Variety, Velocity, Veracity, and Viability: Evaluating the Contributions of Crowdsourced and Professional Fact-checking

Andy Zhao,[1] **Mor Naaman,** [2]

[1] Cornell University
[2] Cornell Tech
dandy@infosci.cornell.edu, mor.naaman@cornell.edu

Abstract

The tension between the increasing need for fact-checking and the limited capacity of fact-check providers inspired several crowdsourced approaches to address this challenge. However, little is known about how effectively crowdsourced fact-checking might perform, and there is no comprehensive framework to evaluate such fact-check providers. We fill this gap by proposing such a framework, using four dimensions (Variety, Velocity, Veracity, and Viability) to assess and compare the contributions of a crowdsourced fact-checking community and professional fact-checking sites. Our analysis shows the different focus these two types of sites have in terms of topic coverage (variety) and demonstrates that while crowdsourced fact-checkers are much faster than professionals (velocity) to answer new requests, these fact-checkers often build on the existing professional knowledge for repeated requests. In addition, our findings indicate that the accuracy of the crowdsourced community (veracity) parallels that of the professional sources; and that the crowdsourced fact-checks are perceived quite close to professionals in terms of objectivity, clarity, and persuasiveness (viability).

Introduction

Fact-checking is perceived as an important counter-strategy to rampant misinformation online, because it may reduce misconceptions in misleading stories, or help people correctly evaluate the veracity of claims (Young et al. 2018; Walter et al. 2020). However, fact-checkers, who are often professional journalists (Graves and Amazeen 2019), face a capacity issue dealing with the breadth of potential misinformation on the Internet (Micallef et al. 2022). One proposed way to address this challenge is crowdsourced fact-checking, where amateur fact-checkers participate in the process of evaluating veracity, and provide their feedback as fact-check (Pennycook and Rand 2019; Allen et al. 2020).

One can consider two models of crowdsourced fact-checking: one that recruits ordinary people as participants to make an aggregated judgment (i.e. "wisdom of the crowds") on the veracity of content (Pennycook and Rand 2019; La Barbera et al. 2020); and another that motivates dedicated individuals to create new knowledge contributions in the form of fact-checks (Priedhorsky et al. 2007). We focus on the latter model in this work. Crowdsourcing services can motivate people to collaborate and create useful knowledge (Sunstein 2006; Kittur and Kraut 2008). The most prominent example is of course Wikipedia, and early studies showed the coverage and the quality of this crowdsourced encyclopedia is comparable to Encyclopedia Britannica (Giles 2005; Sunstein 2006) but at the same time, introducing different biases (Greenstein and Zhu 2018).

Multiple efforts developed crowdsourced approaches for fact-checking. For example, Truthsquad was a 2010 crowdsourced fact-checking experiment, led by fact-checking community Newstrust, where the site examined controversial claims and invited evidence and edits from users (Florin 2010b; Pinto et al. 2019). A more recent crowdsourced fact-checking effort is Twitter's Birdwatch, which is a community-based system to mobilize users to write and rate notes for suspicious tweets (Coleman 2021). Related works suggested that Birdwatch is an effective fact-checking method though it has some shortcomings like low-consensus and partisan focuses (Pröllochs 2022; Allen, Martel, and Rand 2022).

Our ultimate goal in this work is to evaluate whether crowdsourced fact-checking is a promising approach to combat misinformation in the real world. To this end, a comparative study with traditional and professional services is useful because it may expose the advantages and disadvantages of different models. For example, to compare the quality of Wikipedia with traditional encyclopedia, scholars utilized user or expert reviews, as well as behavioral features like the number of edits, word count, user reputation and lexical cues (Giles 2005; Wilkinson and Huberman 2007; Hu et al. 2007; Javanmardi and Lopes 2010; Xu and Luo 2011). However, these metrics and evaluations do not transfer well to the domain of fact-checking. For example, the value of fact-check service is not captured solely by article quality. Nieminen and Sankari (2021) developed detailed criteria to evaluate the different aspects of fact-checking content, which mostly reflects on its veracity. A more comprehensive and quantitatively feasible evaluation framework for fact-checking is necessary to understand the value of such an approach relative to traditional approaches, and perhaps offer implications for improving the crowdsourced practice in fact-checking communities.

In this study, we develop and present a comprehensive framework to evaluate the effectiveness and contribution of crowdsourced fact-checking services. We present a framework, "4V", going beyond a single method and aiming to

measure the value of crowdsourced fact-checking compared to professional fact-checking sites more comprehensively and empirically. The framework's name refers to the four V's that comprise its dimensions:

- Variety, referring to the breadth of fact-check coverage.
- Velocity, representing the speed of fact-checks.
- Veracity, capturing the reliability and objectivity of fact-checks.
- Viability, measuring how persuasive and effective fact-checks are.

Our 4V evaluation framework is not normative: it does *not* consider which dimension is more important, nor does it necessarily judge better or worse performance in each dimension (e.g., variety). Instead, we aim to provide the analytical framework to create a more comprehensive understanding of the value of fact-checking service and their contributions *in relation to* traditional fact-checks.

In this paper, as a case study, we use our proposed framework to compare the Taiwanese crowdsourced fact-checking site Cofacts to two Taiwanese professional fact-checking sites. Cofacts is an online community where motivated participants voluntarily answered fact-checking requests from other users.

Our results (Variety) suggest that two types of fact-checking have distinctive focus: professional fact-checking is more likely to step into fields requiring expertise like medical issues, while the crowdsourced fact-checking service tends to cover politics or common misinformation in everyday life like fraud messages. The results also show (Velocity) that, as expected, crowdsourced fact-checking usually precedes professional fact-checking in responding to requests, though the professional contributions often provided ready-made answers for crowdsourced fact-checkers to use for recurring fact-check requests. We show that (Veracity) the aggregated claims from multiple crowdsourced answers from Cofacts are almost as reliable as professionals. Finally, our analysis (Viability) shows that crowdsourced fact-checks are almost on par with fact-check articles created by professionals regarding persuasiveness and objectivity measures. The crowdsourced article even has a small advantage in clarity.

Related Work

While different scholars may have different opinions about the effectiveness and benefits of fact-checking, several research contributions argued that fact-checking is useful to counter some negative implications of misinformation under different cultural contexts (Porter and Wood 2021). A number of studies addressed the evaluation of professional fact-checks quality. Lim (2018) chose to assess the consensus of fact-checking statement and topic coverage, and she suggested that fact-checkers actually have relatively low topic overlaps and their consensus rates also vary widely due to different conversion methods. Nieminen and Sankari (2021) designed a list of 24 detailed criteria for fact-check practices and manually examined 858 fact-checks from PolitiFact, concluding that PolitiFact is generally of high quality but have a problem of clearness as the complex proposition in fact-check claims may confuse users. We borrowed the ideas and the logic in the above literature into developing our own evaluation framework.

Indeed, professional fact-checking, limited by nature, faces the significant challenge of keeping pace with enormous misinformation online (Allen et al. 2020), leading scholars to turn to crowdsourced fact-checking as a solution. For example, Florin (2010a) suggested that the collaboration between professionals and amateurs could deliver reliable fact-checking results based on Truthsquad experience. But the credibility of crowdsourced fact-checking still remains a concern, and relevant studies have mixed results on this question. La Barbera et al. (2020) suggested that crowds in their experiment exhibit bias in fact-checking, though aggregated conclusion of crowds is close to experts'. Pennycook and Rand (2019) recruited participants from online platforms and asked them to rate news sources, and suggested that laypeople's judgments about news source quality are very effective if aggregated in a balanced manner, though not as good as professional fact-checkers. Godel et al. (2021) chose to recruit ordinary individuals and professional fact-checkers to evaluate popular news stories. They found that while ordinary users cannot reach the level of professional fact-checkers, machine learning models perform better at identifying false news if training only on labels from users with a high level of political knowledge. This result suggests that a selective sample of crowdsourced fact-checkers could be helpful in terms of identifying unreliable news.

As Geiger et al. (2011) suggested, crowdsourcing works can be classified by many characteristics, and contributor group composition and result integration strategy are two important dimensions. However, most studies about crowdsourced fact-checking used experimental settings and recruited ordinary people as participants to make an aggregated judgment (Pennycook and Rand 2019; La Barbera et al. 2020). Although it is a feasible way to assess the potential of crowds, their results are not the same as the contributions of crowdsourced fact-checking in reality, because different coordination methods may have different outcomes (Kittur and Kraut 2008). Or in other words, previous literature focused more on a crowdsourcing model that collects information from a crowd of ordinary individuals and aggregates their results as the final judgments, but this approach is only applicable when most people tend to have a correct answer better than a random guess, as Condorcet's jury theorem suggests, which is not always true (Sunstein 2006).

Therefore, our study here focuses on a different crowdsourcing strategy: motivating dedicated individuals to make meaningful contributions, similar to the model used by Wikipedia, as suggested by Godel et al. (2021). The active users in the crowdsourcing community are usually self-selected and have a wide array of motivations (Oreg and Nov 2008). The aggregation strategy used in the crowdsourced site we study follows neither the simple average of all opinions nor the Wikipedia model of collaborative editing. Instead, it is a model where multiple fact-checks can be contributed, then up- or down-voted. Users may read all answers and selectively accept fact-checks with more upvotes.

Hassan et al. (2019) examined such a fact-checking model on Reddit and suggested that comments from ordinary users did provide informative feedback. Amateur fact-checkers played different roles than professional journalism and coordinated with other users to produce effective answers for fact-checking requests. Hassan et al. (2019) suggested that such a crowdsourced fact-checking model, along with the help from professionals and automation, has strong potential in the future. Saeed et al. (2022) also compared the crowdsourced fact-checks from Birdwatch with experts on ClaimReview. Their study focused on topic selection, evidence sources, and accuracy as well, and suggested the distinctions and edges that crowdsourced fact-checking has.

Current literature in fact-checking overwhelmingly focused on the Western experience, or most specifically, the United States. While many Asia-Pacific countries have rich experience battling against misinformation, for example, establishing special governmental agencies or mobilizing civil societies, their practices are largely understudied (Davis, Crowley, and Corcoran 2019; Cha, Gao, and Li 2020). Taiwan offers a very specific context given its approach and challenges in protecting its information ecosystem. To address its challenges of misinformation and ideology clashes, the Taiwanese government collaborated with civil organizations and tech communities to develop digital tools (Cha, Gao, and Li 2020; Chang, Haider, and Ferrara 2021). Our study object in this paper, Cofacts, is one of the most prominent outcomes of government-civil collaboration efforts.

Method

To compare the contributions of crowdsourced and professional fact-checking, we first collected fact-checks from crowdsourcing and professional sites. We then matched fact-checks about the same requests from both sources to generate comparable pairs for our further analysis. This section details our data and matching strategy, as well as a rating task, performed to understand article perceptions.

Data

Our context of this study is Taiwan, specifically using a popular crowdsourced fact-checking community called Cofacts. We obtained user-generated request and answer data from Cofacts as our dataset of crowdsourced fact-checking. To construct an equivalent dataset of professional fact-checking, we collected fact-check articles from two popular professional fact-checking sites MyGoPen and Taiwan FactCheck Center. All text data are in Chinese.

Cofacts *Cofacts* is an online fact-checking community founded in 2017. It originated from the Taiwanese decentralized civic tech community *g0v*. In this Quara-like community, all users can either make a request to fact-check a suspicious claim or answer any such requests. Besides, users can also second a request, and upvote or downvote an answer.

Cofacts made all data publicly available on GitHub upon request. When we made our data request at the end of July 2021, this site has more than 60,000 fact-checking requests and more than 55,000 fact-checking answers. Except for the early days and the most recent days, their fact-checks maintained a consistent and relatively smooth trend as requested most of the time.

According to our data, 1,823 unique Cofacts users ever answered fact-checking requests. 721 of them answered more than once and 411 of them answered more than twice. Around 6% of all users produced 94% of all replies, and 1% users are responsible for about 90% fact-checks. Cofacts also enables fact-checkers to use different labels to indicate their conclusions, 44.9% fact-checking replies used the label "contains misinformation" and 22.2% used "contains correct information". 19.8% replies were associated with "not fact-checkable" and 13.0% replies were "opinionated".

Taiwanese professional fact-checking sites We chose two popular professional fact-checking sites in Taiwan as our sample of professional fact-checkers: *MyGoPen* and *Taiwan FactCheck Center*. These sites publish fact-checking articles on their websites as most fact-checkers do. We collected all public fact-checking articles from the creation of these two sites, along with necessary features like dates and labels. We crawled 1,687 articles from MyGoPen and 944 articles from Taiwan FactCheck Center by looping their web pages. The dates of these articles ranged from November 2015 to May 2021.

Matching fact-checking requests and answers

Kazemi et al. (2021a) suggested that claim matching between fact-checks of multiple languages is a significant challenge to scaling up global fact-checking. This is also a fundamental challenge for this research, as we need to find comparable pairs between the crowdsourced fact-checks and professional fact-checks on the same topics. This is a necessary step to evaluate the variety, velocity, and veracity in this study.

To achieve this goal, the Cofacts request was seen as a "headline" of fact-check and the Cofacts responses to this request was seen as the articles under this "headline". We then calculated the text similarities between Cofacts requests with the title and summary of professional articles on fact-checking web pages, which returned another article under the same "headline". [1] By doing this, we can find pairs of crowdsourced and professional fact-checks that were under the same "headline", or responded to the same content. For instance, a Cofacts question about a hacking virus on a Christmas greeting picture should be paired with Cofacts responses under this request and professional fact-checks about this specific topic, rather than on Covid virus during Christmas or hacking virus carried by an email.

To capture the subtle distinctions between seemingly identical Chinese text, we chose Jaro-Winkler (JW) similarity, which measures the edit distance between two strings and is ideal for Chinese characters. To test the validity of

[1] Different users may check the claim in different ways with distinctive languages and simply cite a webpage in Cofacts responses, but a relevant fact-check probably cannot avoid the exact terms in the original misinformation. Therefore, we used crowdsourced requests rather than answers for our calculation to retrieve similar professional articles.

JW similarity, we sampled 200 matched article pairs from our dataset (100 positive cases and 100 negative cases with above 0.7 and below 0.6 JW similarity scores respectively). We asked two native Chinese speakers to annotate the homogeneity in the meaning of two articles (eg. talking about the exact same issue). The result indicates that JW has 2 false positives and 37 false negative cases, which means that it has a precision of 0.98 and a recall of 0.73. Since we need to aim for accurate matches and weigh less on recall in this research, the JW algorithm is a suitable tool for us to distinguish the nuanced differences among Chinese text and find identical fact-checks. We also informally evaluated BERT and realized that distinguishing similar Chinese text is a weakness of this language model.

However, the similarity threshold and the time frame to retrieve fact-check candidates may both have great implications on the matching results. Therefore, we tested JW algorithm with different thresholds (0.6 and 0.7) and several time differences (7, 15, 30, 45, and 60) based on our observations and experiences. The results are relatively close, so we chose 0.6 as the similarity threshold and 45-day difference as our matching window to retrieve more fact-checks. Therefore, we identified 1,222 unique professional fact-checks and 1,496 unique crowdsourcing fact-checks on similar issues and posted within 1.5 months from each other (one professional article may match with multiple crowdsourcing fact-checks). These matched fact-checks are made up of 46.4% of all professional articles and 2.4% of all crowdsourced fact-checks. Since we only used matched pairs for a part of our evaluation, our conclusions are insensitive to the match rates and these pairs are still helpful for us to understand the performances of different sources. For example, to evaluate Velocity we evaluate which side is faster to publish a headline from the matched pairs.

Evaluation based on the 4V Framework

Having set up the data sources and some of the data collection details, we now turn to present an evaluation of the crowdsourced to professional fact-checking based on the four dimensions of our framework. For each, we motivate and provide details of how we operationalized the comparison, and then present the outcome of the evaluation.

Variety

Our first dimension, "Variety", represents the topic coverage of fact-checking articles. Our measures for the variety dimension aim to show the differences between the topics covered by the two types. Due to the different volumes of production, we focus on the proportions of topic coverages rather than the topic counts, which are less sensitive to the topic "resolution".

We therefore investigate the variety, operationalized as the diversity of topics, as a measurement dimension. We do this in three different ways to increase the robustness of our findings: first, using user-generated topic labels from crowdsourced fact-checks and supervised learning to predict the topics of professional fact-checks; second, using BERT embedding and unsupervised learning to cluster professional articles and then predict the topics of crowdsourced fact-checks; finally, a "match and assign" strategy with user-generated labels.

Topic Classification For the first approach to understanding the topic coverage of fact-checks from both sources, we built on the topic labels that Cofacts user-generated to describe their posts. In other words, we took advantage of user-curated labels, assigned to fact-check requests by Cofacts community members. Since the professional fact-checks do not have associated topic labels, we used supervised learning to assign them with labels. To this end, we used crowdsourced fact-checks with topic labels as a training dataset to develop a classification model to infer the topics of professional articles. While this one-sided approach will not allow us to study topics that are exclusively covered by professional sites, we do not believe such topics are prevalent given our observations and the wide interests of users.

To have better classification accuracy, we filtered out topics that have a relatively small amount of cases or with an "unclassified" label. We then balanced our training dataset by oversampling smaller topics and undersampling bigger topics. This process resulted in 23,569 fact-checks with seven topic labels. We used 20,000 random fact-checks as our training sample and the remaining data as our evaluation dataset. We embedded fact-check data with BERT and then used a neural network to train our model. The evaluation accuracy on the set of 3,569 article was 0.899 (0.898 precision; 0.903 recall).

The classification results of our model on professional fact-checks are shown in Figure 1, along with the topic distributions of crowdsourced fact-checks. The blue bars represent the number of professional fact-checks in each topic (left y-axis), and the orange bars represent the number of crowdsourced fact-checks (right y-axis). The orange striped blue bars for professional fact-checks indicate the number of professional articles that are referenced by crowdsourced fact-checks, and the blue striped orange bars for crowdsourced fact-checks indicate the number of crowdsourced responses that refer to professional fact-checks. A Chi-squared test comparing the article distributions over these topic labels showed no significant differences between the topic distributions of professional fact-checks and crowdsourced fact-checks. However, the high proportion of crowdsourced fact-checks that refer to professional articles to answer a question about fraud messages also indicates that such answers are largely supported by valuable works from professionals. This phenomenon suggests that while professional fact-checkers did occasionally respond to some requests about fraud messages, recurring needs in this field were usually fulfilled by crowdsourced fact-checkers who helped with the further distribution of valuable fact-checks.

In addition, the high referenced proportion of professional fact-checks in "Health and Food Safety" and "China" indicates the diverse needs of the mass that cannot be satisfied solely by crowdsourced fact-checkers. While not too many crowdsourced fact-checks choose to refer to a professional article in these topics, the contrast between the reference ratios in the two types of fact-checks suggests that the reliance

on crowdsourced fact-checkers on professional sites to answer some less common but more broad requests.

Topic Clustering In the second analysis of the variety of fact-checks, we attempted to use topic clustering to evaluate the fact-check topic distributions from both professional and crowdsourced fact-checkings. We used a BERT model to compute the embedding features of the text of professional (titles and summaries) and crowdsourced (response bodies) fact-checks (Devlin et al. 2018). Then, opposite to the previous step, we trained a model on professional fact-check data with a K-Means clustering algorithm. Attempting different values of k, $k = 5$ gave the highest silhouette score and the best within-cluster cohesion. We then use this unsupervised model to predict the cluster labels of all crowdsourced fact-checks. For presentation, we summarized the fact-check topics by manually examining the contents in each article cluster.

Figure 2 shows the topic distributions of both professional and crowdsourced fact-checks over five topic clusters (a CHI-squared analysis suggests that these two distributions are significantly different, $p<.01$). These topics are summarized after we manually examined the fact-checks in clusters, so it has distinctive topics than Figure 1 (for example, "Health and Food safety" is separated into two different clusters in Figure 2: "Food Security" and "Health and Lifestyle"). The figure suggests that crowdsourced fact-checkers post more on social fraud information: based on our evaluation, these are topics such as anecdotes, store discounts, missing kids, etc. The topic of politics and public policy is another domain where crowdsourced fact-checkers have wrote proportionally more answers. Professional fact-checkers examined more topics like social or international information (for instance, rumors about a Japanese aquarium, counterfeit money, NASA's alien encounters, etc.), food security, and health and lifestyle.

Our supervised and unsupervised learning agree that crowdsourced fact-checking tends to write more on social fraud messages or policies that are relevant to daily information. However, our two methods show different results regarding which source may weigh comparatively more on the topic of health, lifestyle, and food safety, while our findings all confirm that this topic is very popular.

Match with user-generated labels Our third technique to examine topic distributions of fact-checks still utilizes user-generated topic labels on Cofacts but overcomes the lacking of labels in the professional dataset by computationally matching professional articles with crowdsourced requests as explained above in *Method* Section. We assigned the topic labels in Cofacts requests to corresponding crowdsourcing answers *and* the matched professional articles. This analysis aims to understand, for all crowdsourced topic labels, which one also has specific stories covered (more or less) by professionals.

After deduplication at the article-level to exclude recurring requests and repeat answers, we calculated the topic distributions of crowdsourced and professional fact-checking. Table 1 shows the results in the topics with at least 150 fact-checks and suggests that both types have checked many suspicious stories on issues like health and food security, China, and regulations.

Given the fact that professional fact-checking articles can match 2.4% of the crowdsourcing dataset with a 0.6 JW threshold, we treat the topic with higher matching rates as the domain where professionals would give more weight and vice versa. Or in other words, professionals would check proportionally more stories in some topics compared to crowdsourced fact-checkers, and these topics are more likely to have a higher match rate than the expected match rate. Table 1 suggests that professional fact-checkers tend to focus more on COVID-19, technologies and privacy, environment protections, and other medical issues, and crowdsourced fact-checking are more likely to write articles in response to requests on the topics about fraud messages, energies, and political parties. This result is also consistent with our observations in Figure 2.

Overall, three distinct analysis we performed to evaluate the variety of crowdsourced and professional fact-checking show that professional fact-checkers tend to examine the information that requires some knowledge or have bigger implication, for example, medical or health news and international affairs. On the other hand, crowdsourced fact-checkers are proportionally more likely to focus on recurring fraud messages or local political news.

Velocity

Our next dimension, "Velocity", represents the response speed of fact-checking articles. Since the speed of response of fact-checks matters (Brashier et al. 2021), a faster reaction to potential misinformation could be highly valuable.

To compare the response speed of the different services, we again took the Cofacts data as a baseline, using the *requests* for fact-checks as "time zero" for global fact-check needs. On Cofacts, we can take the time difference between the first response to the request and the original request time as its response time. This is because early fact-check responses have the advantage to catch on the spread of misinformation. For the professional fact-checking sites, we identified the articles that match the Cofacts request, and use the time difference between the first Cofacts request and the corresponding professional article as its response time.

However, relying on Cofacts requests is only an approximation of real-world demand for new fact-checks. Because professional fact-check articles could have already existed but were unbeknown to users of Cofacts (if there was a similar request on Cofacts, users may find it with auto-searching during the reporting or notice it in the "similar suspicious message" section). Luckily, the crowdsourcing community itself helps us address this challenge.

We divided our matched fact-checks pairs into two parts: professional articles that existed before the request and those after the request. In the first part, crowdsourced contributors took advantage of fruits planted by professional fact-checkers by responding to requests with a citation to a professional fact-check article. In fact, 454 out of 897 matched fact-check pairs (under 0.6 JW similarity threshold) on Cofacts who had existing answers chose to directly cite a link to MyGoPen or Taiwan FactCheck Center to answer these

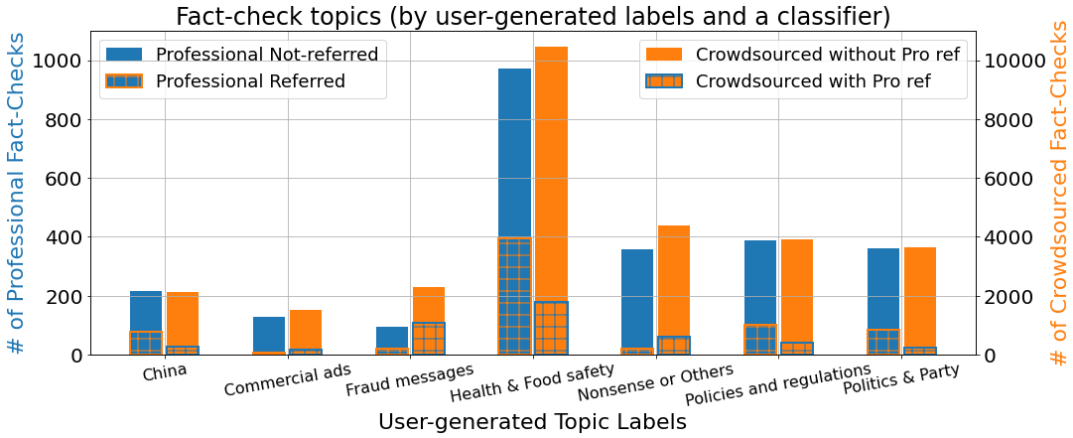

Figure 1: Topic distributions of professional and crowdsourced articles. The orange bars with blue grids represent the crowdsourced fact-checks that refer to professional articles, and the blue bars with orange grids represent the professional fact-checks that were referred by crowdsourcing sites.

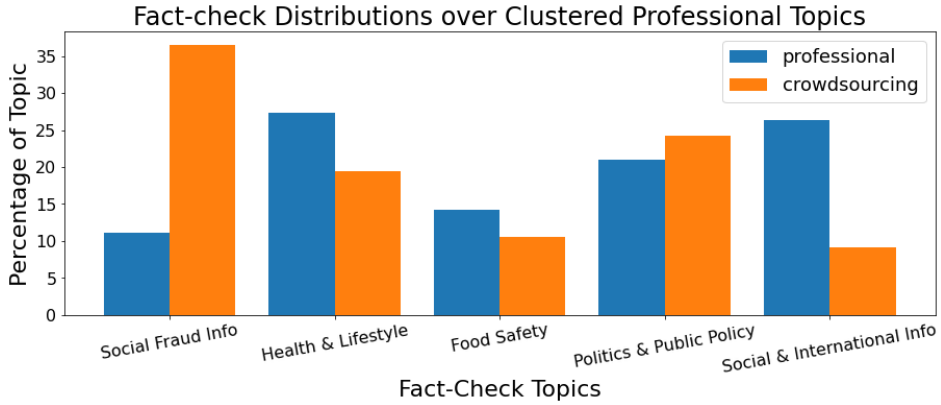

Figure 2: Topic distributions of professional and crowdsourced fact-checking over five topic clusters. The blue bars represent the number of professional fact-checks in each topic and the orange bars represent the number of crowdsourced fact-checks.

outdated requests. This result indicates that crowdsourced fact-checkers rely on their professional counterparts to respond to recurring requests to a large extent.

We now turn to analyze the requests that did not have a ready answer on professional sites. We treat these requests as a "clean slate", assuming there was no previous fact-check on the topic but was also checked by professionals later. Even after excluding those requests in the first part (which is roughly half of all requests on Cofacts), our results indicate a clear advantage for crowdsourcing fact-checking in answering emerging demands. In general, Cofacts is faster in 754 cases out of 879 "clean slate" cases.

Figure 3 shows the day difference distributions between professional and crowdsourced fact-checkers, which also suggests crowdsourced fact-checkers outpaced professionals in "velocity" by a large extent. Table 2 demonstrates the contributions of crowdsourced and professional sites in rapid and slow cases, and crowdsourced fact-checks are earlier than their counterparts in both circumstances. ("Rapid cases" mean that at least one site responded to the fact-checking requests within 24 hours; "Slow cases" represent the situations in which both sites answered the requests after 24 hours; "Tie" suggest that the professionals and crowdsourcing sites answered the requests on the same day.) If we regard rapid cases as easy questions and slow cases as questions that require more effort to explore, our data also implies that professional fact-checkers may have a comparative advantage in requests that may take more time.

In addition, if there was a professional fact-check before the request, the median time for the crowdsourcing community to answer is seventeen minutes; if there was never a professional fact-check after all (even after the request), the median time for the crowdsourcing community to take is about twelve hours longer. However, it is hard to claim any causality here since it could be that professionals accelerated the fact-checking or they tend to avoid the tricky questions.

The velocity differences also vary among different topics. Table 3 shows the comparisons of matched pairs on different topics, where "pre-answered" means the requests that are already answered by professionals. It suggests that the advan-

Domain	# of Cofacts	# matched with Pro	match rate (%)
Covid-19	365	37	10.14
Technologies and privacy	327	21	6.42
Medical issues	81	5	6.17
LGBT and AIDS	286	14	4.90
Environment protection	364	17	4.68
Health and Food security	11,747	418	3.56
Agricultural policy	481	15	3.12
Policies and regulations	4,093	127	3.10
China	2,358	72	3.05
Fraud messages	2,431	72	2.96
Signing and donating	570	14	2.46
Gender issues	240	5	2.08
Politics and parties	4,308	77	1.79
Commercial ads	1,560	21	1.35
Electric and energy	186	2	1.08

Table 1: Topic distributions in Cofacts and matched professional articles

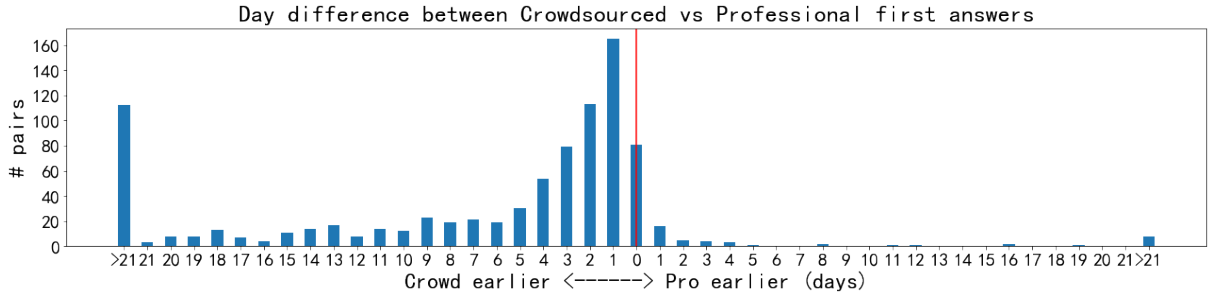

Figure 3: The reaction time differences between professionals and crowdsourcing. The left part of this bar plot represents the cases where crowdsourced fact-checks are faster and the right part indicates the situations where professionals are faster.

win	Crowdsourcing	Tie	Professional
Rapid cases	502	168	25
Slow cases	116	28	40

Table 2: Velocity comparison between crowdsourced and professional fact-checking

tage of crowdsourced fact-checking in velocity still holds in different domains, and recurring requests are more likely to happen for the topics like COVID-19 and fraud messages.

Veracity

Our third dimension is "Veracity", which represents the credibility of fact-checking articles. Fact-checking is expected to be as neutral as possible, but even one of the best crowdsourcing products Wikipedia is not as objective as Encyclopaedia Britannica (Greenstein and Zhu 2018). Therefore, a big concern remains over crowdsourcing fact-checking, naturally, how reliable such contribution might be compared to professional journalists and fact-checkers.

In this section, we now use the *professional* fact-checking as a baseline (or ground truth) for the veracity rating of an article. We measure whether the labels associated with fact-checks by crowdsourced contributions on Cofacts are aligned with the labels assigned to the fact-check article of the same issue by professionals. Only absolute and clear la-

Domain	Crowd	Tie	Pro	Pre-answered
Health and Food security	212	14	10	228
Policies and regulations	73	11	4	54
Fraud messages	25	7	3	45
Politics and parties	32	1	3	43
China	31	5	3	31
Covid-19	23	0	0	20
Commercial ads	14	1	1	11
Environment protection	7	0	2	9
Signing and donating	13	0	0	8
Technologies and privacy	12	1	1	7
Agricultural policy	10	1	0	3
LGBT and AIDS	13	0	0	2
Gender issues	1	0	0	2
Electric and energy	2	0	0	0

Table 3: Velocity of crowdsourced and professional fact-checking over topics

bels on professional fact-checks were taken (in a scale of True and False), which are all consistent among two professional sites and we turned into a numeric scale. For the crowdsourced fact-checking, we selected the majority opinion of all binary True/False labels under each Cofacts request, which is more natural from a user's perspective.

663 question-answer pairs from crowdsourced and professional fact-checkers were identified by the matching method explained above. Out of 643 unique Cofacts requests in matched question-answer pairs, roughly 61% of them received only one answer and 28% of them received two answers. Most of the time, fact-check requests would receive unanimous answers: 546 requests received all False labels and 69 requests got all True labels.

After our initial analysis of label alignment, we manually examined the cases where crowdsourced labels are not consistent with professional labels to validate our results. We found that some disagreements are some Cofacts requests asking to corroborate professional fact-checks, which we would like to refer to as "double-check" cases. In other words, professional fact-checkers debunked a rumor (giving a "False" label); and a Cofacts request asked to verify this fact-checking article; then the crowdsourcing users endorsed the conclusion of this professional article (giving a "True" label). These fact-checks were matched by our JW algorithm because there are identical texts in Cofacts requests and professional fact-checks.

Therefore, we excluded these "double-check" answers and the cases where a professional article existed before a fact-check request was posted on Cofacts, because crowdsourced fact-checkers can simply copy and paste answers from professional sites. The veracity trend remains unchanged in our final result, as shown in the confusion matrix in Table 4. There were only a few disagreements between crowdsourced fact-checking and professional fact-checking.

		Pro	
		True	False
Crowd	True	12	4
	False	0	305

Table 4: Confusion matrix between professional and crowdsourced fact-check labels after data filtering

Another serious concern about amateur contributions in the context of fact-checking is that while it could be trustworthy in most cases, it may be less reliable when evaluating critical issues. We identified four valid cases where professionals and Cofacts users really disagree with each other. The first case is about a church, while MyGoPen and Cofacts have similar answers and references, they put up different labels. The second case is about a haze weather warning. Both sides are correct about this issue, but the timeliness of their answers resulted in different labels. The third case is about the mask policy at the voting stations and two sides disagree on its mandatory. The last case is about a new medical technology, and Cofacts confirmed the existence of this technology on the market while MyGoPen confirmed with a hospital that they didn't have such technology.

In general, while amateur fact-checkers may occasionally disagree with professional fact-checkers, crowdsourcing fact-checking could be as trustworthy as professional fact-checkering most of the time.

Viability

Our final dimension is "Viability", which represents the likelihood of the contributions of fact-checks to be considered valuable by readersWe, therefore, evaluate how readers might perceive the fact checks provided by crowdsourced and professional sources. We used raters to rate a set of articles along multiple dimensions, as described in Section .

Annotating Perceived Quality Fact-checks are read and understood by humans. As part of our evaluation, we used raters to help us estimate the likelihood of the contributions of fact-checkers being perceived by readers along different measures of quality.

To understand the perceived quality of fact-checks, we randomly sampled hundreds of pairs of crowdsourced and professional fact-check articles on the exact same false stories, which were manually examined by our authors. We selected forty article pairs (one crowdsourced fact-check and one professional fact-check on the same topic) for this procedure such that the pairs represent balanced topics (15 for medical and health information, 15 for domestic stories and 10 for international stories), and both articles in each pair are original (not just a reference to another site).

We recruited seven native Taiwanese graduate students as raters to read and evaluate these fact-checks.Participants were randomly exposed to either a crowdsourced or professional fact-check of each of the forty pairs (without knowing its source). Only text and images in the fact-checks were presented to participants, because we want to avoid the influence of website designs and other factors on the perceptions of fact-checks. Participants were asked to read fact-checks carefully and annotate how they perceived the qualities of each article in three measures, on a scale from 1 to 5: persuasiveness, clarity, and objectivity. As each rater reads 40 articles, we obtained 280 responses for each fact-check pair and therefore 140 responses for each (crowdsourced or professional) fact-check article.

Measurement Results The first measure is "objectivity". With this measure, our intent is to understand whether fact-checks from both sources are perceived as objective or neutral from a reader's perspective. Our results show that raters ranked the professional source fact-check articles as somewhat more objective. The mean objective rating for professional fact-check articles was 4.16 (SD=0.94), compared to 3.79 (SD=1.16) for the crowdsourced fact-check articles. The difference between the two sets of articles was significant ($p<0.01$) though the effect size was relatively small (0.34).

Our second dimension, "clarity" aimed to assess whether a fact-check expresses reasoning and outcome in a simple and understandable way. Our results show that, in this case, raters ranked the crowdsourced fact-check (mean=4.24, SD=0.86) as more clear and comprehensive than the professional fact-checks (mean=4.01, SD=1.03). The difference

between the two sets of articles was significant ($p<0.05$) though the effect size was quite small (0.25).

Finally, our third measure of "persuasiveness" aimed to capture whether readers might find a fact-check is strong enough to convince them. Our results show that raters ranked the professional fact-check articles as somewhat more persuasive. The mean rating for professional fact-check articles was 4.14 (SD=0.96), compared to 3.83 (SD=1.01) for the crowdsourced fact-check articles' persuasiveness. The difference between the two sets of articles was significant ($p<0.01$) though the effect size was relatively small (0.32).

In summary, these findings indicate that professional fact-checks articles are found to be more persuasive and more objective than crowdsourced fact-checks by our raters, but that crowdsourced fact-checking articles received higher ratings for their clarity. At the same time, the differences between the two sets of articles, while significant, were not substantial. Moreover, all the measures received rating averages of roughly 4.0 for both sets of articles, indicating that fact-checks from both sources are generally perceived as objective, clear, and persuasive.

Discussion

Using crowdsourcing power to help with fact-checking is an attractive approach for many fact-check needs (Pinto et al. 2019; Allen et al. 2020; Kazemi et al. 2021b), but requires a better understanding of the strengths and weaknesses of crowdsourced fact-checking compared to professional fact-checking. Therefore, we proposed our 4V framework to evaluate the relative contributions of crowdsourced and professional fact-checking. Specifically, we used the framework to compare Cofacts, as an example of a crowdsourced fact-checking community, to two professional fact-checking services. We note, though, that the framework could prove useful to evaluate the contributions of different fact-checking efforts, not limited to the crowdsourcing context.

The Variety analysis showed that crowdsourced fact-checking can cover most topics in professional fact-checks and even provide answers to many issues that professionals may skip, though some questions remain about how to interpret the difference in coverage. In particular, we cannot provide any normative judgment on the topic preferences of crowdsourced and professional fact-checkers topics. The professionals provide a national service and invest more resources into high-visibility issues like COVID. Comparatively, crowdsourced fact-checking is more grassroots, and driven by crowd-based requests, therefore putting more resources into local and daily affairs and common fraud messages.The breadth of response to community needs on varied topics is a critical offering, as it helps mitigate the effects of "everyday misinformation" users encounter (Lu et al. 2020; Wahlheim, Alexander, and Peske 2020), which may equip readers with cognitive defense to ward off potential harms from misinformation, and reduce the dissemination of suspicious stories (Pennycook et al. 2021; Ecker et al. 2022).

At the same time, the responses from crowdsourced fact-checkers also build on the support of abundant fact-checks on professional sites. Roughly half of the responded requests on Cofacts could be answered by simply referring to existing fact-checks. Professional fact-checking certainly does not have the capacity to actively respond to (often recurring) requests. Under this circumstance, professional fact-checking becomes a manufacturer of knowledge, and crowdsourced fact-checkers play the role of distributors, connecting requests and answers in the information market. Standing on the shoulders of professionals gives crowdsourced fact-checkers critical support in directly countering misinformation, more comprehensively, and faster.

We note that citing professional fact-checks could also help the crowdsourcing community bring in the perspective of global cross-language fact-checking. We observed in the Cofacts data that crowdsourced fact-checkers occasionally also refer to English professional fact-checking sites like Snopes, sometimes with translated summaries. This contribution is unique and important because manual claim-matching, though not scaled, can largely help with the knowledge dispersion in a cross-language way and counter the misinformation that is originated from other countries or debunked by other fact-checkers (Kazemi et al. 2021a). Because sometimes only local fact-checkers can have the ability and knowledge to check a story (Ribeiro et al. 2021), reusing fact-checks in other languages can further reduce workloads and increase the capacity of fact-checking.

The Velocity findings follow a similar theme: when crowdsourced fact-checks are not building on the earlier contributions of professionals, we find that they still respond more rapidly to fact-checking needs than professional fact-checkers. Our result holds for both rapid responses (which normally take a few hours) or slow responses (which normally take more than 24 hours), and the advantage of crowdsourced fact-checking is usually as substantial as several days, and is not affected by the topic. Since misinformation usually spread quickly on social media (Vosoughi, Roy, and Aral 2018), a faster fact-check response is necessary to restrain greater damage.

Crowdsourcing power could also help with identifying potential misinformation given its distinctive variety its advantage in velocity. Messaging platforms like Line and WhatsApp even allow users to report suspicious messages to third-party fact-checkers like Cofacts or to platform fact-checkers (Kazemi et al. 2021b), which may improve the efficiency of both crowdsourced and professional fact-checking.

Naturally, the Veracity of crowdsourced fact-checking is one of the most important concerns. Our data suggest that taking professional articles as ground truth (of course, itself a challenging proposition), crowdsourced fact-checking can provide answers almost as reliably as professionals. Our veracity findings are also consistent with the conclusion about crowdsourced contributors with higher political knowledge (Godel et al. 2021): mobilizing a more special and savvy sample of the population can be a great help for fact-checking. This is not surprising because we believe that the self-selective crowdsourcing model in Cofacts would motivate users with more experience and knowledge out of the crowd and provide more reliable answers than average people in experiment settings (Pennycook and Rand 2019;

Godel et al. 2021; Kaufman, Haupt, and Dow 2022). However, as we further discuss below, there could be long-term challenges in engaging these types of individuals and preventing potential bias and manipulation in the future.

Meanwhile, our Viability findings suggest that crowdsourced fact-checking articles are perceived as nearly as persuasive and objective as professional fact-checking, and even perform slightly better on a clarity measure. Those differences were all small in terms of effect size, suggesting that perhaps there is no substantial difference in the qualities between the two types of fact-checking from the perspectives of readers. On the other hand, the significant difference in rates may also imply that the language style, as a medium of fact-checking, could make a difference in convincing readers of its viability (Nieminen and Rapeli 2019). Professional fact-checks are usually longer and contain detailed domain knowledge, which signals their expertise and objectivity as well as creates a barrier for various readers to understand, as simple language could better facilitate the corrections of misbelief (Ecker et al. 2022). However, the contribution of the crowdsourced community could help make professional content more accessible to readers. In our annotation task, we excluded all crowdsourced answers which solely cite professional fact-checks. But in practice, this kind of paraphrase or a summary of a professional article by crowdsourced contributors may provide this desired improved accessibility.

We have two practical limitations in our study, which are due to the methods we used in our approach. We relied on the textual matching method to identify the comparable pairs of crowdsourced and professional fact-checks. This performance of such an approach, with its potential biases, may impact the results. Another practical limitation is our rater task, where we asked raters to provide an evaluation of fact-check articles. Since we lack good tools to reach out to non-Western citizens, our raters were recruited from a student population at a prestigious university, and thus do not provide a good sample of the average Taiwanese users.

Broader Perspectives The data used for our analysis was obtained from publicly shared datasets and public fact-checking websites. There is no personally identifiable information in the data. Our research was also reviewed by the Institutional Review Board at our institution. Overall, we do not believe the data and analysis raise any significant ethical concerns.

Our work highlights the value of crowdsourcing communities like Cofacts which can mobilize dedicated individuals online to counter misinformation on social media along with professional fact-checkers. Our approach assumes that fact-checking is a positive societal contribution, and further assumes that individuals undertake it with a commitment to provide accurate information to the best of their knowledge. However, the concern still remains that a crowdsourced system could be abused by malicious users, as can be observed in many other systems and also acknowledged by Cofacts (Davis, Crowley, and Corcoran 2019). In the extreme, our work here can inform and motivate such users, although we believe the risk of that is low. Instead, we hope this work can encourage support for crowdsourcing services, and highlight the need to protect them e.g. by preventing inauthentic behaviors or information pollution campaigns (Shachaf and Hara 2010; Rawat et al. 2019).

Conclusion

To address the challenge of evaluating the effectiveness of crowdsourced fact-checking, especially in comparison to professionals, we proposed a 4V evaluation framework of variety, velocity, veracity, and viability. We then applied our framework to assess the contributions of a popular crowdsourced fact-checking community and two professional fact-checking sites in Taiwan. Our results were encouraging. We believe that our findings show, at least for the case at hand, that crowdsourced fact-checking offers distinct but quality contributions along the multiple dimensions that are comparable to professional fact-checking efforts. There were key differences, still. While crowdsourcing can provide broader and faster coverage, it also often relies on the professionals' contribution. Our findings provide a hopeful indication that community-based crowdsourced approaches could offer important support to counter online misinformation, helping to advance a society less vulnerable to the challenges of the present and future.

References

Allen, J.; Arechar, A. A.; Rand, D. G.; and Pennycook, G. 2020. Crowdsourced Fact-Checking: A Scalable Way to Fight Misinformation on Social Media.

Allen, J.; Martel, C.; and Rand, D. G. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *CHI Conference on Human Factors in Computing Systems*, 1–19.

Brashier, N. M.; Pennycook, G.; Berinsky, A. J.; and Rand, D. G. 2021. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5).

Cha, M.; Gao, W.; and Li, C.-T. 2020. Detecting fake news in social media: an Asia-Pacific perspective. *Communications of the ACM*, 63(4): 68–71.

Chang, H.-C. H.; Haider, S.; and Ferrara, E. 2021. Digital civic participation and misinformation during the 2020 Taiwanese presidential election. *Media and Communication*, 9(1): 144–157.

Coleman, K. 2021. Introducing Birdwatch, a community-based approach to misinformation.

Davis, R.; Crowley, B. J.; and Corcoran, C. 2019. Civil society: A key player in the global fight against misinformation. *Kennedy School Review*, 19: 171–173.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ecker, U.; Lewandowsky, S.; Cook, J.; Schmid, P.; Fazio, L.; Brashier, N.; Kendeou, P.; Vraga, E.; and Amazeen, M. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Review Psychology*, 1: 13–29.

Florin, F. 2010a. Crowdsourced Fact-Checking? What We Learned from Truthsquad.

Florin, F. 2010b. Introducing Truthsquad.

Geiger, D.; Seedorf, S.; Schulze, T.; Nickerson, R. C.; and Schader, M. 2011. Managing the crowd: towards a taxonomy of crowdsourcing processes.

Giles, J. 2005. Internet encyclopaedias go head to head. *Nature*, 438: 900–901.

Godel, W.; Sanderson, Z.; Aslett, K.; Nagler, J.; Bonneau, R.; Persily, N.; and Tucker, J. 2021. Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*, 1(1).

Graves, L.; and Amazeen, M. A. 2019. Fact-checking as idea and practice in journalism. In *Oxford Research Encyclopedia of Communication*.

Greenstein, S.; and Zhu, F. 2018. Do experts or crowd-based models produce more bias? Evidence from Encyclopædia Britannica and Wikipedia. *Mis Quarterly*.

Hassan, N.; Yousuf, M.; Mahfuzul Haque, M.; A. Suarez Rivas, J.; and Khadimul Islam, M. 2019. Examining the roles of automation, crowds and professionals towards sustainable fact-checking. In *Companion Proceedings of The 2019 World Wide Web Conference*, 1001–1006.

Hu, M.; Lim, E.-P.; Sun, A.; Lauw, H. W.; and Vuong, B.-Q. 2007. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 243–252.

Javanmardi, S.; and Lopes, C. 2010. Statistical measure of quality in Wikipedia. In *Proceedings of the First Workshop on Social Media Analytics*, 132–138.

Kaufman, R. A.; Haupt, M. R.; and Dow, S. P. 2022. Who's in the Crowd Matters: Cognitive Factors and Beliefs Predict Misinformation Assessment Accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–18.

Kazemi, A.; Garimella, K.; Gaffney, D.; and Hale, S. A. 2021a. Claim Matching Beyond English to Scale Global Fact-Checking. *arXiv preprint arXiv:2106.00853*.

Kazemi, A.; Garimella, K.; Shahi, G. K.; Gaffney, D.; and Hale, S. A. 2021b. Tiplines to Combat Misinformation on Encrypted Platforms: A Case Study of the 2019 Indian Election on WhatsApp. *arXiv preprint arXiv:2106.04726*.

Kittur, A.; and Kraut, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 37–46.

La Barbera, D.; Roitero, K.; Demartini, G.; Mizzaro, S.; and Spina, D. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. *Advances in Information Retrieval*, 12036: 207.

Lim, C. 2018. Checking how fact-checkers check. *Research & Politics*, 5(3): 2053168018786848.

Lu, Z.; Jiang, Y.; Lu, C.; Naaman, M.; and Wigdor, D. 2020. The government's dividend: complex perceptions of social media misinformation in China. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Micallef, N.; Armacost, V.; Memon, N.; and Patil, S. 2022. True or False: Studying the Work Practices of Professional Fact-Checkers. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–44.

Nieminen, S.; and Rapeli, L. 2019. Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, 17(3): 296–309.

Nieminen, S.; and Sankari, V. 2021. Checking PolitiFact's Fact-Checks. *Journalism Studies*, 22(3): 358–378.

Oreg, S.; and Nov, O. 2008. Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values. *Computers in human behavior*, 24(5): 2055–2073.

Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. G. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855): 590–595.

Pennycook, G.; and Rand, D. G. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7): 2521–2526.

Pinto, M. R.; de Lima, Y. O.; Barbosa, C. E.; and de Souza, J. M. 2019. Towards fact-checking through crowdsourcing. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 494–499. IEEE.

Porter, E.; and Wood, T. J. 2021. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, 118(37).

Priedhorsky, R.; Chen, J.; Lam, S. T. K.; Panciera, K.; Terveen, L.; and Riedl, J. 2007. Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, 259–268.

Pröllochs, N. 2022. Community-based fact-checking on Twitter's Birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 794–805.

Rawat, C.; Sarkar, A.; Singh, S.; Alvarado, R.; and Rasberry, L. 2019. Automatic Detection of Online Abuse and Analysis of Problematic Users in Wikipedia. In *2019 Systems and Information Engineering Design Symposium (SIEDS)*, 1–6. IEEE.

Ribeiro, M. H.; Zannettou, S.; Goga, O.; Benevenuto, F.; and West, R. 2021. What do fact checkers fact-check when? *arXiv preprint arXiv:2109.09322*.

Saeed, M.; Traub, N.; Nicolas, M.; Demartini, G.; and Papotti, P. 2022. Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare With Experts? In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 1736–1746.

Shachaf, P.; and Hara, N. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3): 357–370.

Sunstein, C. R. 2006. *Infotopia: How many minds produce knowledge*. Oxford University Press.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.

Wahlheim, C. N.; Alexander, T. R.; and Peske, C. D. 2020. Reminders of everyday misinformation statements can enhance memory for and beliefs in corrections of those statements in the short term. *Psychological Science*, 31(10): 1325–1339.

Walter, N.; Cohen, J.; Holbert, R. L.; and Morag, Y. 2020. Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3): 350–375.

Wilkinson, D. M.; and Huberman, B. A. 2007. Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis*, 157–164.

Xu, Y.; and Luo, T. 2011. Measuring article quality in Wikipedia: Lexical clue model. In *2011 3rd Symposium on Web Society*, 141–146. IEEE.

Young, D. G.; Jamieson, K. H.; Poulsen, S.; and Goldring, A. 2018. Fact-checking effectiveness as a function of format and tone: Evaluating FactCheck.org and FlackCheck.org. *Journalism & Mass Communication Quarterly*, 95(1): 49–75.