets will also enable public report cards that assess how well platforms enforce their misinformation policies on public posts. Finally, we will develop an online course to help the public understand the work of platform trust and safety teams and broaden participation in computing by helping people prepare for jobs on those teams.

PROJECT DESCRIPTION

Note: this description is slightly edited from the version submitted.

1 Objectives and Significance of the Proposed Activity

As new communication platforms have opened up the means of publishing and reaching broad audiences, new avenues have opened for misinformation to spread [42; 77]. There are concerns that this may lead to individuals taking actions that harm themselves, such as taking risky treatments for health conditions or losing money to financial scams [78; 17]. There are also concerns that it may erode the public's trust in decisions of governments and other institutions or even erode confidence in the idea of truth itself [43], making it harder for societies to respond to public problems such as pandemics. Four domains where misinformation seems to be causing harm are health [49], elections [61], climate change [21], and inter-group conflicts [72].

There is widespread agreement that communication platforms should take action to identify misinformation and then reduce its impact by marking articles and posts in some way, reducing their distribution, or removing them entirely. For example, upcoming European Union regulations are expected to build on the self-regulatory "Code of Practice on Disinformation" and the upcoming Digital Services Act, with increased onus on large platforms to take action on reported content and increased penalties for non-compliance [12]. In the United States, platform executives have faced questioning from lawmakers about failures to act on foreign disinformation campaigns [31].

The platforms have responded by setting policies and enforcement actions [54; 73; 41]. Though the policies vary and their details are often not made public, at a high level they generally prescribe some kind of enforcement action against posts that are harmfully misleading. One common enforcement action is to alert users that a particular piece of content may contain misinformation, through a label, color coding, or text. Twitter refers to this as warning [75]. Facebook refers to it as an inform [50] action. Another possible action is to downrank a content item so that it appears later in search results or news feeds and thus fewer people encounter it. Many other such strategies are also available in addition to filtering or removing the content [48].

Before they can enforce in any way, however, platforms have to detect the posts that violate their policies, and they miss many of these posts. Enforcement outside of English is especially poor [56; 61; 76]. For example, on one sample of misinformation content, Facebook failed to label 29% of English-language misinformation, 70% in Spanish and 68% in Italian [1].

To understand why, consider how platforms identify violative posts. First, through a combination of end-user flagging, a generic "misinformation model", and searches for posts on particular topics, they identify posts; some posts are handled automatically and some are queued for human review [57; 53; 73]. Reviewers who have access to a detailed codebook then make judgments about whether queued posts violate the platform's policies.

Problems can occur in both steps. First, many violative posts are not detected. End-users may be selective and strategic in what they flag [14]. Generic language models are not good enough at identifying misinformation; if tuned for high recall (identifying a large fraction of violative posts), they would return so many false positives as to overwhelm the capacity of the human reviewers, and so they err on the side of missing violative posts instead. Specialized keyword searches can be effective, but they only tend to be run for the few topics that a trust and safety team is currently focused on, and rarely in languages other than English.

Second, it is difficult to get consistent judgments from human reviewers about whether posts contain misinformation [22; 68; 70; 87]. Codebooks are large and are supplemented by special advisories. Given a post, a reviewer may not recall specific guidance that can be found in the

codebook or advisories. Sometimes there is no specific guidance that covers a post, and reviewers have to make judgments with respect to higher level policies; this typically happens for newly emerging claims that have not yet been fact-checked and for the long tail of persistent but somewhat infrequently shared claims.

The primary objective for WiseDex is to enable more comprehensive, multi-lingual, consistent detection. Our approach expands on the specialized keyword search mechanism; we accumulate a large database of specific claims that are searchable and recognizable to human reviewers. We then support human reviewers by providing software that identifies candidate claims that the post may contain. Enhanced detection will enable better enforcement against misinformation, at large platforms like YouTube, TikTok, and Facebook, mid-sized platforms like LinkedIn and Pinterest, and a host of more niche services. Platforms with smaller Trust and Safety teams may especially benefit from the breadth of the WiseDex database.

A secondary objective is to create benchmark datasets of public posts from social media platforms, annotated with the claims that they contain. These datasets will be useful for training and evaluating enforcement systems. They will also be useful for creating public report cards that assess how well platforms enforce misinformation policies on their public posts.

2 Convergence Research

The WiseDex team draws on expertise in natural language processing, machine learning, crowdsourcing, and political psychology. Our phase I team included academics and a commercial startup company, TrustLab, which sells Trust & Safety services to social media platforms. For phase II, we also welcome non-profit Meedan, which provides technology support services to international fact-checking organizations. Meedan is the inaugural technology provider for third-party fact-checking programs at WhatsApp and another major social media platform. Its fact-checking software Check, is used by more than two dozen fact-checking organizations in India, Brazil, Kenya, Indonesia, Mexico, Spain, Germany, Zambia, France, Columbia, and the United States.

2.1 Claim Profiles for Searchable, Recognizable Claims

Though they vary in exactly how they state their misinformation policies, most social platforms have a high-level policy that prescribes some enforcement action against posts that are harmfully misleading [54; 73; 41]. *Misleading* means that the post would lead readers to believe one or more false claims. *Harmful* means that belief in those false claims could lead to real-world harm for the believer or for others. Some platforms add a third criterion about the stance of the post, for example that it specifically asserts the claim or deliberately attempts to mislead rather than simply having that effect, which distinguishes the concepts of misinformation and disinformation [80]. Interpreting these criteria can be challenging because it may not be obvious whether a particular claim is false or how much harm it would cause if believed.

Our core innovation is to expand abstract misinformation policies into a large database of searchable, recognizable claims. A claim is *searchable* if there are search queries that will return the vast majority of the posts matching the claim (high recall), without returning too many non-matching posts (moderate precision). Each WiseDex claim profile includes keyword expressions, in multiple languages, which trust and safety teams can use as queries to flag posts for human review. WiseDex will also develop a claim matcher API that will classify posts with higher precision and recall than is possible with the queries, as described in Section 2.3.2. A claim is *recognizable*

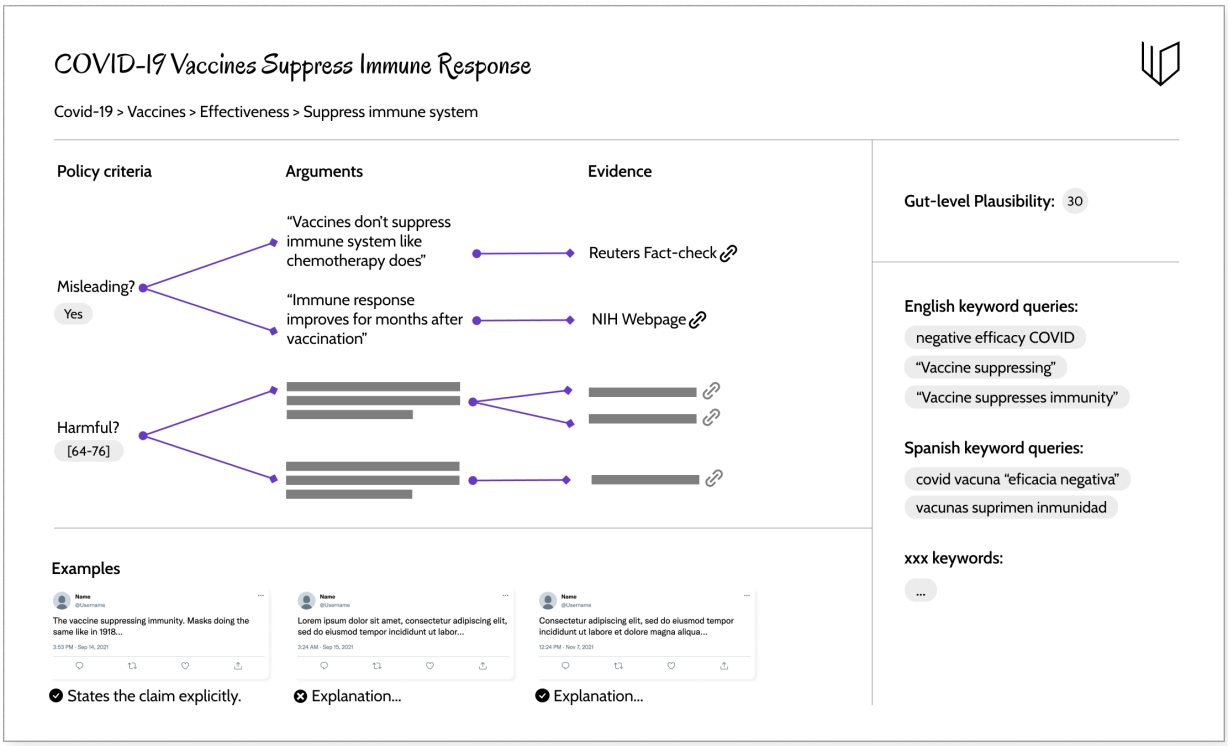

Figure 1: A sample claim profile

if human reviewers, given a post and the claim, can quickly assess whether the post contains the claim and there is high consistency among the reviewers. Each WiseDex claim profile will include instructional example posts that include explanations of why the posts do or do not contain the claim, which will help to improve consistency of assessments among reviewers. Figure 1 illustrates the elements of an example claim profile, one that we will use as a running example.

The industry already relies on similar databases to aid in detection other kinds of harmful content. A database of perceptual hashes [86; 29] of images and videos produced by terrorist entities is maintained by The Global Internet Forum to Counter Terrorism (GIFCT) [27] and the Internet Watch Foundation maintains ones for child sexual abuse images and videos [35]. 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 a process for formulating claims that are searchable and recognizable.

2.2 Policy Trees

Platform policy teams have already taken the first steps toward expanding the abstract criteria into guidelines that specify what we will refer to as a policy tree. Each child node specifies a subtopic of the parent node. For example under the parent topic of harmful COVID misinformation, Twitter's published guideline identifies subtopics including, "deliberate conspiracy by malicious and/or powerful forces" and "misrepresent the protective effect of vaccines" [74]. When a subtopic is so specific that it is defined by a single claim, such as vaccines suppressing immune response, we

will refer to it as a claim topic, or just a claim for short. Platform guidelines sometimes include specific claims as illustrative examples of a category, but we have not seen any that attempt to enumerate a large database of claim topics as an aid to detection.

Part of the art of defining policy trees, as practiced by product policy-makers at the social media platforms, is to subdivide branches until they reach sub-topics that exhibit *policy homogeneity*, meaning that all posts matching that sub-topic are similarly misleading and cause similar levels of harm. For example, Twitter specifies an enforcement action of removal for the branch, "deliberate conspiracy by malicious and/or powerful forces" but a milder action of information labels for the branch, "misrepresent the protective effect of vaccines". Sometimes, homogeneity can be achieved except for an identified sub-subtopic. In that case, a *carveout branch* can override the enforcement action of the parent branch.

To complicate matters, platforms have different policy trees. To be most useful to a platform's trust and safety team, each WiseDex claim profile should exhibit homogeneity for that platform's policies. Our approach is to define a WiseDex unified policy tree based on the union of many platforms' policy trees. Our processes will then define branches that have policy homogeneity with respect to all the policies, but not necessarily lead to the same enforcement action for all platforms. A WiseDex policy tree node might be violative for one platform but not for another, but will not be heterogeneous for any platform.

We will convene working groups to define unified policy trees in four areas: health, election information, climate change, and intergroup conflict. Each working group will have 8–10 people, including representatives from policy teams of platforms and outside experts from academia and civil society. A working group will assemble criteria articulated in published guidelines from platforms, from the Facebook Oversight Board [19] and from other published works of academics and advocacy groups. In addition, platforms will be invited to offer their more detailed private guidelines for consideration, under a non-disclosure agreement. The working group will not judge which criteria should be used to determine policy violations; all available criteria will be included. We will provide a tool that makes it easy for each platform to mark any criteria or topical subdivisions that are irrelevant to its policies, thus leading to a reduced tree. The platform can then associate enforcement actions with branches of the reduced tree.

When setting enforcement policies, platforms need to account for not only the harms caused by misinformation that their platform helps to amplify, but also the costs in terms of user trust in the platform that may occur from over-enforcement. People of different races, genders, and political orientations experience disproportionate enforcement [32] and frustration with the platforms is especially high among those who do. An ideal system would enforce on every post that violates a platform's policy and no others, but some enforcement errors will be inevitable.

To gain insights into the public's tolerance for over-enforcement versus under-enforcement, we will conduct survey experiments that present scenarios about, hypothetical content enforcement systems and ask respondents whether they would want those systems turned on or off. The hypothetical scenarios will vary on two dimensions of interest: (1) the specific topic of the content and (2) the proportion of true and false positives and true and false negatives that would occur if the enforcement system were used. This will yield insight into the public's tolerance for over-enforcement versus under-enforcement, and examine whether such preferences change depending on content type. We will contract with YouGov to collect data from a representative sample of U.S. residents to ensure the external validity of our findings.

2.3 Using the Claim Profiles

Before describing the processes by which WiseDex will accumulate a database of claim profiles, we first describe services that make use of those claim profiles. To help social media platforms enforce their policies more comprehensively, we will develop claim matching software and a reviewer assistant tool. Commercial partner TrustLab will develop and sell commercial versions of these services and research prototypes will be made available for low-volume, non-commercial use. We will also simulate the operation of a trust & safety team on a pool of content from platforms that have primarily public posts, such as reddit, Twitter, and YouTube. This will produce public datasets that can be used as benchmarks for evaluating vendors' misinformation monitoring tools and for producing public report cards about the frequency of enforcement actions. The datasets will respect the platforms' policies, in particular by linking to content on the sites rather than copying it.

2.3.1 Surfacing Content: Search Terms

In our stakeholder interviews, we heard that trust and safety teams have limited resources at their disposal for development of custom software or machine learning models, despite sitting inside huge technology companies. That was also the experience of senior personnel Loney when he led trust and safety teams. They do, however, usually have access to a search tool, using keyword or regular expression queries. When a misinformation topic is of special concern, they use that tool to conduct searches, with the results being queued for human review. The WiseDex database makes it easy to scale up that process to more topics and more languages, using the search terms that are included in claim profiles.

We will create a machine-readable format in which to publish the claim profiles. Based on reviewer labeling done at the time of profile creation (see Section 2.4.5), each search term in the profile will have an associated precision estimate, the fraction of returned posts from public platforms that match the claim topic. We will regularly run each of the queries on a random sample of posts from public platforms to estimate the current prevalence of the claim topic. The machine readable profiles with their embedded estimates will make it easy for platforms to create automated pipelines to run all of the queries that meet thresholds they set for precision and prevalence.

2.3.2 Surfacing Content: Claim-matcher API

Some platforms may opt for greater integration, where they call on a low-latency, high-throughout claim-matcher Application Programming Interface (API) to process some or all of their posts. The input is post text, and the output of this API is the set of claims that are best matches to the post, if any, with confidence scores. The claim-matcher API will offer some advantages over surfacing content through keyword search. One is that the API can use more sophisticated natural language processing techniques than keyword or regular expression search, and thus achieve higher precision and recall. Another is efficiency: each post is only processed once, rather than potentially being considered separately in many keyword searches.

TrustLab already offers a version of this API, but it currently matches posts to published fact-checks rather than to WiseDex claim profiles. While some WiseDex claim profiles will be based on published fact-checks, some claim topics may not map cleanly to fact-checks, some claims may be profiled earlier than published fact-checks appear, and WiseDex will map claims to platform policy criteria. Thus, having an API that matches posts to WiseDex claim profiles, in addition to published fact-checks, will be valuable.

Recent advances in Natural Language Processing (NLP) research will drive the development of the claim-matching software, including transformer-based architectures [15; 13] and embedding similarity search models [69; 38]. Commercial partner TrustLab will develop a commercial version, focusing on the challenges of achieving low latency and high throughput that are required by social media platforms to handle their post volume.

2.3.3 Reviewer Assistant Tools

WiseDex will contribute to the speed and consistency of human reviewer decisions. Recognition tasks ("is this an X?") are easier for people than recall tasks ("what is this?") [33]. Existing tools require human recall; a reviewer must remember or search for particular policy guidance that applies to the post at hand. Specialization by geography and topic makes this easier, by restricting the set of policy rules that may be relevant for a particular reviewer on a given day, but does not eliminate the problem.

The WiseDex approach is to provide a reviewer assistant tool that turns the recall problem for reviewers into a recognition problem, by automatically searching for the claim profiles that are potential matches to the current post. The reviewer then only needs to decide whether any of the potential matches are correct. Moreover, if the reviewer is unsure, documentation about the claim topic has already been retrieved and is available for viewing. This documentation includes explained examples that have been selected to help illuminate the decision boundary, as described in Section 2.4.5. The net result is that reviewers should be able to decide more quickly, and raters will be more likely to agree with each other.

Our Phase I interviews revealed that each platform has its own tool, typically web-based, for presenting content to reviewers and eliciting their judgments. They often hire vendors such as Accenture, Concentrix, and Vaco to recruit and supervise the reviewers, but the vendors usually do not provide the reviewing software.

When we presented mockups to both platform trust and safety teams and external reviewer vendors, some expressed great interest. Six potential customers, including two very large platforms and a vendor that manages contract reviewers, indicated interest in trialing the WiseDex reviewer tool in a production environment. For the external vendors, they thought that it could help them improve speed and inter-rater reliability, metrics on which they are judged by the platforms that contract with them.

They raised a concern, however, that it would require a large up-front commitment to modify the platform's reviewer tool. Thus, we will create a stand-alone version of a reviewer assistant tool. A reviewer opens it in a separate tab or window. The reviewer pastes in some text, is shown potential matching claims, and can explore the instructive examples and other material in the claim profile. The reviewer then returns to the main reviewing interface to enter a judgment.

We will also offer several levels of optional integration for the assistant tool. One is to embed it in a browser plug-in. With the tool loaded, a reviewer is automatically presented with a visual indicator highlighting likely misinformation content directly in their review queue (similar to how 3rd party plugins like "Grammerly" highlight lexical errors in text across platforms). A higher-level of integration would come in the form of a platform-specific browser plug-in that would further improve reviewer efficiency. Based on parsing the HTML in that platform's reviewer tool, it will pre-load the assistant's content for each post and automatically fill in judgments when the reviewer selected a matching claim in the assistant tool. The highest level of integration would come from the platform incorporating the assistant features directly into their tool, calling APIs that we provide.

2.3.4 Benchmark Datasets

We will create monthly benchmark datasets, each containing a sample of public social media posts, labeled with any WiseDex claim profiles that they match. The posts will come from platforms that have primarily public posts, such as reddit, Twitter, and YouTube; we will oversample for content likely to be related to our primary domains of health, elections, climate change, and intergroup conflict. These datasets will be useful for training and evaluation of our claim-matcher API as well as models produced by others.

To label the posts, we will simulate the review process of a trust and safety team. That is, we will create our own reviewer tool, with the reviewer assistant built in. We will invite bids from the same kind of external vendors that the platforms use, such as Accenture, Concentrix, and Vaco, and contract with them to set up reviewer queues.

In addition to generating benchmark datasets, this process will feed information into the claim profile creation process described in Section 2.4. If reviewers disagree about whether a post matches a particular claim profile, it can trigger additional work on defining the claim topic (2.4.3) or the instructive boundary examples (2.4.5). If a reviewer thinks the post matches a claim profile but not the policy assessment associated with it, this can initiate review of whether the claim topic needs to be split in order to maintain policy homogeneity (2.4.3). Finally, if the reviewer thinks the post does not match any claim profile, but may be considered harmful misinformation by some people, they can flag it for escalation, contributing to the pool of new candidate content (2.4.2).

2.3.5 Enforcement Report Cards

We will also create public report cards about the misinformation enforcement of platforms that have public, searchable posts. We will track whether and when a platform begins enforcement against each of post in a benchmark dataset. To do that, the posts in the benchmark dataset will be selected through the platform's streaming API, so that they are selected soon after they are first posted. We will then re-fetch them periodically and check for removal or inclusion of warning labels. A report card, then, can include metrics such as the fraction of posts that should have received warning labels, according to the platform's policy, that had them within one hour, one day, or three days, the fraction of posts that should have been removed that were and, for those that were removed, the median time from posting to removal. We will provide these metrics per language, and these metrics can provide concrete evidence for advocacy groups and policy makers that have raised concerns about disparate enforcement in different languages, and clear evidence of improvement that should be celebrated when that occurs.

Unfortunately, it will not be possible to create report cards based on external audits of platforms whose posts are not public. However, there is growing pressure for platforms to provide transparency. We hope that our report cards will serve as a model for information that transparency reports should include. We will also offer to work with any platform with non-public content under an NDA to produce public report cards while keeping the content private.

2.4 Articulating and Profiling Claims

In order to scale up the generation of claim profiles to thousands of claims, our process requires careful integration of large groups of people as well as machines, producing a collective intelligence beyond the capabilities of any one individual [51]. The challenge in the design of crowdsourced workflows is to decompose a complex task into a set of component tasks, provide software supporting each task that enhances worker capabilities, and provide software that coordinates the

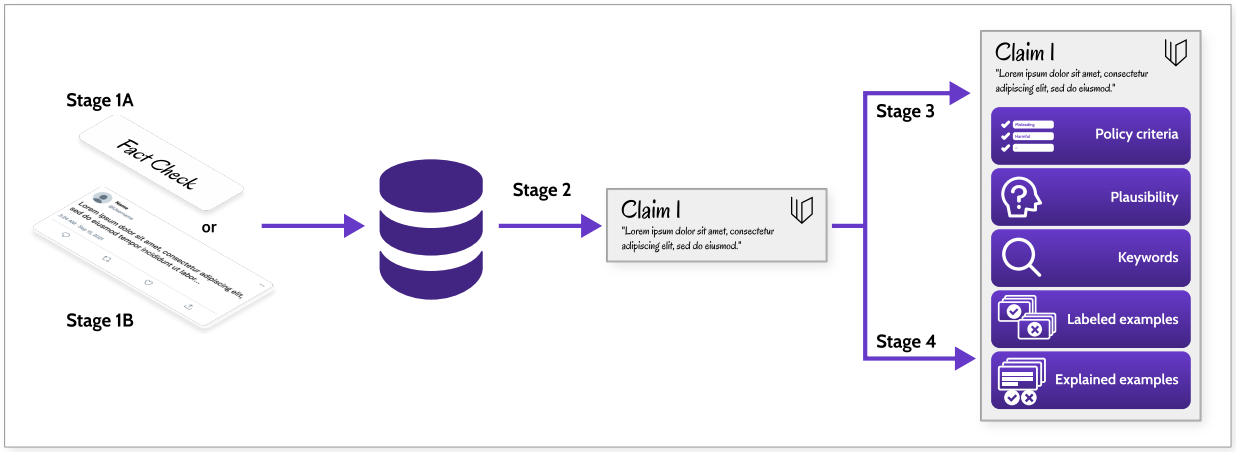

Figure 2: Articulating and Profiling Claims

workflow, assigning tasks to appropriate people and aggregating their outputs [46]. Some common patterns have emerged such as find-fix-verify, in which one group of workers finds problems that need to be fixed, another group proposes solutions, and a third group filters invalid solutions [4]. These and other patterns have been applied to increasingly complex tasks [40], resulting in increased efficiency [8] and consistency [7]. Other work has explored workflows that incorporate differences in skills and perspectives from the crowd, including expert crowdsourcing systems [65; 39] and systems that account for diverging laypeople assessments [37; 28; 7]. Recent work has examined the respective strengths of experts and laypeople in content moderation and misinformation identification [5; 59], suggesting the promise of hybrid approaches. Through our design of structured crowdsourced workflows, we will efficiently coordinate both experts and laypeople towards the novel task of articulating and profiling claims.

We have designed a four-stage process, as illustrated in Figure 2. In the first stage, candidate content is collected from published fact-checks (1A, Section 2.4.1), and from several sources of posts (1B, Section 2.4.2). In Stage 2 (Section 2.4.3), workers articulate specific claims that cover some of the candidate content. In Stage 3 (Section 2.4.4) and 4 (Section 2.4.5) an expert, a lay jury, and crowd workers fill in elements of the claim profile. For most tasks, the workflow includes a review step where a second worker provides feedback or signs off on the first worker's result. Workers completing tasks in all four stages will be supported by Natural Language Processing capabilities that extract keywords from a text corpus and that cluster similar documents. In Stage 4, Machine Learning techniques will be used to select queries to explore, select posts to be assigned to human reviewers, and select posts to use for training human reviewers.

2.4.1 Stage 1A: New Fact Checks

One source of new candidate content is fact-checks and other published reports from external organizations. These are especially useful because they tend to be created for claims that are circulating on social media and they provide much or all of the information needed for the casebook that an expert assembles in Stage 3. Meedan will continue to expand its **Fetch** software that aggregates fact-checks from over 40 sources including other publicly accessible (but incomplete) aggregators and websites of fact-checking organizations. The service cleans the text, identifies data errors, and exposes the fact-checks in the ClaimReview structured format defined at Schema.org [10; 11]. We will also process a feed of misinformation reports related to U.S. elections provided by the Election

Integrity Partnership (eip.net).

When a new fact-check appears, a worker will check whether it matches with an existing claim profile in the database. If so, it will be added to the claim profile and, if needed, that claim profile will be referred for reconsideration of the judgment process of Section 2.4.4 and/or the keywords and examples accumulation of Section 2.4.5. New fact-checks that do not match existing claim profiles will be marked as candidate content for consideration in later stages.

2.4.2 Stage 1B: User Flags, Escalations and Searches

A second kind of candidate content is posts that people have flagged for consideration. There may not (yet) be a published fact-check covering the specific claims in such posts. Such posts may still be useful fodder for identifying patterns in stage 2, where claims are articulated. We envision three sources of such posts. The first is escalations from reviewers employed to label content for our benchmark data sets, as described in Section 2.3.3.

The second is direct submissions by users. Meedan currently operates tiplines on behalf of two dozen fact-checking organizations around the globe, where end-users can suggest content for the fact-checking organizations to consider. The software checks incoming tips against existing fact-checks to help the organization prioritize whether to respond. Meedan will build the ability to allow the fact-checking organizations to optionally choose to share data with WiseDex.

The third is posts found through a search task that we pay crowd workers to perform. A worker is asked to provide a search query where at least two of the first ten returned posts are worthy of coverage by a new claim profile. To be worthy, a post must not match any existing claim profile and the worker must think that some people might consider it to contain harmful misinformation. Workers may get inspiration for their queries from as-yet-unprocessed fact checks and posts and from exploration of the existing policy tree. We will provide a custom search tool that searches across public tweets on Twitter and public videos on YouTube; when the worker tries a query, results that match existing claim profiles are marked as such, with the keywords that match those claim profiles highlighted.

We also enhance the worker's natural capabilities by providing a tool that suggests possible keywords to use as queries, drawing on a keyword extraction tool that we will develop. We will leverage unsupervised algorithms that combine statistical properties of entities and their relationships. Building upon our previous work on graph-based keyword extraction [55], and the more recent unsupervised algorithm based on local statistics features [6], we will develop a pipeline consisting of three main steps: (1) candidate keyword extraction, where the claim will be preprocessed and individual words and phrases will be extracted as candidates; (2) candidate scoring, where a score will be calculated for each candidate, combining statistical features [6], graph scoring [55], and positional features [23]; and (3) candidate ranking, where keywords will be selected based on their score.

2.4.3 Stage 2: Formulating Claims

Skilled workers will perform the task of proposing a refinement of the existing policy tree that would cover one or more of the fact checks and posts from stage 1. For example, we can imagine a worker noticing some posts in the candidate pool related to negative efficacy of the vaccine, and also noticing the Reuters fact-check. The worker might then have been inspired to articulate the false claim in Figure 1, that COVID vaccines suppress people's immune response.

This will be a challenging task requiring sophistication about misinformation policies, logical thinking skills, and a good understanding of the entire WiseDex claim profiling process. We ex-

pect workers to graduate to this task after demonstrating skill at the search task described in the previous section and the reviewing task of recognizing whether posts match existing claims.

A software tool to support workers in completing this task will include several components:

- A policy tree exploration tool to help identify the closest parent and turn it into a checklist of policy criteria.
- An interactive clustering system, built by our NLP team on top of Meedan's **Alegre** similarity matching system [52]. The similarity matching will leverage the recent context-aware neural models for text representations [16; 47], which we will fine-tune to the specifics of our domain. Moreover, through interaction, the clustering algorithm will move away from the generic one-size-fits-all methods and seamlessly integrate worker feedback [85]. This system will suggest groupings of candidate content that has been collected through the Stage 1 activities, as well as content that the worker has surfaced through queries. The worker can then adjust the groupings. The keyword extraction tool described in the previous section will be used to suggest keyword labels for organizing the groups. The system will also include a component that allows for precedent exploration in order to examine the neighborhood of past decisions on posts that share properties with a group of posts. By selecting and comparing against existing decisions, workers will construct a structured explanation for establishing a new claim, as well as what existing claims are most similar.
- A version of the query tool from previous section. This will help the worker to check how widely a claim is circulating on social media platforms.
- Nudges to search for counter-example posts that would render a proposed claim topic heterogeneous with respect to the policy criteria; only after failing to find such counter-examples should the worker deem the claim topic ready for consideration in Stage 3.

2.4.4 Stage 3: Expert Casebooks and Lay Panel Assessments

Given a proposed claim generated in stage 2, the task in the third stage is to assess it with respect to corresponding policy criteria. The harm that would be caused by exposure to the information is always one of the criteria, but more specific criteria may also be defined in the policy tree.

For this process, a domain expert first formulates arguments for or against the claim topic meeting the associated policy criteria and the evidence supporting the arguments. For example, for the criterion of whether claiming that vaccines suppress immune response is misleading, Figure 1 shows two arguments that an epidemiologist surfaced during a pilot test of this task: that vaccines don't suppress the immune system like chemotherapy does; and that immune response has been shown to improve rather decline for at least several months after the vaccine. In assembling evidence, experts are asked to follow the Wikipedia principle of "no original research" [84], meaning that they link to trustworthy sources, including fact-checks, but do not interview people or write their own reports. For example, in the pilot test, the epidemiologist provided a link to a Reuters fact-check page for the first argument and a CDC page for the second. The arguments and evidence constitute a *casebook*.

The casebook is sent to a layperson jury that will deliberate and make judgments about each of the criteria. We design our layperson jury process to include features that improve perceived public legitimacy and consistency according to prior literature, including incorporating deliberation [34; 20], expert-assembled evidence [59], and prior examples [7]. In addition, we select jury

participants to be representative of important viewpoints. For instance, for topics that are politically polarized in the United States, that would include both Republicans and Democrats; for other topics and in other countries, other representation constraints may apply.

Prior to consideration of the casebook, the jurors individually report a gut-level plausibility assessment, the probability that the claim is true or false. Platforms may not be so concerned about claims that would be harmful or problematic if believed, but are totally implausible (e.g., the President is actually a robot controlled by Martians); whereas claims that are misleading but more plausible may be a larger concern.

The jury then deliberates about the arguments and evidence and renders final judgments about each of the criteria. Most judgments are binary yes–no. For the general criterion of "harm if posts related to the claim are widely distributed", the jury renders a judgment as an interval on a 1–100 scale. For that, they will use the Goldilocks system [7], comparing the claim to other claims in order to define a lower and upper bound on its harm score.

During this stage, if a proposed claim is found to not exhibit all the features needed for a claim profile, it is returned to an earlier stage for further refinement. For example, if the expert or the jury finds that the claim is heterogeneous (a policy criterion applies to some but not all posts promoting the claim), then the claim is sent back to stage 2 to be further refined, along with the problematic examples. Similarly, if the jury finds that the casebook is incomplete or biased, they can send it back to the content experts for amendment.

The public nature of the WiseDex casebooks will enable an appeals process that leads to continuous improvement, as well as public legitimacy of the assessments in the claim profiles. An apellant will be asked to select among possible reasons for the appeal: arguments and evidence that should have been included in the casebook but were omitted; newly emerging evidence that contradicts evidence in the casebook; examples of posts that match the claim but do not violate the policy criteria under which the claim was judged. The appellant will also be expected to provide links to evidence or example posts to support their appeal. A small panel will assess whether the appeal is frivolous; if it is not, an expert will consider amending the casebook and a new jury will be convened to reassess the claim.

2.4.5 Stage 4: Search Terms, Labeled Examples, and Explained Examples

Given a claim that exhibits policy homogeneity, the goal in this stage is to make the claim searchable and recognizable. Successful completion of this stage will result in the addition of the following elements to the claim profile:

- **Search Terms** A set of keywords and regular expressions, in multiple languages, that each has a precision of at least 10% when run as search queries on publicly queryable platforms, meaning that at least 10% of returned posts would be judged by human reviewers as matching the claim topic.
- **Explained Examples** A small set of "instructive boundary posts," in multiple languages, that include explanations generated by juries about why those posts do or not qualify as matching the claim topic. Human reviewers, after completing a training program involving the instructive boundary posts, should have high inter-rater reliability at classifying posts as matching or not matching the claim topic.
- "**Labeled Posts**," A larger set of posts, in multiple languages, with correct labels but without explanations, suitable for training an automated classifier.

To select search terms, we will build upon our prior work for model-in-the-loop re-ranking [45; 44]. From the topic description and initial samples in each language, we will generate queries appropriate to the tools used by reviewers, using a hybrid approach that combines keyword extraction and generation. The keyword extraction tool described in Section 2.4.2 will generate an initial set of candidate queries. Queries will be *generated* by training a new T5 model [64] on the data from Section 2.4.2 to generate queries from the claim document, an approach that has seen high performance in recent benchmarks for its ability to generalize [e.g., 90; 36]; because the model is trained directly from how real reviewers search, these generated queries may take on complex forms such as regular expressions or using Boolean search operators (e.g., conjunction).

The candidate queries with high precision are retained. To estimate precision, queries are run and passed through a claim-specific classifier. The classifier is trained on posts that human raters have labeled, using the same reviewer pool used for creating benchmark datasets as described in 2.3.4. Posts are selected for review base on classifier uncertainty, a technique known as active learning [67; 18]. As more matching posts are accumulated, the keyword extraction/generation processes are rerun, possibly generating additional candidate queries to evaluate. This process yields not only a set of queries to run but also a set of positive and negative examples and a classifier trained on those examples; these are all included in the claim profile.

In order to achieve rater consistency, human reviewers may need to see some instructive examples with explanations. We will select the examples through a semi-automated process we have developed recently called Active Crowd Training, inspired by the technique of active learning for AI models, but in this case selecting based on human disagreement rather than the uncertainty of automated classifiers. It proceeds in multiple rounds. In each round, several reviewers are trained using the instructive examples selected from previous rounds; then they label new posts. The post on which the reviewers disagree most is added as an extra instructive example in the next round. Each new example is sent to a jury of raters, who make a decision about whether the example should be treated as a match to the claim topic or not, and write an explanation for why or why not. Through this process, we aim to develop contrasting example pairs with high agreement but where the examples have high textual similarity with each other (or with the misinformation claim) but only one is an example of the claim. These pairs will help serve as explained examples of what is or is-not a part of the claim and can also help better refine policy; such contrastive examples pairs are also extremely useful for training machine learning models [79; 58].

2.4.6 Evaluation

We will conduct an iterative process of evaluation and improvement of all the elements of the four-stage intake of new claim profiles. For the software supporting worker tasks such as searching for novel posts, formulating claim topics, and creating casebooks, we will evaluate the usability of the tools through feedback from workers and the utility of the tools by the frequency with which work outputs from one stage are returned for rework by workers at a later stage. The keywords produced for the claim profiles will be evaluated based on the precision of their novel results, meaning the fraction of results matching the associated claim topic, among all results that are not returned by other queries. The quality of the topic descriptions and instructive examples will be jointly evaluated by measuring the inter-rater reliability of reviewers who label posts as either matching the claim topic or not.

We will also conduct an experiment to check whether the plausibility ratings elicited in Stage 3 are predictive of effectiveness of accuracy nudges. We expect accuracy nudges to be more effective for claims that are less plausible [63] so this will serve as a check on the reliability of the plausibility ratings. We will recruit participants from Lucid (a service for recruiting online survey participants)

and show them each claim (e.g., in the form of a Facebook post), and ask them how likely they would be to share the claim if they saw it online. Half the participants will receive an accuracy prompt at the outset of the study that makes the concept of accuracy top of mind for them (e.g., having them judging the accuracy of a random banal claim). We will then compare the average sharing intention for each headline in the control (no accuracy prompt) versus treatment (accuracy prompt) conditions, and correlate this difference with the jury plausibility ratings.

In addition to iterative evaluation and improvement of particular components, we will collect two measures of the effectiveness of the overall intake pipeline. The first is whether difficulty of the search task for novel posts increases as the claims database grows. Early on, searchers will find low-hanging fruit of common claims, guided by existing fact checks that have not yet been profiled. Even though searcher skill may improve with practice, as the database grows it should take longer to find novel posts and these should increasingly reflect newly emerging claims that were not widely shared previously.

The second indicator of overall effectiveness is how long it takes to create claim profiles for fresh claims. One of the challenges platforms have is that it can take weeks to mobilize enforcement against a fresh claim, by which time quite a bit of harm has already occurred. We will measure time to completion as the elapsed time from earliest identification of matching content through Stage 1A or 1B and the end of the Stage 4 claim profiling process. We will also estimate the total number of matching public posts on social media sites prior to completion of the claim profile; this number should be minimized if we are effective at focusing attention first on claims with higher prevalence.

3 Partnerships and Roles

Key team members and their project roles are described in Table 1.

We will establish partnerships with some of the fact-checking organizations that use Meedan's software to run tiplines. We have also established a partnership with the Election Integrity Partnership, which will provide a data stream of misleading content related to the 2022 U.S. election that will serve as one source of claims to profile, as described in Section 2.4; Kate Starbird has provided a letter of collaboration.

We ran an advisory board in Phase I and we expect many of the participants to return. Members act in an advisory capacity and do not have responsibility for carrying out any of the work activities described for the projects. Phase I advisors included (affiliations listed for identification purposes only; participation did not signify institutional endorsement from their employers): Michael Bernstein (Stanford), Mary Beth Hunzaker (Twitter), Jakub Kocinski (TikTok then Meta), Justin Kosslyn (Google), Rachel Larimore (The Dispatch), Lydia Laurenson (The New Modality), Carmen Scurato (Free Press), and Jeffrey Sorensen (Google Jigsaw). Jonathan Zittrain from Harvard has provided a letter of collaboration indicating that he will join the advisory board for Phase II.

4 Coordination Plan

The project team will work together through a Slack workspace, Zoom, and email. Project manager James Park will be responsible for personnel and procurement issues, and for tracking progress toward all deliverables.

Four teams have responsibility for specific components, and each sits solely within a single organization: NLP at the University of Michigan, crowd-sourcing at the University of Washington;

Name	Organization	Role	Responsibility
Paul Resnick	Michigan	PI	Overall leadership
Adam Berinsky	MIT	Co-PI	Survey on public preferences for error tradeoffs of policy enforcement (Section 2.2)
James Park	Michigan	Other personnel	Project Manager
Benji Loney	Trust Lab	Senior Personnel	Product management and customer discovery
Amy Zhang	Washington	Senior personnel	Crowdsourced Workflows (Sections 2.3.4, 2.4.2, 2.4.3, 2.4.4, and 2.4.5)
Rada Mihalcea	Michigan	Senior personnel	Natural Language Processing (Sections 2.3.2, 2.4.2, 2.4.3, and 2.4.5)
Veronica Perez-Rosas	Michigan	Senior personnel	Natural Language Processing (Sections 2.3.2, 2.4.2, 2.4.3, and 2.4.5)
David Jurgens	Michigan	Senior personnel	Natural Language Processing (Sections 2.3.2, 2.4.2, 2.4.3, and 2.4.5)
Scott Hale	Meedan	Senior personnel	Intake of of fact-checks and user flags from external sources (Sections 2.4.1 and 2.4.2) and supervising software development for research prototypes across the project
David Rand	MIT	Senior personnel	Evaluation of credibility scores by correlation with effectiveness of accuracy nudges (Section 2.4.4)

Table 1: Roles and Responsibilities

surveys at MIT; external intake of fact-checks and opt-in users flags at Meedan. PI Resnick will continue the practice of joining weekly team meetings for the components via Zoom.

Benji Loney will hire and supervise a product manager at TrustLab who will coordinate software development work across all the teams. The product manager will conduct customer interviews and develop a product roadmap. The product manager will also be responsible for software project management, following an Agile/scrum methodology. This includes writing product specifications in the form of story cards with acceptance criteria, prioritizing work for sprints and conducting sprint post-mortems. The software developers will be hired by Meedan and Michigan; both organizations have experience with geographically distributed teams for software development.

The full project team will have an annual retreat at the University of Michigan, for which subcontractors have budgeted travel. There will also be a weekly check-in meeting for all the leads of component activities.

The advisory board will have one-hour remote meetings monthly. They will also be invited to travel to the University of Michigan for an annual one-day in-person meeting, in conjunction with the full project team retreats.

5 Phase I Portfolio

The Phase I NSF curriculum and weekly convenings had three important impacts on the project. The first was the stakeholder interview process, which led to a fairly major pivot in project direction, as described below. The second was the communication tasks and exercises. Our stakeholder interviews steeped us in jargon familiar to people working in trust and safety roles at platforms. However, in feedback about an oral pitch and about a draft of our written project overview for the expo, it became clear that we needed to speak more plainly to be understood by a broader audience. Boiling down our post-pivot ideas to a three-minute video led to increased clarity within the project team as well as providing a quick way to introduce potential new partners to our plan. The third was that the weekly meetings connected us to many others doing complementary work in this space, notably Scott Hale from Meedan who has joined our team for phase II. Below we describe five project activities during Phase I, and their results.

Legitimacy Surveys Across two online studies (Ns = 676 and 660; quota matched to national distribution on age, gender, ethnicity, and geographic region), we presented participants with a series of possible 'juries' of decision-makers tasked with evaluating whether some piece of online content should be labeled as harmfully misleading on a social media platform. Each jury varied on three key dimensions: (i) who is evaluating the online content (e.g., laypeople juries such as random people online or nationally representative; expert juries such as professional fact-checkers or domain experts), (ii) the qualifications of jury members (e.g., no qualifications or minimum news knowledge or domain knowledge for laypeople), and (iii) whether or not juries were allowed to discuss content with each other during the evaluation process. For each jury, participants answered a series of questions about how legitimate they perceived the jury to be. Participants also were asked to choose which of a pair of juries they preferred to make decisions for several pairs of juries presented, via a choice-based conjoint format [2]. As expected, clearly inappropriate juries (e.g., coin flip, roll of the dice) were evaluated as less legitimate than other juries. Next, despite evidence of recent heightened distrust of experts in the U.S., layperson juries (without minimum qualifications or discussion) were perceived as less legitimate than expert juries (also without discussion). However, adding minimum qualifications (news or domain knowledge) and the opportunity to discuss allowed lay juries to match the legitimacy of expert juries (without deliberation). Republicans and Democrats largely followed the same pattern of preferences, although Republicans reported systematically lower trust in all juries. One difference was that Republicans showed a clear preference for juries that included both Democrats and Republicans, whereas Democrats found such juries to be no better than randomly selected juries. These findings have informed the two-stage design of our claim profile assessment process (Section 2.4.4); experts prepare casebooks to ensure that jurors are well informed, but final judgments are made by juries that can be selected to be more representative of the population.

Stakeholder Interviews We conducted stakeholder interviews with four journalist/fact-checkers, one leader of an industry consortium, one leader of an advocacy group, two academics, and eleven people associated with technology platforms. While many were intrigued with our original focus on golden sets, judged by representative juries, as a way for platforms to claim legitimacy for their misinformation enforcement activities, it became clear that public legitimacy of their decisions was not their highest priority. Both insiders and outsiders agreed that handling emerging, fresh misinformation, and improving enforcement coverage, especially for non-English posts, were more critical needs. Thus, we pivoted. Multiple interviewees also pointed to a potential inspiration: the Global Internet Forum to Counter Terrorism (GIFCT) industry consortium's database of perceptual hashes of images and videos produced by terrorist entities [27]. WiseDex claim profiles are the analog of GIFCT image hashes, although much more is required to make WiseDex claims

be automatically searchable and recognizable to human reviewers. When we speak to industry insiders, we now refer to WiseDex informally as "GIFCT for misinfo."

Feedback on Reviewer Assistant Mockups Feedback on the new approach has been very positive. Six potential customers, including two very large platforms and a vendor that manages contract reviewers, indicated interest in trialing the WiseDex reviewer tool in a production environment. Several, however, specified that a key to their interest was the ability to try it without requiring integration with existing reviewing software. This inspired our plan to offer a standalone version of the reviewer assistant first, with a browser plugin offering an intermediate level of integration, as described in Section 2.3.3.

Pilot Test of Expert Casebook Assembly As a pilot test of the casebook assembly process described in Section 2.4.4, we hired six epidemiologists on UpWork to assemble casebooks. They all understood the task and were able to complete it in an average of 49 minutes each, and they generally rated each other's casebooks favorably, with occasional suggestions for more arguments or evidence to avoid an appearance of bias. One important design insight we gained was that casebooks will sometimes need to record arguments with links to the best available evidence but a negative assessment of that evidence, rather than simply omitting arguments for which no trustworthy evidence was found.

Pilot Test of Keyword Extraction As a pilot test of crowd-generated search query generation, we hired crowd workers on Prolific to find queries related to six claim topics, in ten languages (English, Spanish, Arabic, Hindi, Tagalog, and others). They were able to generate new claim-specific queries in all languages, including the keywords shown in Figure 1. A separate group of annotators found that 15–60% of tweets retrieved by some of these queries were directly related to the claims, demonstrating that crowd workers can suggest good query terms and underscoring the global, multilingual nature of identifying misinformation.

6 Timeline of Milestones and Deliverables

Figure 3 shows a roadmap of the project milestones with the target completion times. Numbers in italics refer to proposal section numbers where corresponding activities are described.

All of the software deliverables will be generated during the two years of Phase II. We expect to populate the claim profiles database with at least 400 claims, each in English and Spanish, by the end of the first year, with an additional 1000 in the second year. In subsequent years, only software maintenance, not major development, will be required. In years 3–5, WiseDex will expand the number of claims in the database, and expand to additional languages beyond English and Spanish.

7 Deliverables

The project deliverables are:

- **Data Products** The central deliverable is the database of claim profiles, as described in Section 2.1. These will be made available in machine-readable format. Using the database, we will also produce benchmark datasets (Section 2.3.4) and report cards (Section 2.3.5).
- **Policy Trees** As described in Section 2.2, working groups will create unified policy trees for four domains: health, elections, climate change, and intergroup conflict.

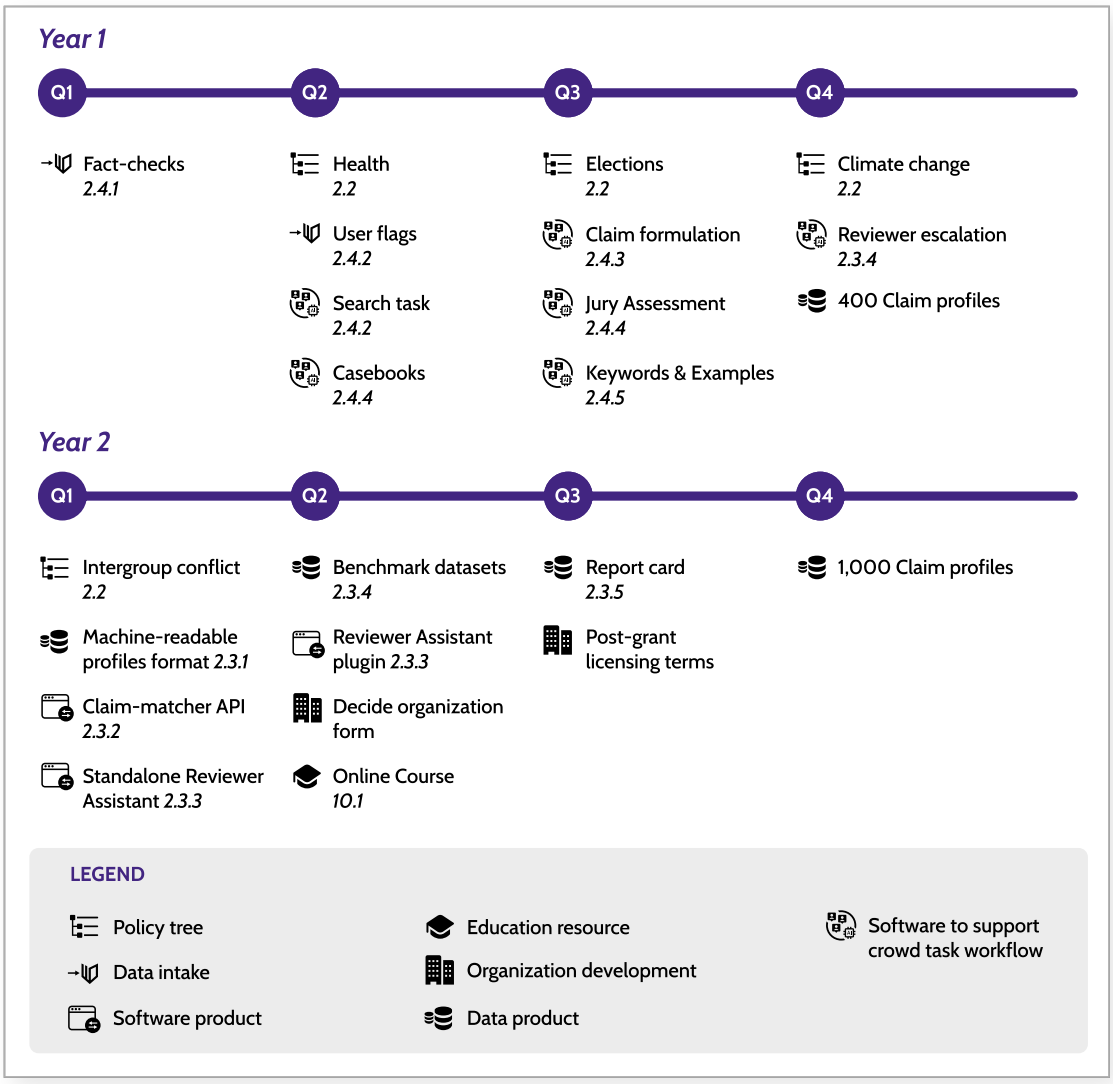

Figure 3: Roadmap of Project Milestones

- **Data Intake Processes** As described in Sections 2.4.1 and 2.4.2, Meedan will develop automated pipelines to take in published fact-checks and flagged content from user tiplines.
- **Software Products** As described in Sections 2.3.2 and 2.3.3, we will develop both research prototypes for non-commercial use and commercially supported software-as-a-service for claim matching and reviewer assistance.
- **Software to Support Crowd Task Workflows** Software to support crowd workflows, in some cases powered by natural language processing and other AI tools, will be developed to support the four stages of claim profile creation (Section 2.4). This software will be made publicly available for non-commercial use.
- **Organization Development** One of the key activities for year two will be establishing the organizational form for WiseDex subsequent to NSF funding. The most likely is either

a stand-alone nonprofit organization or becoming a project of an existing non-profit organization or industry consortium. Revenue would come either directly from social media platforms or indirectly from platforms via one or more commercial trust and safety vendors who pay licensing fees for use of the claim profiles database or software.

- **Education Resources** We will develop an online course to help people understand the work of platform trust and safety teams (see below in Section 9.1) and prepare for jobs on such teams.

8 Track Alignment

The overarching goal of Track F is to "develop prototype(s) of novel research platforms forming integrated collection(s) of tools, techniques, and educational materials and programs to support increased citizen trust in public information of all sorts (health, climate, news, etc.), through more effectively preventing, mitigating, and adapting to critical threats in our communications systems." The WiseDex project will help communication platforms to mitigate one critical threat, the dissemination of misinformation.

WiseDex will be complementary to other Track F projects that work either to identify misinformation narratives and sources or projects that develop resilience and resistance to misinformation in end-users and communities. Misinformation identified through other projects can be a source for our intake process, as described in Sections 2.4.1 and 2.4.2.

In the other direction, information produced as part of our claim profiles may be a useful resource to other project teams. For example, the casebooks of arguments and evidence links assembled by expert panels for claim assessment, as described in Section 2.4.4, may be useful to teams developing misinformation training materials for the general public. As another example, the keyword search terms that we surface as part of each claim profile may be used to search for examples or to conduct audits that yield report cards of the kind described in Section 2.3.5, but specialized to the specific topics of interest to communities served by other project teams. As described in the Intellectual Property section (11), we will provide royalty-free non-commercial licenses to the full database of claim profiles.

Finally, other projects that are working with localized communities or training end-users may lead some people to be interested in an entry path to trust and safety work at social media platforms. WiseDex will provide such a path through paid work, starting with low-level reviewing tasks and graduating to more complex tasks of searching for misinformation posts not yet caught by WiseDex (Section 2.4.2) and formulating subtopics that exhibit policy homogeneity (Section 2.4.3). As described in Section 9.1, we will also develop a short online course about platform trust and safety policy making and enforcement, which would help help people prepare for jobs in the trust and safety industry.

9 Broader Impacts

The most important broader impacts will come directly from the project deliverables. The claim matcher and reviewer assistant tools, powered by the claim profiles database, will enable faster, more comprehensive, and more consistent detection of harmful misinformation. That will enable platforms to take enforcement actions that reduce the audience that misinformation reaches and the impact that it has on them, while also reducing false positives that threaten free speech and create a public backlash against enforcement. The benchmark datasets will help to advance

development of misinformation detection algorithms and the enforcement report cards will provide better empirical evidence for policy makers and advocates trying to understand the state of misinformation detection and enforcement.

9.1 Broadening Participation Plan

This project will contribute to broadening participation in computing in two ways. The first is by creating a pathway into Trust and Safety work for people who complete our crowdwork tasks. Some scholars have criticized low-level, highly-structured reviewing tasks for rarely serving as a stepping-stone to higher level positions in the industry [30; 66]. However, Senior Personnel Loney, who previously managed platform trust and safety teams, reports hiring many people who had previously served as reviewers, if they showed a higher level understanding of the policy-making, detection, and enforcement process of which reviewing is a component. To help open this door, we will develop an online course that teaches people about key concepts and terminology of the Trust and Safety industry, including the policy-making process, how engineering teams work with DevOps teams to enqueue content for human review, and measurements of precision, recall, and inter-rater reliability. We have experience creating online courses: the PI has a Coursera MOOC course that more than 50,000 learners have completed. Reviewers who work for 100 hours or more will be eligible for ten hours of paid time while completing the course. To measure the success of this course, we will conduct an exit survey asking participants about their interest level in computing careers, especially in Trust and Safety, and whether the course has increased their confidence in their ability to pursue such careers.

The second way that we will contribute to broadening participation is through inclusion of undergraduates in the project. We will recruit from pools that are currently under-represented in computing. MIT runs the Summer Research Program (MSRP) which seeks to improve the research enterprise through increased diversity; Senior personnel Rand has supervised students through this program in previous summers. Senior personnel Zhang runs the DUB summer internship program for Human Computer Interaction & Design at the University of Washington, which explicitly seeks broad participation, including members of underrepresented groups. Senior personnel Mihalcea runs the Girls Encoded initiative aiming to increase the number of women in computing, which includes the Explore Computer Science Research program for research experiences for underrepresented groups; the Discover CS introductory class to recruit more women into computing; and the Women in Computing seminar at Michigan. At the University of Michigan, we will hire an intern from either the Bachelor or Master of Science in Information programs; more than half the students in the HCI tracks in both programs identify as women.

10 Results from Prior NSF Support

Paul Resnick was PI and **Amy Zhang** was co-PI on the Phase I grant for this project, "NSF Convergence Accelerator Track F: Misinformation Judgments with Public Legitimacy" (**2137469**, $750,000 for 10/01/2021 - 09/30/2022). *Intellectual Merit*: This project explored how to create "golden sets" for misinformation, backed by jury judgments that would be perceived as publicly legitimate, and how to create a forecasting service that would predict jury judgments. We conducted public surveys to determine conditions of legitimacy and developed an initial version of the forecasting service before deciding to pivot to the Phase II approach. *Broader Impacts*: The intention was that providing a way to make misinformation enforcement decisions that would be perceived as publicly legitimate would free social media platforms to innovate in their detection procedures,

leading to improvements in enforcement and reducing the spread of harmful misinformation.

Adam Berinsky is PI and **David Rand** is co-PI on an NSF-funded project, "Promoting Accurate Information on Social Media" (**FAIN-2047152**, $737,381 for 07/01/2021 - 06/30/2023). *Intellectual Merit*: This project aims to unify hitherto separate lines of work studying the relationship between belief in (mis)information and cognitive reflection (typically found by cognitive scientists to be associated with having more accurate beliefs) versus political knowledge (typically found by political scientists to be associated with more polarized, rather than more accurate, beliefs). We are also testing the efficacy of interventions that prompt people to consider accuracy before sharing news. *Broader Impacts*: This proposal to help create a bridge between cognitive scientists and political scientists, who rarely interact despite often studying similar topics.

David Jurgens was PI on an NSF-funded project, "Explainable Recognition of Social Relationships from People's Linguistic Interactions" (RI-1850221, $174,999 for 3/15/2019 - 3/14/2022). *Intellectual merit*: This project investigated social relationships and their impact on language. The project introduced new NLP methods for recognizing the type of social relationship from interactions and explaining the sociolinguistic cues that signal these relationships. Results from the project led to seven papers [9; 62; 88; 89; 3; 71; 60]. *Broader Impact:* This project enabled testing and extending social science theories about relationships and social organization and developed new NLP models for recognizing veiled incivility in cases where the linguistic cues violate social norms for a relationship.

Rada Mihalcea was PI on the NSF-funded grant "RI: Small: Demographic-Aware Lexical Semantics" (NSF award IIS-1815291, 09/01/2018-08/31/2021, $450,000). This project explores methods where people-centric information complements language-based information. *Intellectual Merit*: We created demographic-aware word representations, and explored the extent to which they have an impact on certain user groups (e.g., minority vs majority groups). We introduced a methodology to create simple or compositional demographic word embeddings [26; 82; 81]. We also explored the role played by these representations in downstream tasks, including part-of-speech tagging and parsing [25], language models [83], and humor generation [24]. *Broader Impacts*: The project has had an impact on the performance of NLP applications that make use of word embeddings, and has positively impacted the performance of such applications on minority groups.

Scott Hale was PI on "NSF Convergence Accelerator Track F: FACT CHAMP - Fact-checker, Academic, and Community Collaboration Tools: Combating Hate, Abuse, and Misinformation with Minority-led Partnerships" (**NSF-49100421C0035**, $749,974 for 9/15/2021–9/14/2022). *Intellectual Merit* The team conducted interviews and pilot experiments to understand how to better support misinformation response effects in Asian-American and Pacific Islander (AAPI) communities. The team identified misinformation targeting AAPI communities that is off the radar of mainstream fact-checkers and journalists, built claim matching models in Asian languages, and supported pilot activities. *Broader Impacts* AAPI and other minoritized communities are underserved by current efforts. The project envisioned combinations of human and machine intelligence to better empower AAPI community organizations working to counter misinformation.

Benji Loney and **Veronica Perez-Rosas** do not have prior NSF funding.

References

[1] AVAAZ. How facebook can flatten the curve of the coronavirus infodemic. https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/, 2020.

[2] BANSAK, K., HAINMUELLER, J., HOPKINS, D. J., YAMAMOTO, T., DRUCKMAN, J. N., AND GREEN, D. P. Conjoint survey experiments. *Advances in Experimental Political Science 19* (2021).

[3] BAO, J., WU, J., ZHANG, Y., CHANDRASEKHARAN, E., AND JURGENS, D. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference* (2021).

[4] BERNSTEIN, M. S., LITTLE, G., MILLER, R. C., HARTMANN, B., ACKERMAN, M. S., KARGER, D. R., CROWELL, D., AND PANOVICH, K. Soylent: a word processor with a crowd inside. *Communications of the ACM 58*, 8 (2015), 85–94.

[5] BHUIYAN, M. M., ZHANG, A. X., SEHAT, C. M., AND MITRA, T. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction 4*, CSCW2 (2020), 1–26.

[6] CAMPOS, R., MANGARAVITE, V., PASQUALI, A., JORGE, A., NUNES, C., AND JATOWT, A. YAKE! keyword extraction from single documents using multiple local features. *Inf. Sci. 509* (Jan. 2020), 257–289.

[7] CHEN, Q., WELD, D., AND ZHANG, A. Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty. *Proc. ACM Hum.-Comput. Interact.*, CSCW (Oct. 2021).

[8] CHILTON, L. B., LITTLE, G., EDGE, D., WELD, D. S., AND LANDAY, J. A. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), ACM, pp. 1999–2008.

[9] CHOI, M., BUDAK, C., ROMERO, D., AND JURGENS, D. More than meets the tie: Examining the role of interpersonal relationships in social networks. In *Proceedings of the International AAAI Conference on Web and Social Media* (2021), vol. 15, pp. 105–116.

[10] CLAIMREVIEW. Claim review project. https://www.claimreviewproject.com/.

[11] CLAIMREVIEW. Claim review schema. https://schema.org/ClaimReview.

[12] COMMISSION, E. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions: European commission guidance on strengthening the code of practice on disinformation. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2021:262:FIN, 2021.

[13] CONNEAU, A., KHANDELWAL, K., GOYAL, N., CHAUDHARY, V., WENZEK, G., GUZMÁN, F., GRAVE, E., OTT, M., ZETTLEMOYER, L., AND STOYANOV, V. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), Association for Computational Linguistics, pp. 8440–8451.

[14] CRAWFORD, K., AND GILLESPIE, T. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society 18*, 3 (2016), 410–428.

[15] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.

[16] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North Association of Computational Linguistics* (2019).

[17] DI DOMENICO, G., SIT, J., ISHIZAKA, A., AND NUNAN, D. Fake news, social media and marketing: A systematic review. *Journal of Business Research 124* (2021), 329–341.

[18] DOR, L. E., HALFON, A., GERA, A., SHNARCH, E., DANKIN, L., CHOSHEN, L., DANILEVSKY, M., AHARONOV, R., KATZ, Y., AND SLONIM, N. Active learning for bert: an empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020), pp. 7949–7962.

[19] FACEBOOK. Facebook oversight board. `https://oversightboard.com/`, 2022. [Accessed: May. 17, 2022].

[20] FAN, J., AND ZHANG, A. X. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (2020), pp. 1–14.

[21] FARRELL, J., MCCONNELL, K., AND BRULLE, R. Evidence-based strategies to combat scientific misinformation. *Nature climate change 9*, 3 (2019), 191–195.

[22] FLINTHAM, M., KARNER, C., BACHOUR, K., CRESWICK, H., GUPTA, N., AND MORAN, S. Falling for fake news: investigating the consumption of news via social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–10.

[23] FLORESCU, C., AND CARAGEA, C. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. *Proceedings of the 55th Annual Meeting* (2017).

[24] GARIMELLA, A., BANEA, C., HOSSAIN, N., AND MIHALCEA, R. Judge me by my size (noun), do you? yodalib: A demographic-aware humor generation framework. In *Proceedings of the International Conference on Computational Linguistics (COLING 2020)* (2020).

[25] GARIMELLA, A., BANEA, C., HOVY, D., AND MIHALCEA, R. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019).

[26] GARIMELLA, A., BANEA, C., AND MIHALCEA, R. Demographic-aware word associations. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)* (Copenhagen, Denmark, 2017).

[27] GIFCT. What is hash sharing? `https://vimeo.com/564638166?embedded=true&source=vimeo_logo&owner=143145801`.

[28] GORDON, M. L., LAM, M. S., PARK, J. S., PATEL, K., HANCOCK, J., HASHIMOTO, T., AND BERNSTEIN, M. S. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–19.

[29] GORWA, R., BINNS, R., AND KATZENBACH, C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society 7*, 1 (2020), 2053951719897945.

[30] GRAY, M. L., AND SURI, S. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

[31] GUARDIAN. Facebook, google and twitter grilled by congress over russian meddling – as it happened. "`https://www.theguardian.com/technology/live/2017/oct/31/facebook-google-twitter-congress-russian-election-meddling-live`", 2017.

[32] HAIMSON, O. L., DELMONACO, D., NIE, P., AND WEGNER, A. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW2 (2021), 1–35.

[33] HAIST, F., SHIMAMURA, A. P., AND SQUIRE, L. R. On the relationship between recall and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition 18*, 4 (1992), 691.

[34] HU, X. E., WHITING, M. E., AND BERNSTEIN, M. S. Can online juries make consistent, repeatable decisions? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–16.

[35] IWF. Internet watch foundation intelligrade. `https://www.iwf.org.uk/our-technology/intelligrade/`.

[36] JU, J.-H., YANG, J.-H., AND WANG, C.-J. Text-to-text multi-view learning for passage reranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021), pp. 1803–1807.

[37] KAIRAM, S., AND HEER, J. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (2016), pp. 1637–1648.

[38] KAZEMI, A., GARIMELLA, K., GAFFNEY, D., AND HALE, S. Claim matching beyond English to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Online, Aug. 2021), Association for Computational Linguistics, pp. 4504–4517.

[39] KIM, J., STERMAN, S., COHEN, A. A. B., AND BERNSTEIN, M. S. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing* (2017), pp. 233–245.

[40] KITTUR, A., SMUS, B., KHAMKAR, S., AND KRAUT, R. E. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology* (2011), ACM, pp. 43–52.

[41] KRISHNAN, N., GU, J., TROMBLE, R., AND ABROMS, L. C. Research note: Examining how various social media platforms have responded to covid-19 misinformation. *Harvard Kennedy School Misinformation Review 2*, 6 (2021), 1–25.

[42] LAZER, D. M. J., BAUM, M. A., BENKLER, Y., BERINSKY, A. J., GREENHILL, K. M., MENCZER, F., METZGER, M. J., NYHAN, B., PENNYCOOK, G., ROTHSCHILD, D., SCHUDSON, M., SLOMAN, S. A., SUNSTEIN, C. R., THORSON, E. A., WATTS, D. J., AND ZITTRAIN, J. L. The science of fake news. *Science 359*, 6380 (2018), 1094–1096.

[43] LEWANDOWSKY, S., ECKER, U. K., AND COOK, J. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition 6*, 4 (2017), 353–369.

[44] LI, C., RESNICK, P., AND MEI, Q. Multiple queries as bandit arms. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (2016), pp. 1089–1098.

[45] LI, C., WANG, Y., RESNICK, P., AND MEI, Q. Req-rec: high recall retrieval with query pooling and interactive classification. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (2014), pp. 163–172.

[46] LITTLE, G., CHILTON, L. B., GOLDMAN, M., AND MILLER, R. C. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (2010), ACM, pp. 57–66.

[47] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[48] LO, K. Toolkit for civil society and moderation inventory. `https://meedan.com/reports/toolkit-for-civil-society-and-moderation-inventory/`, 2020.

[49] LOOMBA, S., DE FIGUEIREDO, A., PIATEK, S. J., DE GRAAF, K., AND LARSON, H. J. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour 5*, 3 (2021), 337–348.

[50] LYONS, T. Hard Questions: What's Facebook's Strategy for Stopping False News? `https://about.fb.com/news/2018/05/hard-questions-false-news/`, May 2018.

[51] MALONE, T. W., AND BERNSTEIN, M. S. *Handbook of collective intelligence*. MIT press, 2015.

[52] MEEDAN. Alegre. `https://github.com/meedan/alegre`.

[53] META. Detecting violations. `https://transparency.fb.com/enforcement/detecting-violations/`.

[54] META. Facebook community standards: Misinformation. `https://transparency.fb.com/policies/community-standards/misinformation/`.

[55] MIHALCEA, R., AND TARAU, P. Textrank: Bringing order into text. *Proceedings of the 2004 conference on* (2004).

[56] MOLTER, V. Platforms of babel: Inconsistent misinformation support in non-english languages. https://www.eipartnership.net/policy-analysis/inconsistent-efforts-against-us-election-misinformation-in-non-english, 2020.

[57] MOSSERI, A. Working to stop misinformation and false news. https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news, 2017.

[58] NANGIA, N., SUGAWARA, S., TRIVEDI, H., WARSTADT, A., VANIA, C., AND BOWMAN, S. R. What ingredients make for an effective crowdsourcing protocol for difficult nlu data collection tasks?, 2021.

[59] PAN, C. A., YAKHMI, S., IYER, T. P., STRASNICK, E., ZHANG, A. X., AND BERNSTEIN, M. S. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction 6*, CSCW1 (2022), 1–31.

[60] PARK, C. Y., MENDELSOHN, J., RADHAKRISHNAN, K., JAIN, K., KANAKAGIRI, T., JURGENS, D., AND TSVETKOV, Y. Detecting community sensitive norm violations in online conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (2021), pp. 3386–3397.

[61] PARTNERSHIP, E. I. The long fuse: Misinformation and the 2020 election. https://purl.stanford.edu/tr171zs0069, 2021.

[62] PEI, J., AND JURGENS, D. Quantifying intimacy in language. In *Proceedings of EMNLP* (2020).

[63] PENNYCOOK, G., EPSTEIN, Z., MOSLEH, M., ARECHAR, A. A., ECKLES, D., AND RAND, D. G. Shifting attention to accuracy can reduce misinformation online. *Nature 592*, 7855 (2021), 590–595.

[64] RAFFEL, C., SHAZEER, N., ROBERTS, A., LEE, K., NARANG, S., MATENA, M., ZHOU, Y., LI, W., AND LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research 21* (2020), 1–67.

[65] RETELNY, D., ROBASZKIEWICZ, S., TO, A., LASECKI, W. S., PATEL, J., RAHMATI, N., DOSHI, T., VALENTINE, M., AND BERNSTEIN, M. S. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology* (2014), ACM, pp. 75–85.

[66] ROBERTS, S. T. *Behind the Screen*. Yale University Press, 2019.

[67] SCHRÖDER, C., AND NIEKLER, A. A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267* (2020).

[68] SCHWARZ, N., AND JALBERT, M. When (fake) news feels true: Intuitions of truth and the acceptance and correction of misinformation. In *The Psychology of Fake News*. Routledge, 2020, pp. 73–89.

[69] SHAAR, S., BABULKOV, N., DA SAN MARTINO, G., AND NAKOV, P. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), Association for Computational Linguistics, pp. 3607–3618.

[70] SIMKO, J., RACSKO, P., TOMLEIN, M., HANAKOVA, M., MORO, R., AND BIELIKOVA, M. A study of fake news reading and annotating in social media context. *New Review of Hypermedia and Multimedia 27*, 1-2 (2021), 97–127.

[71] SKY, C.-W., AND JURGENS, D. Using sociolinguistic variables to reveal changing attitudes towards sexuality and gender. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), pp. 9918–9938.

[72] STEVENSON, A. Facebook admits it was used to incite violence in myanmar. *New York Times* (2018).

[73] TWITTER. How we address misinformation on twitter. `https://help.twitter.com/en/resources/addressing-misleading-info`.

[74] TWITTER. Covid-19 misleading information policy. `https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy`, 2021. [Accessed: May. 17, 2022].

[75] TWITTER. Synthetic and manipulated media policy. `https://help.twitter.com/en/rules-and-policies/manipulated-media`, 2021. [Accessed: Dec. 28, 2021].

[76] VALENCIA, S. Misinformation online is bad in english. but it's far worse in spanish. `https://www.washingtonpost.com/outlook/2021/10/28/misinformation-spanish-facebook-social-media/`, 2021.

[77] VOSOUGHI, S., ROY, D., AND ARAL, S. The spread of true and false news online. *Science 359*, 6380 (2018), 1146–1151.

[78] WANG, Y., MCKEE, M., TORBICA, A., AND STUCKLER, D. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine 240* (2019), 112552.

[79] WANG, Z. J., CHOI, D., XU, S., AND YANG, D. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing* (Online, 2021), Association for Computational Linguistics, pp. 47–52.

[80] WARDLE, C., AND DERAKHSHAN, H. Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. In *Journalism,'fake news'& disinformation*, C. Ireton and J. Posetti, Eds. Unesco, Paris, 2018, ch. 2, pp. 43–54.

[81] WELCH, C., KUMMERFELD, J., PEREZ-ROSAS, V., AND MIHALCEA, R. Compositional demographic word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2020).

[82] WELCH, C., KUMMERFELD, J., PEREZ-ROSAS, V., AND MIHALCEA, R. Exploring the value of personalized word embeddings. In *Proceedings of the International Conference on Computational Linguistics (COLING 2020)* (2020).

[83] WELCH, C., PEREZ-ROSAS, V., KUMMERFELD, J. K., AND MIHALCEA, R. Learning from personal longitudinal dialog data. *IEEE Intelligent Systems 34*, 4 (2019).

[84] WIKIPEDIA. Wikipedia:no original research. `https://en.wikipedia.org/wiki/Wikipedia:No_original_research/`, 2022. [Accessed: May. 17, 2022].

[85] YANG, W., WANG, X., LU, J., DOU, W., AND LIU, S. Interactive steering of hierarchical clustering. *IEEE Trans. Vis. Comput. Graph. 27*, 10 (Oct. 2021), 3953–3967.

[86] ZAUNER, C. Implementation and benchmarking of perceptual image hash functions. Master's thesis, Upper Austria University of Applied Sciences, 2010.

[87] ZHANG, A. X., RANGANATHAN, A., METZ, S. E., APPLING, S., SEHAT, C. M., GILMORE, N., ADAMS, N. B., VINCENT, E., LEE, J., ROBBINS, M., BICE, E., HAWKE, S., KARGER, D., AND MINA, A. X. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018* (2018), pp. 603–612.

[88] ZHOU, N., AND JURGENS, D. Condolence and empathy in online communities. In *Proceedings of EMNLP* (2020).

[89] ZHU, J., AND JURGENS, D. The structure of online social networks modulates the rate of lexical change. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2021), pp. 2201–2218.

[90] ZHUANG, S., LI, H., AND ZUCCON, G. Deep query likelihood model for information retrieval. In *European Conference on Information Retrieval* (2021), Springer, pp. 463–470.