Crowdsourced Fact-Checking: How is BirdWatch Doing?

Mohammed Saeed
EURECOM, France
mohammed.saeed@eurecom.fr

Gianluca Demartini
University of Queensland, Australia
demartini@acm.org

Paolo Papotti
EURECOM, France
papotti@eurecom.fr

Introduction

Fact-checking is one of the prominent solutions in fighting the spread of online misinformation, which carries risks for the democratic process (Starbird, 2019). 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 (Hassan et al., 2015). As computational fact-checking is still not ready for adoption (Nakov et al., 2021), crowdsourcing has been proposed to tackle this challenge (Roitero et al., 2020a). Using the crowd for this task scales with a smaller cost, but has always been studied in controlled environments. Twitter has started BIRDWATCH as the first large-scale effort of crowdsourced fact-checking in January 2021 (BWP, 2021). BIRDWATCH adopts a community-driven approach for fact-checking by allowing selected Twitter users to identify fallacious information.

In this talk, we look at how crowdsourced fact-checking works in practice when compared with human experts. We report on the following research questions:

RQ1 How are check-worthy claims selected by BIRDWATCH users? Can the crowd identify check-worthy claims before experts do?

RQ2 What sources of information are used to support a fact-checking decision in BIRDWATCH and how reliable are they? Can the crowd be considered as "independent fact-checkers"?

RQ3 Are crowd workers able to reliably assess the veracity of a tweet? Is their assessment considered helpful by others?

In the rest of the talk proposal, we give background information on fact-checking and crowdsourcing, introduce the datasets collected in our study, and report our main results.

Background

The fact-checking process starts with identifying check-worthy claims and ends with a label about their veracity. Labels vary across services but can be divided into four categories: true, partially-true, false, or not enough evidence to judge. Given an input textual tweet, both BIRDWATCH crowd and expert checkers follow three main steps.

1. Claim Selection. Deciding whether a claim is worth checking is similar to the task of judging the relevance of a document w.r.t. a search query. The crowdsourced annotation of content on social networks is an activity across all platforms. 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 (Dori-Hacohen et al., 2021). For human fact-checkers the selection of the claims to verify is driven by journalistic principles, such as importance or if the claim contains a verifiable fact.

2. Evidence Retrieval. The crowd makes use of expert fact-checking outcomes when available, otherwise they use Web evidence with the risk of being influenced by their own personal belief and context (Roitero et al., 2020b). Expert fact-checkers instead rely on their training to identify verified, transparent, and accountable evidence, sometimes involving third-party domain experts.

3. Claim verification. When misinformation is identified on social media, crowd users tend to counter it by providing evidence of it being misleading (Micallef et al., 2020). 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, such those that are part of the International Fact-Checking Network, which sets editorial standards on the verification protocol.

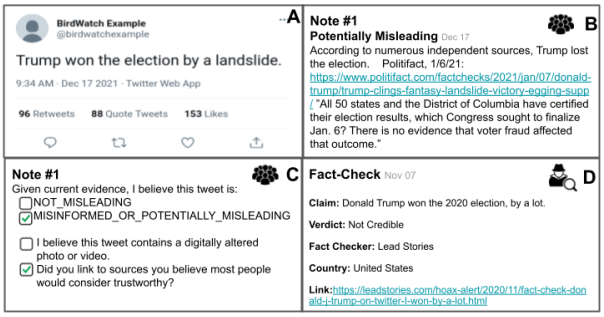

Figure 1: BIRDWATCH note and CLAIMREVIEW fact-check examples. (A) tweet. (B) note with the assigned label to such tweet. (C) sample of questions when submitting a note. (D) fact-check delivered by an expert.

Data

Community-driven fact-checking on Twitter is governed by the BIRDWATCH initiative (BWP, 2021), while checks written by journalists and expert checkers are curated using the CLAIMREVIEW schema (CRP, 2021). We describe the datasets and how to identify claims verified by both parties.

Birdwatch. In the BIRDWATCH program, participants identify misleading tweets and provide context using *Notes* and *Ratings*. Participants can add notes to any tweet, e.g., Figure 1(A). Their notes contain two elements. First, a classification label indicating whether the tweet is misinformed/misleading (MM) or not misleading (NM) with a text field where they justify their label and include links to sources, e.g., Figure 1(B). Second, answers to multiple-choice questions about their decision, e.g., Figure 1(C). The key data we use from the notes are three. The *Classification Label*: Whether the tweet is misleading or misinformed (MM) or not (NM) according to the user. The *Note Text* with the user justification for the label. The *Timestamps* at which the note was written.

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. The *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. The *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 BIRDWATCH data up to September 2021. The dataset contains 87k ratings for 12k tweets (15k notes) from 5k unique participants.

ClaimReview. The CLAIMREVIEW schema is used to publish fact-checking articles by organizations and journalists. Our dataset is a collection of items following this schema, collected from various sources (Mensio and Alani, 2019). 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).

Matched Data. To study how the judgements of the crowd compare to those of expert checkers, we matched claims from both datasets with a combination of computational methods and human annotators. After running the exercise, we are left with 2.2k tweets verified by BIRDWATCH that are also matching with the CLAIMREVIEW fact-checks. An example of a tweet matching a CLAIMREVIEW is shown in Figure 1.

Results

RQ1 BIRDWATCH users and CLAIMREVIEW experts show correlation in claim selection decisions w.r.t. major news and events, but with important differences due to the circulation of claims that have been already debunked by experts. The crowd seems to be effective also in identifying tweets with misleading claims even before they get fact-checked by an expert. Also, both popular and non-popular tweets get verified by BIRDWATCH users. Computing the check-worthiness of a tweet does not lead to effective results using current methods.

RQ2 Expert fact-checkers rely on a relatively small set of high-quality sources to verify claims, while BIRDWATCH participants provide a variety of sources that seem to be neglected by fact-checkers. While most of these sources are evaluated as credible (by journalists) and useful (by the BIRDWATCH crowd), malicious users might game the algorithm and effectively label notes as unhelpful according to their ideology and beliefs.

RQ3 BIRDWATCH users show high enough levels of agreement to reach decisions in the vast majority of cases. The crowd focuses mostly on misleading tweets and shows high agreement with expert fact-checkers in terms of classification label.

References

2021. Claimreview project. https://www.claimreviewproject.com/the-facts-about-claimreview.

2021. Introducing birdwatch, a community-based approach to misinformation. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

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*, page 2627–2628. Association for Computing Machinery, New York, NY, USA.

Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. In *Proceedings of the 2015 computation+ journalism symposium*.

Martino Mensio and Harith Alani. 2019. Misinfome: Who is interacting with misinformation? In *ISWC*, volume 2456 of *CEUR Workshop Proceedings*, pages 217–220. CEUR-WS.org.

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 *2020 IEEE International Conference on Big Data (Big Data)*, pages 748–757.

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*.

Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020a. Can the crowd identify misinformation objectively? the effects of judgment scale and assessor's background. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 439–448.

Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020b. The covid-19 infodemic: Can the crowd judge recent misinformation objectively? In *CIKM*, CIKM '20, page 1305–1314, New York, NY, USA. Association for Computing Machinery.

Kate Starbird. 2019. Disinformation's spread: bots, trolls and all of us. *Nature*, 571(7766):449–450.