Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare With Experts?

Mohammed Saeed
Eurecom
Biot, France
mohammed.saeed@eurecom.fr

Nicolas Traub
The University of Queensland
Brisbane, Australia
n.traubdamico@uq.net.au

Maelle Nicolas
Eurecom
Biot, France
Maelle.Nicolas@eurecom.fr

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

Paolo Papotti
Eurecom
Biot, France
paolo.papotti@eurecom.fr

ABSTRACT

Fact-checking is one of the effective solutions in fighting online misinformation. However, traditional fact-checking is a process requiring scarce expert human resources, and thus does not scale well on social media because of the continuous flow of new content to be checked. Methods based on crowdsourcing have been proposed to tackle this challenge, as they can scale with a smaller cost, but, while they have shown to be feasible, have always been studied in controlled environments. In this work, we study the first large-scale effort of crowdsourced fact-checking deployed in practice, started by Twitter with the Birdwatch program. Our analysis shows that crowdsourcing may be an effective fact-checking strategy in some settings, even comparable to results obtained by human experts, but does not lead to consistent, actionable results in others. We processed 11.9k tweets verified by the Birdwatch program and report empirical evidence of i) differences in how the crowd and experts select content to be fact-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.

ACM Reference Format:
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 *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3511808.3557279

1 INTRODUCTION

The spread of online misinformation carries risks for the democratic process and for a decrease in public trust towards authoritative sources of news [52]. Fact-checking is one of the prominent solutions 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 [29]. Automated methods and crowdsourcing have been proposed to tackle this challenge [41, 48, 54], as they can scale with a smaller cost, but have always been studied in controlled environments. Current approaches focus either on fully automated machine learning methods [38, 57] or on hybrid human-machine approaches making use of crowdsourcing to scale-up human annotation efforts [49].

The first large-scale effort of crowdsourced fact-checking was piloted by Twitter with the Birdwatch program on the 23rd of January 2021 [8]. Birdwatch 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 Birdwatch users as being helpful or not. In this setting, any user can create a *note* for a tweet (providing some metadata about the annotations) and other users can up/down *rate* such note. Multiple users can check the same content independently.

In this study, we perform an analysis of how crowdsourced fact-checking works in practice when compared with human experts and automated fact-checking methods. To this end, we perform an analysis of the grass-root fact-checking process in Birdwatch, including which content is selected to be fact-checked, which sources of evidence are used, and the fact-checking outcome. We also look at possible bias in terms of volume and topics as compared to experts. To enable a fair comparison across the three fact-checking approaches (i.e., computational methods, crowd, experts), we collected a dataset of 11.9k tweets with Birdwatch checks and identified 2.2k tweets verified both by Birdwatch users and expert journalists. This dataset enables us to analyze and contrast the three approaches across the main dimensions in the standard fact-checking pipeline (see Figure 1). We focus 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? Does the crowd always rely on data previously fact-checked by experts, or can they be considered as "independent fact-checkers"?

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

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557279

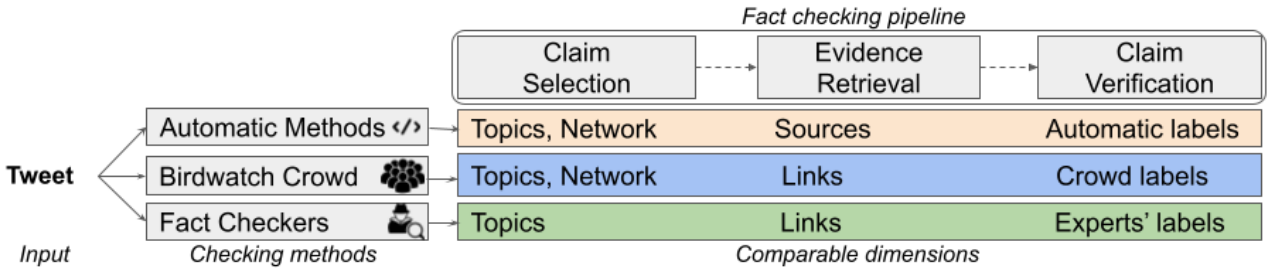

Figure 1: Given input tweets, the three alternative checking methods (automatic, crowd, professional checkers) are analyzed across their comparable dimensions according to a standard fact-checking pipeline.

Our results reveal insights from real data to answer these questions. As automatic methods are still not competitive for checking the truthfulness of online content, we focus on how the crowd can fact-check claims and the way they do it as compared to experts. The main contribution of this work is an in-depth data-driven study of how crowdsourced fact-checking can work in practice, as compared to expert fact-checking over a number of dimensions such as topics, sources of evidence, timeliness, and effectiveness.

Claim Selection. The first step in the process is deciding which claims, out of the very many produced on Twitter, should be fact-checked. This is similar to the process of assessing relevance in a search task, as users are looking for a piece of content that is valuable and that satisfies their requirements (e.g., an information need or potential harm caused by the piece of content if misleading). Regarding the selection of the claims to check, we show that the crowd mostly matches the claims selected by expert fact-checkers in terms of topics, and it is not strongly influenced by properties of the social network, such as popularity of tweets. Moreover, we analyze the responsiveness of crowd and experts with respect to fresh tweets and found that in some cases the crowd is orders of magnitude faster in generating a correct fact-checking outcome.

Evidence Retrieval. In terms of sources, Birdwatch users and fact-checkers rely on different set of online resources, with only few reference websites in common. For the sources used by the crowd and the experts, we also compare the quality perceived by the Birdwatch community against the quality ratings obtained from a professional journalistic tool. The two scoring methods show correlation, but also remarkable bias in source quality assessment by the crowd on some topics related to politics.

Claim Verification. In terms of effectiveness of the claim verification, we show that crowdsourcing may be an effective checking strategy in most settings, even comparable to the results obtained by human experts, but does not lead to consistent, actionable results for some topics. We also analyze the agreement among Birdwatch users and the use of different scoring functions to aggregate their feedback, including the one used in production by Twitter.

Our observations show how crowdsourcing fact-checking in practice can bring an added value as compared to expert fact-checkers or computational methods used in isolation. Additionally, we release the first dataset of tweets with labels from expert fact-checker, crowd, and computational methods.

In the rest of the paper, we discuss related work in fact-checking and crowdsourcing (Section 2), introduce the datasets collected and crafted for our study (Section 3), and discuss the empirical results for our analysis (Section 4). Finally, we discuss the main challenges and opportunities for crowdsourced fact-checking (Section 5) and conclude the paper with some open research questions (Section 6).

2 BACKGROUND AND RELATED WORK

Fact-checking requires a chain of steps that starts with identifying check-worthy claims and ends with a label about the veracity of the claim. Labels vary across services but usually can be divided into four popular categories: true, partially-true, false, or not enough evidence to judge. The top of Figure 1 shows a generic high-level fact-checking pipeline [41]. The three considered checking methods are then reported, specifically automatic methods, Birdwatch crowd, and expert fact-checkers. Given an input textual tweet, every method can be used to assess if it is worth checking and eventually verified. For every checking method, we also report the dimensions that can be used to compare and contrast the alternative methods. We discuss next the main steps in the pipeline, their related work, and their implementation in the different methods.

Claim Selection. For claim selection we can use automatic methods, the crowd, or experts. Given a sentence, an automatic method scores if it contains check-worthy factual claims [16, 28, 30]. A model trained on annotated sentences gives low scores to non-factual and subjective sentences. 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. In Information Retrieval evaluation, well-trained experts (e.g., NIST assessors) may be used to produce judgements of relevance following guidelines, or be instead substituted by crowd workers who receive simple instructions. Similarly, check-worthiness may be performed by a panel of experts or crowdsourced, like done in Birdwatch. The crowdsourced annotation of textual content on social networks is a widely supported 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 [10, 23]. For expert, human fact-checkers the selection of the claims to verify is driven by journalistic principles, e.g., claims should contain *verifiable* facts [9]. Experts also assess if a claim is *important*, with a definition that changes according to the public and the mission of the organization, e.g., voters and elections [1].

The crowd may have different criteria and priorities in deciding which claims to fact-check and a definition of check-worthiness that takes into account the topic, the timeliness, and their own personal points of view. Previous research in crowdsourced fact-checking (e.g., [44]) has not looked in detail at how the crowd may perform this step of the pipeline, and it is something that instead we do in this work.

Evidence Retrieval. For computational methods, we distinguish the task of detecting previously fact-checked claims and the task of gathering evidence to support the verification step. As false claims are often repeated across platforms and over time, independently of available fact-checks, claim matching aims at automatically identifying an existing debunking article for the claim at hand [11, 19, 50]. Claim matching is feasible at scale because websites use the schema.org standard ClaimReview metadata to share their checks [4]. For fresh claims, which have not been debunked yet, several methods aim at finding external evidence to help fact-checkers and computational methods deciding on the veracity of a claim [55]. The output is usually a ranking of retrieved documents or specific passages [14]. The crowd makes use of expert fact-checking outcomes when available. Indeed, Roitero et al. [48] removed expert outcomes from the search results used by the crowd in their fact-checking task to avoid influencing crowd worker judgments. Expert fact-checkers instead rely on their training to identify proven, verified, transparent, and accountable evidence [6], sometimes involving third-party domain experts [7].

Claim verification. A large body of research focus on developing and evaluating automatic solutions for fact-checking [19, 20, 34, 41, 45, 53, 56]. However, there are coverage and quality issues with automated systems [15], and thus a pragmatic approach is to build tools to facilitate human fact-checkers [56]. At the same time, effort in artificially creating rumors and misinformation has been shown to be effective [33]. The crowd makes use of evidence from the Web and is influenced by their own personal belief and context [17, 49]. Interestingly, when misinformation is identified on social media, users tend to counter it by providing evidence of it being misleading [40]. This shows an intrinsic motivation that certain members of the crowd have to contribute to the checking process. An approach for crowd-sourced fact-checking is using tools that surface relevant evidence for their judgement [25]. This however comes with the risk of over relying on such tools to make judgements [43].

Finally, there has been some early analysis of the Birdwatch data [13, 46], but they focus only on the tweets and notes content, while we rely on the manually aligned expert claim reviews to compare Birdwatch results against the best solution in this space. A related study has looked at a Reddit community involved in the fact-checking process using a crowdsourced approach [31].

3 DATA

Community-driven fact-checking on Twitter is governed by the Birdwatch initiative [8], while fact-checks written by journalists and expert fact-checkers are curated using the ClaimReview schema [4]. In this section, we describe both datasets and how to match similar claims identified by both parties. Approval from authors' institution research ethics committee to perform this study has been obtained prior to commencing.

3.1 Birdwatch

Misinformation on Twitter can be mitigated through the Birdwatch program, where participants can identify misleading tweets and provide more context [8]. Currently, Birdwatch is only available to participants in the US, where users can identify misleading information using two core elements: *Notes* and *Ratings*.

Notes. Participants in the Birdwatch program can add notes to any tweet. 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, (ii) answers to several multiple-choice questions about their decision [3], and (iii) an open text field where participants can justify their choice of the label and possibly include links to sources that prove their point. An example of a note is shown in Figure 2 (B,C). The key data we use from the notes are the following:

- *Classification Label*: Whether the tweet is misinformed (MM) or not (NM) according to the Birdwatch user (Section 4.3).
- *Note Text*: the text given by the user with the justification for the label (Sections 3.4 and 4.1).
- *Timestamps*: time at which the note was written (Section 4.1).

Ratings. Participants rate the notes of other participants. Ratings help identify which notes are most helpful. A user rates a note by providing answers to a list of questions [3]. An example of a rating is shown in Figure 2 (D). Out of these questions, we focus on the following:

- *High-quality Sources:* The user answers the yes/no question 'Is this note helpful because it cites high-quality sources?'. We use this information to assess whether Birdwatch users distinguish credible sources (Section 4.2.2).
- *Helpfulness Label:* The user answers the question 'Is this note helpful?'. The possible answers are (i) not helpful, (ii) somewhat helpful, and (iii) helpful. We use this information to compute an helpfulness score for notes (Section 4.3).

All Birdwatch notes start with a 'Needs More Rating' status until enough ratings are achieved according to a platform defined threshold (currently set to 5). Once achieved, these ratings are aggregated and weighted by a 'Rater Score' to compute the 'Note Helpfulness Score'. A higher rater score gives more weight to participants (i) whose notes are found helpful by other participants, and (ii) whose ratings align with the final rating outcome. A higher note helpfulness score means that many participants found a note adequate, and it would likely hold a valid classification label.

Descriptive Statistics. We use the Birdwatch data up to September 18^{th} 2021. The dataset contains 86,924 ratings for 15,445 notes on 11,871 tweets from 5124 unique Birdwatch participants. Bar plots of the number of notes and ratings are shown in Figure 3 and Figure 4, respectively. Most tweets have only one or two notes, while the tweet with the most notes has 61. The majority of notes have less than five ratings, and the most rated note has 184 ratings. The user with most notes checked 656 tweets with around 71% related to US Politics. Among these 656 tweets, 643 do not have any other note. The user with most notes in common with other users shares 85 notes (on 85 tweets) with 217 other users.

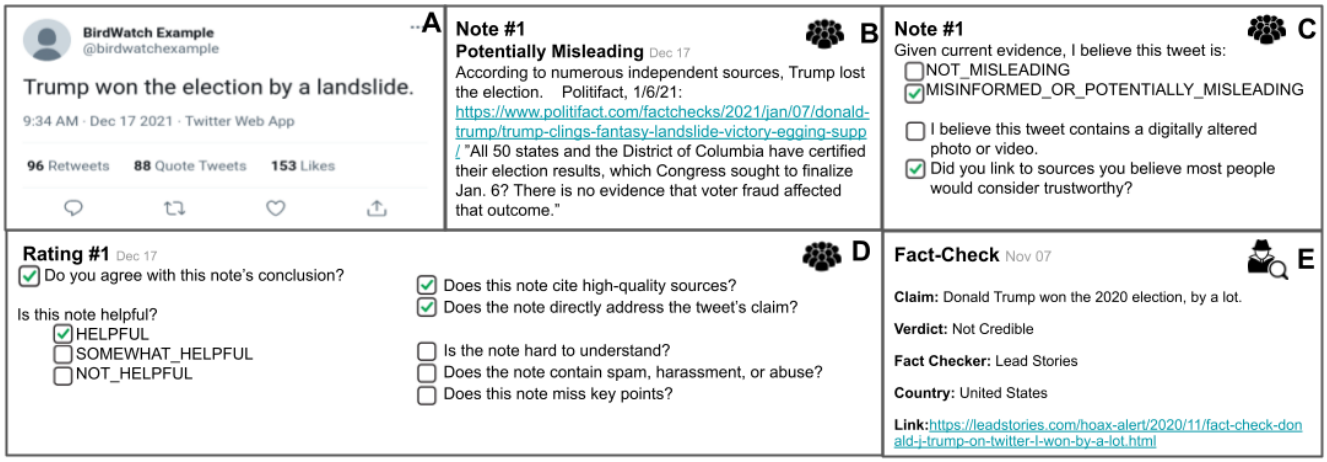

Figure 2: Birdwatch note and ClaimReview fact-check Example. (A) shows a tweet. (B) is the note with the assigned label to such tweet. (C) is a sample of questions when submitting a note. (D) is a sample of questions when submitting a rating. (E) shows a fact-check delivered by an expert.

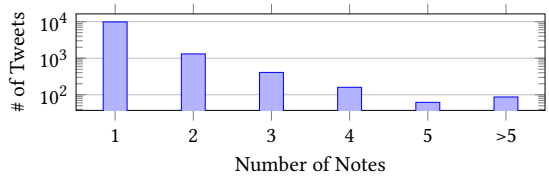

Figure 3: Bar plot of the number of notes per tweet.

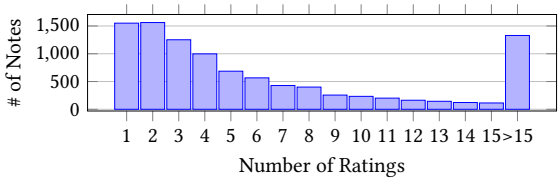

Figure 4: Bar plot of the number of ratings per note.

3.2 ClaimReview

The ClaimReview project [4] is a schema used to publish fact-checking articles by organizations and journalists. The schema defines mark-up tags that are used in web pages so that search engines identify the information in a debunking article, such as text claim, claim label, and author [5]. Our dataset is a collection of items following the ClaimReview schema, collected from various sources [39]. Each item, or *fact-check*, is a (claim, label) pair produced by a professional journalist or fact-checking agency. We assume that professional fact-checkers do not overlap with Birdwatch participants, as the former have no interest in doing their work without retribution. Since different fact-checkers use different labels, the data is normalized into a smaller subset of labels (credible, mostly credible, uncertain, unverifiable, not credible). In addition to the claim and the label, the checks also contain a link to the fact-checking article. Note that checked claims in this dataset could occur anywhere on the web and need not be only on Twitter. We use a dataset containing 76,769 fact-checks. Examples of the data are shown in Figure 2 (E) and Table 2.

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. As the automatic matching is imperfect, we used the Amazon Mechanical Turk crowdsourcing platform [2] for matching the text in the tweets checked by the Birdwatch crowd with the claim text in the ClaimReview fact-checks. When workers accepted a Human Intelligence Task (HIT), they were shown (i) the tweet that is to be matched and (ii) the top-10 similar ClaimReview checks provided by SentenceBERT using a bi-encoder with cosine similarity between the text of the tweet and that of the claim in ClaimReview fact-checks [47]. We also add a 'None of the Above' option for cases where the worker could not find a match. A manual inspection of the matches showed that the vast majority of tweets with a score below 0.6 do not have matching ClaimReview checks. We therefore run the annotation for tweets with at least 0.6 as top-1 similarity score. The workers were required to have at least 500 approved HITs to access our task, which comprised of 5322 tweets to be matched. Each tweet was shown to 3 workers, similar to previous work [35, 37]. The hourly rate based on median completion time was 12.41$.

To measure the quality of the worker annotations, we manually annotated the top-500 tweets in terms of matching score. Among these 500 tweets, we manually identified 75 with a matching ClaimReview check. Workers correctly matched 63/75 tweets according to our ground truth, while the baseline method choosing the highest SentenceBERT score correctly matched 59/75 tweets. After running the study over the 5322 tweets, we obtain 2208 tweets (3043 notes) matching with ClaimReview checks. An example of a tweet matching a ClaimReview check is shown in Figure 2 (A,E).

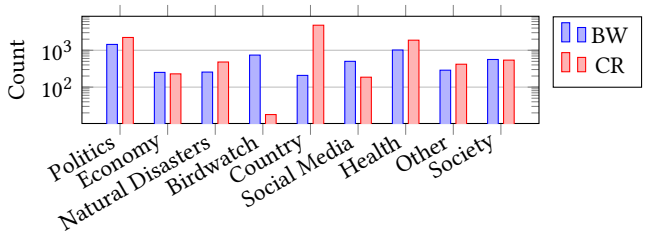

Figure 5: Bar plot of tweets checked by Birdwatch (BW) and ClaimReview (CR) fact-checks for 2021 divided by topic.

More examples of matched tweets, Birdwatch notes, and ClaimReview checks are in Table 2. Our dataset containing matched tweets to ClaimReview checks alongside labels from Birdwatch and ClaimReview and code relevant to the paper are available at https://github.com/MhmdSaiid/BirdWatch.

3.4 Topics

We analyze how Birdwatch notes and ClaimReview checks compare in terms of covered topics. We use *BertTopic*, a topic-modeling technique that utilizes transformers and TF-IDF for clustering [27], to predict the topic of every Birdwatch tweet and ClaimReview claim for the year 2021 and report their frequency distributions in Figure 5. *Politics* and *Health* have high counts in both. Topic *Country*, which includes news about countries all over the world, has higher counts for ClaimReview data since Birdwatch is deployed in the US only. Birdwatch notes cover mostly tweets in English and is biased towards US related news, whereas the ClaimReview data contains fact-checks in different languages and from local fact-checking agency, thus explaining the high number of country-related tweets.

4 RESULTS

We report results in addressing our three research questions next.

4.1 RQ1: Claim Selection

We analyze how Birdwatch participants effectively identify check-worthy claims in a comparison with fact-checking experts. We also compare Birdwatch users, who do not necessarily have journalistic training, against computational methods for this task.

4.1.1 Topic Analysis. After predicting the topic of every Birdwatch tweet and ClaimReview fact-check, we plot the frequency distribution of four topics showing interesting trends, on a monthly basis, in Figure 6. The high count of Birdwatch tweets and ClaimReview fact-checks covering political tweets show that they both consider the *Politics* topic important. The similar trends for this topic suggest that both methods react similarly to news and major events in terms of claim selection. For example, the peak in *Politics* for both methods in August is related to the Taliban take-over of Afghanistan. We observe the same trend for the topics *Economy* and *Natural Disasters*.

However, for the *Health* related tweets, we observe an abrupt change in the trends from July 2021. This is due to the emergence of the COVID-19 Delta variant in US, which triggered more tweets about the topic, mainly discussing masks/vaccines issues, and more Birdwatch notes on this topic. This is accompanied by a decrease in the number of health-related fact-checks, which can be explained by multiple reasons. One explanation is that the most important issues about masks and vaccines had already been debunked before the Delta variant. This shows that despite fact-checks are available online, numerous social network users keep spreading false claims that have previously been debunked (see also Section 4.3).

Topic selection also reflects the different geographic focus of the two methods. For example, the Birdwatch peak in February in topic *Economy* is due to the Texas power crisis, a US-specific event. Despite the differences, our results show that both Birdwatch participants and ClaimReview experts pick the content to verify in response to the events happening in reality, independently from the specific topic.

4.1.2 Computational Methods. We report on the ClaimBuster API for claim check-worthiness [30]. Given a sentence, the API provides a score between 0.0 and 1.0, where a higher score indicates that the sentence contains check-worthy claims. We run the API on Birdwatch tweets and the claims in the ClaimReview fact-checks, with the associated box plots for the scores in Figure 7 (A). The results show a check-worthiness median score at around 0.4, for both sets of claims, while in ClaimBuster the suggested threshold for check-worthiness is 0.5 [30]. One explanation of the difficulties of computational methods for claim selection is the bias in the training data used to build them. Indeed, most available datasets for this task are of high-quality text, coming from articles or political speeches, while the text used on Twitter is usually much noisier, e.g., due to the use of slang.

4.1.3 Tweet Popularity. We check whether the claim selection process of Birdwatch users is affected by the popularity of a tweet. For every tweet, we retrieve the number of retweets and favorites and sum them to obtain a quantifiable popularity score. As expected, Figure 8 shows that popular tweets receive more activity than others from the Birdwatch community, i.e., have more notes and ratings. However, there are popular tweets with low Birdwatch activity and unpopular tweets with high number of notes and ratings.

4.1.4 Temporal Analysis. We analyze tweets (T), Birdwatch 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 Birdwatch note, then ClaimReview fact-check (TBC), (ii) Tweet then ClaimReview fact-check then Birdwatch note (TCB), and (iii) ClaimReview fact then Tweet then Birdwatch note (CTB).

TBC: There are 129/2208 tweets in our matched data for this case. In all tweets, Birdwatch users provide a response much faster than experts. On average, a Birdwatch provides a response 10X faster than an expert. These examples show how Birdwatch participants can fact-check claims with reliable sources without the need of ClaimReview fact-checks such as ID #4 in Table 2.

TCB: In our dataset, a ClaimReview rarely occurs after a tweet and before a Birdwatch. We observe faster responses from ClaimReview than Birdwatch users for 26/2208 tweets. Since the granularity of the ClaimReview is days while that of Birdwatch is

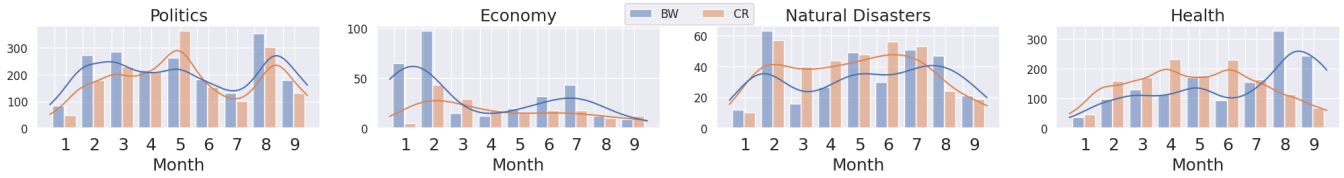

Figure 6: Per-topic frequency histograms and KDE Plots for Birdwatch (BW) notes and ClaimReview (CR) fact-checks (month granularity).

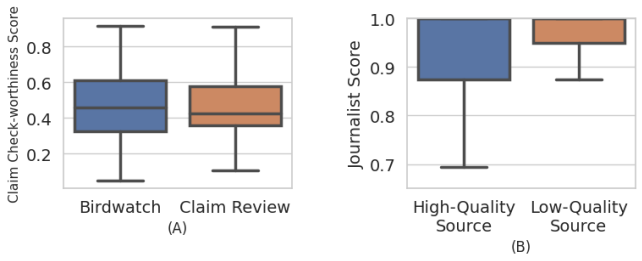

Figure 7: (A) shows a box plot of claim check-worthiness scores of Birdwatch tweets and the claims in the ClaimReview fact-checks. (B) shows a box plot of journalist scores compared to the final verdict of Birdwatch users (x-axis).

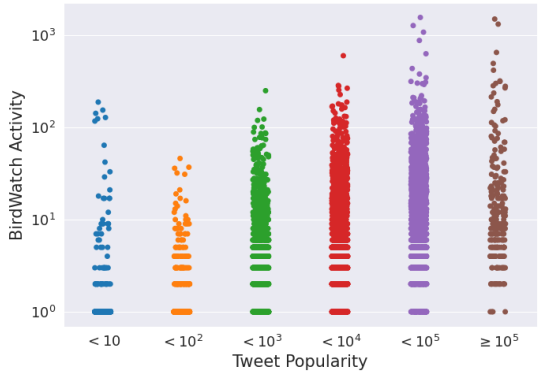

Figure 8: Tweet popularity and Birdwatch activity.

seconds, there are also 17/2208 tweets that occur on the same day, and we cannot state which of the two was actually faster.

CTB: The majority of the matched tweets follow this 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, specifically those related to Trump winning the elections. We discuss more this issue in Section 5.

Claim Selection Take-away Message: 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 off the shelf APIs.

4.2 RQ2: Evidence Retrieval

Both crowd checkers and experts report the sources used in their verification process. We analyze such sources and then contrast their quality according to an external journalistic tool.

4.2.1 Descriptive Statistics. We extract all links from Birdwatch notes. We find a total of 12,909 links covering 2,014 domains. Unsurprisingly, the top cited links are those coming from journalistic and fact-checking sites (Politifact, Reuters, NYtimes) and governmental websites, such as USGS and CDC. The distribution of the links is right-skewed, where half of the links are from only 29 domains. ClaimReview checks contain 76,769 links covering only 73 domains of fact-checking groups and journalists. The distribution of links shows less skewness than that of Birdwatch.

Birdwatch participants use only 17 domains in common to those of the ClaimReview experts. The other 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 Birdwatch focuses on US. Birdwatch sources are a larger number as they range from Wikipedia and YouTube videos to medical websites and research papers.

4.2.2 Expert Judgement of Source Quality. We compare ratings of source quality of Birdwatch users to those of expert fact-checkers. To assess the quality of web sources, we rely on an external tool that provides a score (between 0 and 100) where the higher the score, the higher the quality of the source[1]. The score is obtained by journalists manually reviewing every website, and we refer to it as the *journalist score*.

For every note in our matched dataset, we first compute a Birdwatch score indicating whether the links are high-quality sources or not by performing a majority voting on the ratings of the note, and we then compute the average journalist score of every link in the note. Out of 3043 notes, 2231 contained links. We obtained results for 656, while the others either had (i) no ratings (363/2231), (ii) no journalist scores (698/2231), (iii) nor both (309/2231), or (iv) there was no majority in the ranking votes (205/2231).

A box plot of journalist scores and Birdwatch labels is shown in Figure 7 (B). For note links rated as high-quality by Birdwatch users (with majority voting), we observe high journalist scores. The majority of tweets of the notes are related to US elections and COVID-19, with Birdwatch users citing sources such as Politifact

[1] https://www.newsguardtech.com

and CDC. Some sources in the notes have been classified as being high-quality by Birdwatch users but low-quality w.r.t. the journalist scores. Those notes share mainly COVID-19 studies such as Mayoclinic.org, a nonprofit American medical center, and fda.gov, the US food and drug administration, that are regarded as reliable sources in the US but do not meet all the requirements for high journalist score.

238/656 notes contain sources that are rated as low-quality, but have a high journalist score. These notes are debunking news about US politics, specifically about Trump winning the 2020 elections and misinformed COVID-19 content. These tweets include links to reliable sources, but a significant fraction of Birdwatch users labeled such links as low-quality. This shows how some Birdwatch users convey partisanship, forming a group of people trying to deceive the Birdwatch program to serve their common interest, such as supporting a political party in social media.

Such groups can be effective in "gaming" the algorithm [24], ultimately having a profound effect on Birdwatch since (i) the biased Birdwatch participants can steer the ultimate label of a note to their favor, thus spreading misinformation, and (ii) by increasing their weight in the Birdwatch platform since if one's ratings match those of the ultimate rating, they will get a higher weight. As an example, for the tweet `'Joe Biden is President In Name Only. #PINO'`, a certain note replied that Biden is indeed the president with links from *Politifact* and *APNews*, both having journalistic scores of 100/100 and 95/100 respectively while 12/14 of the raters identified such sources as unreliable.

We also compute journalist scores for links in Birdwatch and ClaimReview data. As Birdwatch users use many links, we only computed scores for the top-100 occurring links that form 68.6% of the data. While both distributions of Birdwatch and ClaimReview link scores attain a median of 1.0, links by ClaimReview fact-checks have lower variance with a minimum of 0.875, while that of Birdwatch notes is 0.495.

Evidence Retrieval Take-away Message: 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.

4.3 RQ3: Claim Verification

We ponder whether Birdwatch 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 report results for computational methods.

4.3.1 Internal Agreement. We use the participants' classification labels to see whether the tweet is classified as misinformed or not. To compute agreement, we use the standard metrics Krippendorff's alpha [36] and Fleiss's kappa [26]. However, due to the large sparsity in the data and the huge number of missing values, both metrics fail to provide meaningful results [21]. We then compute the variance as a metric for agreement. Lower variance means that all Birdwatch participants agree on the classification label.

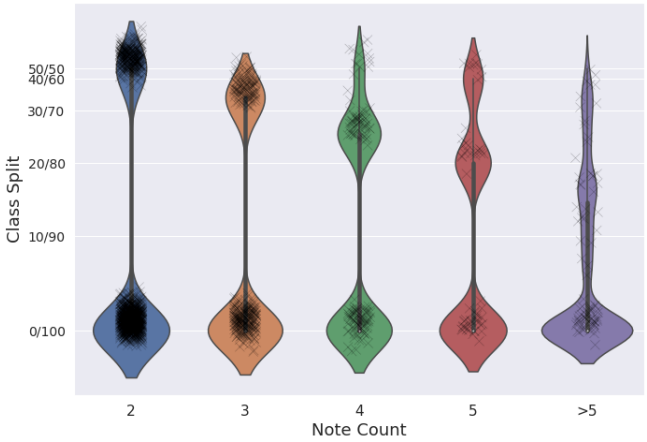

Figure 9: Violin Plot of note counts and class splits of the classification labels of notes. The figure shows the Kernel Density Estimation (KDE) plot of every note count, with their respective data points. Tweets with two notes are dominant, with most notes agreeing on the label. Other note counts do show full agreement on the label (0/100), with more cases of disagreement as the number of notes increase.

		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 Birdwatch and ClaimReview on the note level and the tweet level (obtained through majority voting). Agreement in bold.

A violin plot is shown in Figure 9 for tweets with various note counts. On the y-axis, we report the density of the class splits, where a class 0/100 indicates full agreement across the users and 50/50 indicates full disagreement. We see that most tweets have two notes and the majority of users perfectly agree on the final classification label. The same applies to tweets with more note counts, where most of the notes agree on the final label, with conflicts happening on some tweets but with a small subset with full disagreement. A topic analysis of tweets shows that 48.3% of tweets with full disagreement are related to either politics or COVID-19.

4.3.2 External Agreement. After matching Birdwatch data with ClaimReview fact-checks, we compare their labels. Table 1 shows that the majority of ClaimReview labels match the Birdwatch ones. Specifically, in terms of notes, there are 2022 cases (25+14+1983) where they agree and 449 cases of disagreement. In terms of tweets, there are 1492 (9+7+1476) decisions with the same classification

ID	Tweet	BW Note		CR Fact		Comment
		Text	Label	Text	Label	
#1	Pregnant women, please don't take this vaccine. https://t.co/4KKlnMIbl7	Updated CDC guidance, and newly accepted and reviewed medical research, has stated there are no safety concerns for pregnant women to be vaccinated against COVID-19. *(links omitted for brevity)*	MM	The vaccine is not safe for pregnant women or women planning on becoming pregnant within a few months of taking the vaccine... We are the lab rats.	NC	The Birdwatch user provides proof of why the claim is False. The fetched fact-check has a label of being not credible.
#2	The mass shooting at Marjory Stoneman Douglas High School in Parkland, Florida was real and not staged.	That this continues to be debated is astounding. Yes, this really happened. Here is a link: https://en.m.wikipedia.org/wiki/Stoneman_Douglas_High_School_shooting	NM	Say David Hogg is a crisis actor.	NC	BW note confirms the tweet, thus the label was not misleading. The CR check states a claim opposite to the tweet and its label is not credible.
#3	Chicago PD Says Enhanced Vid Shows Gun in 13-Year-Old Adam Toledo's Hand https://t.co/B0Twu733RL	Chicago Mayor Lori Lightfoot said Adam Toledo had a gun in his hand when he was fatally shot by a police officer, or words to that effect.	MM	Adam Toledo did not have the gun in his hand when he was approached by the police who shot him. He has his arms up and complied. The gun was on the floor, not his hands.	CR	Difference in granularity of the claim. For CR check, the claim is whether Toledo was holding a gun; while for the BW note, the claim was whether Chicago's mayor *said* that Toledo hold a gun.
#4	New poll indicates Biden approval at 11%. The LOWEST approval rating of ANY president in American history. Gallup via Daily Caller	https://fivethirtyeight.com/features/how-were-tracking-joe-bidens-approval-rating/	MM	Biden approval at 11%	NC	The Birdwatch participant provided a reliable source (score of 92.5) 7 days before a fact-check by an expert was available.
#5	Biden thinks he came to the US Senate 120 years ago?!?	US President Joe Biden made a clear joke at his first press briefing since his inauguration, in which he said he went to the Senate 120 years ago. This is a self-deprecating joke and shouldn't be taken seriously.	MM	Joe Biden said, 'With regard to the filibuster, I believe we should go back to the position of the filibuster that existed just when I came to the United States Senate 120 years ago.'	CR	The Tweeter took a joke seriously, which was interpreted as misleading by the Birdwatch participant.

Table 2: Examples of Tweets, Birdwatch Notes, and matched ClaimReview fact-checks. Birdwatch uses labels misleading/potentially misinformed (MM) or not misleading (NM), while ClaimReview uses credible (CR) or not credible (NC).

label and 232 (126+44+62) with different labels. For 69 (9+5+55) tweets there is a tie in the voting across Birdwatch users. For completeness, we report also the numbers for other ClaimReview labels ('not_verifiable' and 'uncertain'), even if they have no mapping to Birdwatch classification labels. We did some analysis to understand the cases where the labels are not aligned, some examples are reported in Table 2. Among the 209 notes that are labeled as credible by the ClaimReview fact-checks and misinformed by the Birdwatch 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 (ID #3). In other cases, tweets are mistakenly labelled as misinformed, e.g., because a joke is taken seriously by a Twitter user (ID #5). Finally, assuming correct ClaimReview labels, we believe in some cases the mismatch is due to biased Birdwatch users. For the tweets labeled as not credible by ClaimReview fact-checks and not misleading by Birdwatch notes, we observe cases where a Birdwatch note is the negated version of the ClaimReview fact-check (ID #2), thus producing opposite labels. There are also mismatch of labels, even though the Birdwatch user provides evidence from a link that has a high journalistic score (0.875).

4.3.3 Note Helpfulness Score. In the real-world production setting, not all Birdwatch notes are used for finding the ultimate label that gets exposed on the platform. In fact, a *note helpfulness score* is computed by the platform for each note, and those having a high enough score are used for computing the ultimate label. Birdwatch exposes the code for computing such score, however, the public code does not include raters' scores into consideration. We use the available code and filter out notes that are not helpful for the final label, using a Twitter-defined threshold for the note helpfulness score (0.84). We are left with 533 tweets (over 2208) that pass the threshold, with 333 notes labeling the tweets according to ClaimReview checks. About 95% of notes label the tweets as misleading, thus indicating that Birdwatch users tend to rate misleading tweets more than non-misleading ones, in agreement with previous work [46]. Of course, malicious ratings of the classification labels can steer the note helpfulness score in misleading directions, similarly to the judgement for source quality as discussed in Section 4.2.2.

4.3.4 Computational Methods. We compare our matched data with labels coming from computational fact-checking systems. We use again ClaimBuster, as it can also verify claims [32], and E-BART [20]. ClaimBuster provides correct results for 118 out of 2208 tweets, where 2090 tweets have no output from the model with an F1-score of 0.042. E-BART correctly labels 369 (over 2208) and does not produce a decision for 59 tweets with an F1-score of 0.17. A random classifier produces an average F1-score of 0.333 with 0.008

standard deviation. As for claim selection, tweets are harder to handle for computational methods than news articles and quotes from politicians, which are the bulk of content in training corpora.

Claim Verification Take-away Message: Birdwatch users show high enough levels of agreement to reach decisions in the vast majority of cases. The Birdwatch crowd focuses mostly on misleading tweets and shows high agreement with expert fact-checkers in terms of classification label. Computational methods have room for improvement in automatically verifying tweets.

5 DISCUSSION

5.1 Collaborative solutions

The analysis of the quality of the Birdwatch users shows that crowdsourced fact-checking is a promising and complementary solution, with results that correlate with those of professional fact-checkers. However, we argue that a crowd-based solution should not be considered to replace experts, but rather as a tool in collaborative effort where, for example, the crowd helps flagging content and creating links to more sources of trustable evidence. Indeed, our results show that the crowd can be even more reactive than experts to a new false claim and is able to identify a large array of high quality sources of evidence. This is especially important, as there is evidence that fact-checking interventions are significantly more effective in novel news situations [42].

Looking forward, and assuming we can characterize the trust and the cost level for all involved actors (i.e., crowd, experts, and computational methods), there is an opportunity to design novel hybrid human-machine solutions that coordinate this joint effort in order to combine the benefits of the different approaches. The role of automatic tools can be that of providing real-time and scalable fact-checking for all posted content. Platform users can then intervene quickly in a more focused manner to provide a first line of defense on potentially harmful content. This can then be followed by quality in every step of the fact-checking pipeline, with humans collectively processing evidence for the final labeling.

5.2 Hard to verify claims

The matching process of tweets and claim-review checks led us to recognize the difficulty of this task. The first challenge is the semantic match in terms of content, but in many cases where the match is clear, the problem is hard even for humans. Several problems, such as sarcasm and vagueness, are known in general for the detection of worth-checking claims [16]. However, another problem is the granularity of the tweet. Even a very short tweet may contain two interleaved claims, such as "Mike said: the earth is flat" (see also, e.g., Tweet #3 in Table 2). Assume there are two claim reviews, one checks that Mike made a claim about the earth (labeling the matching tweet as true), and the other checks about the fact that the planet is not flat (labeling the tweet as false).

This suggests the challenge of being able to identify the textual claims where both crowdsourcing and computational fact-checking methods are most likely to fail short. This can be modeled as a new supervised classification task aiming at predicting when a claim cannot be verified effectively without experts. As an orthogonal approach, this is also an opportunity for automatic controversy detection methods (e.g., [22]) to play a complementary role in supporting the crowd, making them aware of potential controversies during their verification tasks.

5.3 Stale claims

We found clear evidence that claims that have been already verified by expert checkers keep circulating and spreading on Twitter, even months after the publication of their debunking [51]. Unfortunately, we have also observed that automatic methods still fail short in matching with high accuracy tweets that contain such "stale" claims, likely because of the peculiar language used in tweets. In such a setting, Birdwatch users can play an important role in quickly and effectively recognizing these cases. Indeed, the significant difference in the health-related Birdwatch notes and ClaimReview fact-checks is explained by the increase of tweets spreading already fact-checked claims. Stale claims are a good fit for the role of Birdwatch participants, especially when automatic matching methods [12, 50] fail, while fresh claims might require proper forensic processes and need the expertise of journalists.

6 CONCLUSION

In this paper, we present a data-driven analysis of the Birdwatch program through the lens of the three main components of a fact-checking pipeline: claim detection based on check-worthiness, evidence retrieval, and claim verification. This is also the first study that bridges real data from a large-scale crowdsourced fact-checking initiative with the debunking articles produced by professional fact-checkers. While our analysis has been limited to Birdwatch participants from the US, we hope the Birdwatch initiative can be deployed globally for a more comprehensive analysis.

Our study shows that Birdwatch notes are effective in terms of claim verification, with encouraging results, in contrast with the negative results obtained by previous crowdsourcing efforts [18]. However, we also show that in Birdwatch a group of users sharing a common goal could potentially steer the final classification label, such as in the case of source credibility. This suggests that more attention is needed in identifying harmful groups by profiling their activity and by incorporating their biases in the note ranking system. Another approach would be to calibrate the selection of Birdwatch participants to enforce high diversity to mitigate this issue [24].

Our results indicate that the Birdwatch program is a viable initial approach towards crowd-based fact-checking, which can help complement the work of expert fact-checkers.An interesting open question is how to develop a collaborative platform involving the different fact-checking methods (crowd-based, computational, and experts). Given different trust and cost profiles for the different methods, a model to assess the difficulty of validating a claim (either data-driven or crowd-driven as in Birdwatch), and a certain budget, what is the optimal way to assign the claim to each fact-checking method so that the number of verified claims is maximized with a high level of trust?

Acknowledgments. This work is partially supported by an ARC Discovery Project (Grant No. DP190102141), by the ARC Training Centre for Information Resilience (Grant No. IC200100022), by gifts from Google and by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003).

REFERENCES

[1] [n.d.]. About the Fact Checker. https://www.washingtonpost.com/politics/2019/01/07/about-fact-checker.

[2] [n.d.]. Amazon Mechanical Turk. https://www.mturk.com.

[3] [n.d.]. BirdWatch Data. https://twitter.github.io/birdwatch/contributing/download-data/.

[4] [n.d.]. ClaimReview Project. https://www.claimreviewproject.com/the-facts-about-claimreview.

[5] [n.d.]. ClaimReview Schema. https://schema.org/ClaimReview.

[6] [n.d.]. Evaluating Evidence and Information Sources. https://kit.exposingtheinvisible.org/en/how/evaluate-evidence.html.

[7] [n.d.]. Full Fact Frequently asked questions. https://fullfact.org/about/frequently-asked-questions/.

[8] [n.d.]. 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.

[9] [n.d.]. The Principles of the Truth-O-Meter. ttps://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i.

[10] [n.d.]. Third party fact-checking. https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works.

[11] Ben Adler and Giacomo Boscaini-Gilroy. 2019. Real-time Claim Detection from News Articles and Retrieval of Semantically-Similar Factchecks. In *NewsIR'19 Workshop at SIGIR*.

[12] Naser Ahmadi, Hansjorg Sand, and Paolo Papotti. 2022. Unsupervised Matching of Data and Text. In *ICDE*. IEEE, 1058–1070. https://doi.org/10.1109/ICDE53745.2022.00084

[13] Jennifer N L Allen, Cameron Martel, and David G Rand. 2021. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. https://doi.org/10.31234/osf.io/57e3q

[14] Milad Alshomary, Nick Düsterhus, and Henning Wachsmuth. 2020. Extractive Snippet Generation for Arguments. In *SIGIR*. 1969–1972.

[15] Phoebe Arnold. 2020. *The challenges of online fact checking*. Technical Report. Full Fact.

[16] Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 1: Check-Worthiness. In *Working Notes of CLEF (CLEF '19)*.

[17] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In *ECIR (Lecture Notes in Computer Science, Vol. 12036)*. Springer, 207–214. https://doi.org/10.1007/978-3-030-45442-5_26

[18] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proc. ACM Hum. Comput. Interact.* 4, CSCW2 (2020), 93:1–93:26. https://doi.org/10.1145/3415164

[19] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. BRENDA: Browser Extension for Fake News Detection. In *SIGIR*. 2117–2120.

[20] Erik Brand, Kevin Roitero, Michael Soprano, and Gianluca Demartini. 2021. E-BART: Jointly Predicting and Explaining Truthfulness. In *TTO*. 18–27.

[21] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

[22] Shiri Dori-Hacohen, David Jensen, and James Allan. 2016. Controversy Detection in Wikipedia Using Collective Classification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 797–800. https://doi.org/10.1145/2911451.2914745

[23] 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*. Association for Computing Machinery, New York, NY, USA, 2627–2628. https://doi.org/10.1145/3404835.3464926

[24] Ziv Epstein, Gordon Pennycook, and David G. Rand. 2020. Will the Crowd Game the Algorithm?: Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *CHI*. ACM, 1–11. https://doi.org/10.1145/3313831.3376232

[25] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating Fact Checking Briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7147–7161.

[26] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (1971), 378–382.

[27] Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. https://doi.org/10.5281/zenodo.4381785

[28] Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. 2019. Neural Check-Worthiness Ranking with Weak Supervision: Finding Sentences for Fact-Checking. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 994–1000. https://doi.org/10.1145/3308560.3316736

[29] 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*.

[30] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *KDD*.

[31] Naeemul Hassan, Mohammad Yousuf, Md Mahfuzul Haque, Javier A. Suarez Rivas, and Md Khadimul Islam. 2019. Examining the Roles of Automation, Crowds and Professionals Towards Sustainable Fact-Checking. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1001–1006. https://doi.org/10.1145/3308560.3316734

[32] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: The First-ever End-to-end Fact-checking System. *Proc. VLDB Endow.* 10, 12 (2017), 1945–1948. https://doi.org/10.14778/3137765.3137815

[33] Larry Huynh, Thai Nguyen, Joshua Goh, Hyoungshick Kim, and Jin B Hong. 2021. ARGH! Automated Rumor Generation Hub. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3847–3856.

[34] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. 2020. Scrutinizer: A Mixed-Initiative Approach to Large-Scale, Data-Driven Claim Verification. *Proc. VLDB Endow.* 13, 11 (2020), 2508–2521.

[35] Gabriella Kazai. 2011. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval - Volume 6611* (Dublin, Ireland) *(ECIR 2011)*. Springer-Verlag, Berlin, Heidelberg, 165–176. https://doi.org/10.1007/978-3-642-20161-5_17

[36] Klaus Krippendorff. 2011. *Computing Krippendorff's Alpha-Reliability*. Technical Report.

[37] Di Liu, Randolph G. Bias, Matthew Lease, and Rebecca Kuipers. 2012. Crowdsourcing for usability testing. *Proceedings of the American Society for Information Science and Technology* 49, 1 (2012), 1–10. https://doi.org/10.1002/meet.14504901100 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/meet.14504901100

[38] Yang P. Liu and Yi fang Brook Wu. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. In *AAAI*.

[39] Martino Mensio and Harith Alani. 2019. MisinfoMe: Who is Interacting with Misinformation?. In *ISWC (CEUR Workshop Proceedings, Vol. 2456)*. CEUR-WS.org, 217–220.

[40] 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)*. 748–757. https://doi.org/10.1109/BigData50022.2020.9377956

[41] 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).

[42] Dorit Nevo and Benjamin D. Horne. 2022. How Topic Novelty Impacts the Effectiveness of News Veracity Interventions. *Commun. ACM* 65, 2 (jan 2022), 68–75. https://doi.org/10.1145/3460350

[43] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe It or Not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 189–199. https://doi.org/10.1145/3242587.3242666

[44] Marcos Rodrigues Pinto, Yuri Oliveira de Lima, Carlos Eduardo Barbosa, and Jano Moreira de Souza. 2019. Towards Fact-Checking through Crowdsourcing. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 494–499. https://doi.org/10.1109/CSCWD.2019.8791903

[45] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. *Vera: Prediction Techniques for Reducing Harmful Misinformation in Consumer Health Search*. Association for Computing Machinery, New York, NY, USA, 2066–2070. https://doi.org/10.1145/3404835.3463120

[46] Nicolas Pröllochs. 2021. Community-Based Fact-Checking on Twitter's Birdwatch Platform. *CoRR* abs/2104.07175 (2021). arXiv:2104.07175 https://arxiv.org/abs/2104.07175

[47] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[48] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. 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*. 439–448.

[49] 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* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1305–1314. https://doi.org/10.1145/3340531.3412048

[50] 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.

[51] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2017. Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *New Media & Society* 19, 8 (2017), 1214–1235. https://doi.org/10.1177/1461444816634054 arXiv:https://doi.org/10.1177/1461444816634054

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

[53] Ting Su, Craig Macdonald, and Iadh Ounis. 2019. Ensembles of Recurrent Networks for Classifying the Relationship of Fake News Titles. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 893–896. https://doi.org/10.1145/3331184.3331305

[54] James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3346–3359.

[55] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The Fact Extraction and VERification (FEVER) Shared Task. In *FEVER*. 1–9. https://doi.org/10.18653/v1/W18-5501

[56] Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking URL recommendation to combat fake news. In *SIGIR*. 275–284.

[57] Penghui Wei, Nan Xu, and Wenji Mao. 2019. Modeling Conversation Structure and Temporal Dynamics for Jointly Predicting Rumor Stance and Veracity. In *EMNLP-IJCNLP*. ACL, 4787–4798. https://doi.org/10.18653/v1/D19-1485