The Community Notes Observatory: Can Crowdsourced Fact-Checking be Trusted in Practice?

Luca Righes
EURECOM
France
luca.righes@eurecom.fr

Mohammed Saeed
EURECOM
France
mohammed.saeed@eurecom.fr

Gianluca Demartini
University of Queensland
Australia
demartini@acm.org

Paolo Papotti
EURECOM
France
papotti@eurecom.fr

ABSTRACT

Fact-checking is an important tool in fighting online misinformation. However, it requires expert human resources, and thus does not scale well on social media because of the flow of new content. Crowdsourcing has been proposed to tackle this challenge, as it can scale with a smaller cost, but it has always been studied in controlled environments. In this demo, we present the Community Notes Observatory, an online system to evaluate the first large-scale effort of crowdsourced fact-checking deployed in practice. We let demo attendees search and analyze tweets that are fact-checked by Community Notes users and compare the crowd's activity against professional fact-checkers. The attendees will explore evidence of i) differences in how the crowd and experts select content to be checked, ii) how the crowd and the experts retrieve different resources to fact-check, and iii) the edge the crowd shows in fact-checking scalability and efficiency as compared to expert checkers.

KEYWORDS

crowdsourcing, fact-checking, misinformation, information quality

1 INTRODUCTION

The spread of online misinformation carries risks for the democratic process [14]. Fact-checking is a prominent solution in fighting online misinformation. However, traditional fact-checking is a process requiring scarce expert human resources, and thus does not scale well to social media because of the continuous flow of new content. Automated methods and crowdsourcing have been proposed to tackle this challenge [10, 11], as they scale with a smaller cost, but have always been studied in controlled environments.

Twitter has started Community Notes as the first large-scale effort of crowdsourced fact-checking in January 2021 [3][1]. Community Notes adopts a community-driven approach for fact-checking by allowing selected Twitter users to identify fallacious information by (i) classifying tweets as misleading or not, accompanied by a written review, and by (ii) classifying reviews of other Community Notes users as being helpful or not.

In this demo, we show how crowdsourced fact-checking works in practice when compared with human experts. We provide a Web interface[2] where users can search the tweets fact-checked both by Community Notes users and professionals mitigators. Our tool allows the attendees to compare how the two approaches differ in how content is selected to be checked, which sources of evidence are used, and the ultimate fact-checking outcome.

In the rest of the paper, we first give some background information on fact-checking and crowdsourcing (Section 2), we introduce the datasets collected for our study (Section 3), and we report our main results (Section 4). Finally, we describe the application and the interactive use cases that we will demonstrate (Section 5).

2 BACKGROUND

The fact-checking process starts with identifying check-worthy claims and ends with a label about their veracity. Labels vary across services but usually can be divided into four popular categories: true, partially-true, false, or not enough evidence to judge. Given an input textual tweet, both Community Notes crowd and expert fact-checkers can be used to assess if it is worth checking and eventually verified. We describe next the three main steps in the pipeline.

Claim Selection. Users label content that violates the guidelines of the site, such as hate speech and misinformation. This process triggers the human verification with moderators hired by the platform [4]. For human fact-checkers the selection of the claims to verify is driven by journalistic principles, such as importance and if the claim contains a verifiable fact.

[1]The program changed name to from *Birdwatch* to *Community Notes* in December 2022. Our data has been collected before the name change.
[2]Prototype available at https://birdwatch.eurecom.fr

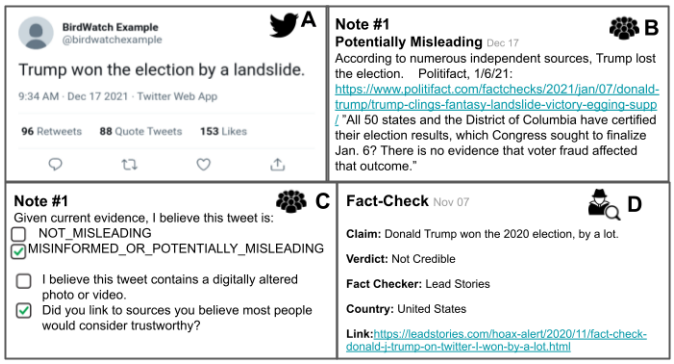

Figure 1: Community Notes note and ClaimReview fact-check example. (A) shows a tweet. (B) is the note with the assigned label to such tweet. (C) is a sample of questions when submitting a note. (D) shows a fact-check delivered by an expert.

Evidence Retrieval. The crowd makes use of expert fact-checking outcomes when available, otherwise they use evidence from the Web with the risk of being influenced by their own personal belief and context [11]. Expert fact-checkers instead rely on their training to identify proven, verified, transparent, and accountable evidence, sometimes involving third-party domain experts.

Claim verification. When misinformation is identified on social media, crowd users tend to counter it by providing evidence of it being misleading [9]. This shows an intrinsic motivation that certain members of the crowd have to contribute to the fact-checking process. Most expert fact-checkers work within organizations that are part of the International Fact-Checking Network, which sets editorial standards on the verification protocol.

3 DATA

Community-driven fact-checking on Twitter is governed by the Community Notes initiative [3], while fact-checks written by journalists and expert fact-checkers are curated using the ClaimReview schema [2]. We describe both datasets and how to match similar claims identified by both parties in this section. More details on the data collection process are in the full paper for this project [12].

3.1 Community Notes

In the Community Notes program participants identify misleading tweets and provide feedback using *Notes* and *Ratings*.

Notes. Participants can add notes to any tweet (example in Figure 1(A)). Their notes are formed from: (i) a classification label indicating whether the tweet is misinformed/misleading (**MM**) or not misleading (**NM**) according to their judgement with and an open text field where they justify their label and possibly include links to sources Figure 1(B); (ii) answers to multiple-choice questions about their decision (example in Figure 1(C)). The key data we use from the notes are:

- *Classification Label*: Whether the tweet is misleading or misinformed (MM) or not (NM) according to the user.
- *Note Text*: the text with the user justification for the label.
- *Timestamps*: time at which the note was written.

Ratings. Participants rate the notes of other participants to help identify which notes are helpful. A user rates a note by providing answers to a list of questions; we focus on two of them:

- *High-quality Sources:* The user answers the question 'Is this note helpful because it cites high-quality sources?'. We use this information to assess if users distinguish credible sources.
- *Helpfulness Label:* The user answers the question 'Is this note helpful?'. We use this information to compute a helpfulness score for notes.

We use the Community Notes data up to September 2021. The dataset contains 87k ratings for 12k tweets (15k notes) from 5k unique participants.

3.2 ClaimReview

The ClaimReview project [2] is a schema used to publish fact-checking articles by organizations and journalists. Our dataset is a collection of items following this schema, collected from various sources [8]. Each item, or *fact-check*, is a (claim, label) pair produced by a professional journalist or fact-checking agency. Since different fact-checkers use different labels, the data is normalized into a smaller subset of labels. We use a dataset containing 77k fact-checks. A fact-check is shown in Figure 1 (D).

3.3 Matched Data

To study how the judgements of the crowd compare to those of expert fact-checkers, we match claims from both datasets, obtaining 2208 matches. We rely on an unsupervised text-to-text matching algorithm [13] to bootstrap the pairing of the text of the tweets and the claim in ClaimReview fact-checks. These candidate matches are then manually verified with the Amazon Mechanical Turk crowdsourcing platform [1]. Each tweet was shown to 3 workers, similar to previous work [6, 7]. An example of a tweet matching a ClaimReview is shown in Figure 1.

4 RESULTS

Our dataset of manually annotated tweets, verified both by Community Notes users and expert journalists, enables us to analyze and contrast the approaches across the main dimensions in the standard fact-checking pipeline.

4.1 Claim Selection

We show how Community Notes users identify check-worthy claims in comparison with fact-checking experts. We predict the topic of every Community Notes tweet and ClaimReview fact-check using BerTopic [5], and plot the frequency distribution of four topics, on a monthly basis, in Figure 2. The high count of Community Notes tweets and ClaimReview fact-checks covering political tweets show that they both consider the *Politics* topic important. The similar trends for all topics suggest that both methods react similarly to news and major events in terms of claim selection.

4.2 Evidence Retrieval

Both crowd checkers and experts report the sources used in their verification process. Community Notes participants use only 17 domains in common to those of the ClaimReview experts. The other

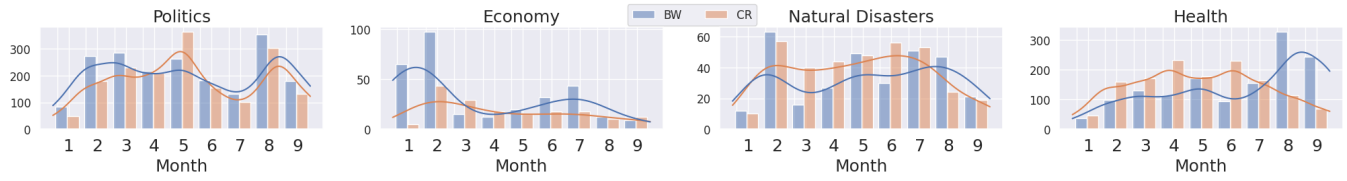

Figure 2: Per-topic frequency histograms and KDE Plots for Community Notes (BW) notes and ClaimReview (CR) facts.

		BirdWatch				
		Notes		Tweets		
		MM	NM	MM	NM	Tie
Claim Review	credible	209	**25**	126	**9**	9
	mostly_credible	56	**14**	44	**7**	5
	not_credible	**1983**	184	**1476**	62	55
	not_verifiable	300	25	225	8	9
	uncertain	225	22	156	8	9

Table 1: Matching the classification labels across Community Notes and ClaimReview on the note level and the tweet level (obtained through majority voting). Agreement in bold.

56 ClaimReview domains, which are not in the overlap, include 53 local resources, such as news outlets, for non US countries as fact-checking organizations work at a global scale and Community Notes focuses on US. Community Notes sources are a larger number as they range from Wikipedia and YouTube videos to medical websites and research papers.

We also compare ratings of source quality of Community Notes users to those of expert fact-checkers using an external tool (https://www.newsguardtech.com), and we refer to it as the *journalist score*. For note links rated as high-quality by Community Notes users, we observe high journalist scores, however, a considerable number of sources rated as low-quality have a high journalist score. These notes are debunking news about US politics and are effectively reliable sources, but a significant fraction of Community Notes users labeled such links as low-quality. This shows how some Community Notes users convey partisanship, forming a group of people trying to deceive the Community Notes program to serve their common interest.

4.3 Claim Verification

We ponder whether Community Notes participants provide accurate judgements. We first compare agreement (i) among themselves and then (ii) with ClaimReview expert fact-checkers. We then analyze different scoring functions for note aggregation, and finally discuss the temporal relationships between notes and fact-checks.

4.3.1 Internal Agreement. We use the participants' classification labels to see whether the tweet is classified as misinformed or not. We observe that most tweets have two note counts and the majority of users agree on the final classification label.

4.3.2 External Agreement. After matching Community Notes data with ClaimReview fact-checks, we compare their labels. Table 1 shows that the majority of ClaimReview labels match the Community Notes ones. There are 1492 (9+7+1476) decisions with the same classification label and 232 (126+44+62) with different labels. Among the 209 notes that are labeled as credible by the ClaimReview fact-checks and misinformed by the Community Notes participants, the most common cause are texts with multiple claims, i.e., multiple facts are reported in a tweet and the fact-checked claims differ.

4.3.3 Temporal Analysis. We analyze tweets (T), Community Notes notes (B), and ClaimReview fact-checks (C) time-wise. As a note can only occur after a tweet, we have three different configurations: (i) Tweet occurs first, then Community Notes note, then ClaimReview fact-check (TBC), (ii) Tweet then ClaimReview fact-check then Community Notes note (TCB), and (iii) ClaimReview fact then Tweet then Community Notes note (CTB). The most interesting results in that for **TBC**, there are 129/2208 tweets for which Community Notes users provide a response much faster than experts. On average, a Community Notes provides a response 10X faster than an expert. These examples show how Community Notes participants can fact-check claims with reliable sources without the need of ClaimReview fact-checks. For **TCB**, a ClaimReview rarely occurs after a tweet and before a Community Notes. The majority of the matched tweets follow the **CTB** pattern, with most of them related to US politics and COVID-19. As Twitter is an open space, several users tend to spread false news even after they have been fact-checked.

5 DEMONSTRATION

In our demonstration, we use a Web service. The backend has been implemented using the Flask framework for exposing a RESTful API to the frontend and MongoDB for storing the datasets containing the notes and the claim reviews. The interactions between the two entities is managed through the Pymongo library that enables the exchange of data across the database and the backend.

The users will be able to explore and analyze tweets, Community Notes notes, and their corresponding matched fact-checks as shown in Table 1. We will organize the demonstration around four main challenges presented next. To address the challenges, the users can explore the tweets by looking up keywords in the search bar (Figure 3). For every tweet, we display (i) the associated tweets and their classification label, and (ii) the matched claim reviews. Rolling over the text with the pointer will enable users to dig into the details for tweets, notes and claims reviews, such as the timestamp, the author, the journalistic and high-quality crowd scores for the sources, and the helpfulness score.

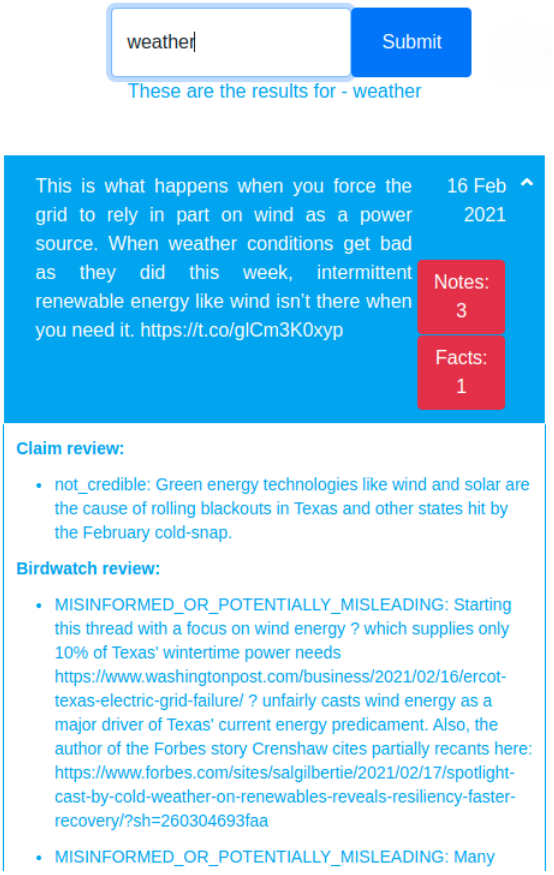

Figure 3: Results for a keyword search in our application showing both Community Notes notes and ClaimReview checks for a tweet.

1. Claim selection challenge. What is neglected by professional fact-checkers? We will ask the attendees to identify an example of claim that is checked only by the crowd and an example of claim checked only by experts. Analyzing the search results, users will observe that crowd and experts align in terms of topics for the claim verification.

2. Source quality challenge. What are the good sources for fact-checking? We will ask attendees to identify an example of claim for which the crowd provides better sources of evidence than the ones reported for the same claim in the expert fact-check. Looking at the source scores, attendees will learn that expert fact-checkers rely on a relatively small set of high-quality sources to verify claims, while Community Notes participants provide a variety of sources that seem to be neglected by fact-checkers. However, we will show with examples that, while most of the crowd's sources are evaluated as credible (by journalists) and useful (by the Community Notes user ratings), malicious users might game the algorithm and effectively label notes as unhelpful according to their ideology and beliefs.

3. Disagreement challenge. What are the most controversial topics that divide the crowd? We will ask the attendees to identify an example of a divisive claim. Attendees will see how Community Notes participants show high levels of agreement in the final classification label in the majority of cases. However, the Community Notes crowd do gather conflicting notes on specific topics, such as elections.

4. The time challenge. Can the attendee find a claim that is debunked by the crowd before a professional fact-checker debunk it? To guide the attendees, we will remark that the system allows the search according to timestamps for tweets, Community Notes notes, and ClaimReview fact-checks. Moreover, this pattern can be observed only for events happened after January 2021 (starting month for Community Notes). The attendees will notice that the crowd can identify and check tweets with misleading claims even before they get fact-checked by an expert. However, the majority of the matched tweets have an already fact-checked claim review, with most of them related to US politics and COVID-19. As Twitter is an open space, several users tend to spread false news even after they have been fact-checked.

Beyond the standard interface, we will show to the visitors more details about the process of matching tweets to claim reviews, with examples showing the limits of the ClaimReview method. Moreover, attendees with more time will be able to analyze the different topics that Community Notes participants and ClaimReview experts portray. We provide graphs of the variation of a certain topic (such as Politics and Health) over time, as shown in Figure 2.

Acknowledgments. This work is partially supported by gifts from Google, by CHIST-ERA within the CIMPLE project (CHISTERA-19-XAI-003), and by the ARC Training Centre for Information Resilience (Grant No. IC200100022).

REFERENCES

[1] [n.d.]. Amazon Mechanical Turk. https://www.mturk.com.
[2] [n.d.]. ClaimReview Project. https://www.claimreviewproject.com/the-facts-about-claimreview.
[3] [n.d.]. Community Notes. https://communitynotes.twitter.com/.
[4] Shiri Dori-Hacohen, Keen Sung, Jengyu Chou, and Julian Lustig-Gonzalez. 2021. *Restoring Healthy Online Discourse by Detecting and Reducing Controversy, Misinformation, and Toxicity Online.* ACM, 2627–2628.
[5] Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. https://doi.org/10.5281/zenodo.4381785
[6] Gabriella Kazai. 2011. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *ECIR*. Springer-Verlag, 165–176.
[7] Di Liu, Randolph G. Bias, Matthew Lease, and Rebecca Kuipers. 2012. Crowdsourcing for usability testing. *Proc. Assoc. Inf. Sci. Technol.* 49, 1 (2012), 1–10.
[8] Martino Mensio and Harith Alani. 2019. MisinfoMe: Who is Interacting with Misinformation?. In *ISWC (CEUR Workshop Proceedings, Vol. 2456)*. 217–220.
[9] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. In *Big Data*. 748–757.
[10] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *IJCAI* (2021).
[11] Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020. The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?. In *CIKM*. ACM, 1305–1314.
[12] Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare With Experts?. In *CIKM*. ACM, 1736–1746. https://doi.org/10.1145/3511808.3557279
[13] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *ACL*. 3607–3618.
[14] Kate Starbird. 2019. Disinformation's spread: bots, trolls and all of us. *Nature* 571, 7766 (2019), 449–450.