Diffusion of Community Fact-Checked Misinformation on Twitter

CHIARA DROLSBACH, JLU Giessen, Germany
NICOLAS PRÖLLOCHS, JLU Giessen, Germany

The spread of misinformation on social media is a pressing societal problem that platforms, policymakers, and researchers continue to grapple with. As a countermeasure, recent works have proposed to employ non-expert fact-checkers in the crowd to fact-check social media content. While experimental studies suggest that crowds might be able to accurately assess the veracity of social media content, an understanding of how crowd fact-checked (mis-)information spreads is missing. In this work, we empirically analyze the spread of misleading vs. not misleading community fact-checked posts on social media. For this purpose, we employ a dataset of community-created fact-checks from Twitter's "Birdwatch" pilot and map them to resharing cascades on Twitter. Different from earlier studies analyzing the spread of misinformation listed on third-party fact-checking websites (e. g., snopes.com), we find that community fact-checked misinformation is less viral. Specifically, misleading posts are estimated to receive 36.62 % fewer retweets than not misleading posts. A partial explanation may lie in differences in the fact-checking targets: community fact-checkers tend to fact-check posts from influential user accounts with many followers, while expert fact-checks tend to target posts that are shared by less influential users. We further find that there are significant differences in virality across different sub-types of misinformation (e. g., factual errors, missing context, manipulated media). Moreover, we conduct a user study to assess the perceived reliability of (real-world) community-created fact-checks. Here, we find that users, to a large extent, agree with community-created fact-checks. Altogether, our findings offer insights into how misleading vs. not misleading posts spread and highlight the crucial role of sample selection when studying misinformation on social media.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Social media**; • **Information systems** → **Crowdsourcing**.

Additional Key Words and Phrases: social media, misinformation, fact-checking, crowd wisdom, information diffusion

1 INTRODUCTION

There are widespread concerns that misinformation on social media is damaging societies and democratic institutions [17]. In recent years, viral misinformation on social media has been observed repeatedly, especially during elections and crisis situations [1, 5, 22, 25]. In order to identify and eventually curb the spread of misinformation, experts fact-checkers on various third-party fact-checking organizations (e. g., snopes.com, politifact.com, factcheck.org) regularly investigate the veracity of social media rumors [45, 47]. However, due to the limited amount of fact-checks that can be performed by these organizations, they are unable to accommodate the amount and speed of content creation on social media. Misinformation thus often continues to circulate and may only be detected when a tremendous amount of attention is paid to it [11]. Furthermore, about 50% of all Americans have concerns regarding the independence of the experts' assessment, i. e., distrust professional fact-checkers [29]. Given these challenges, the real-world impact of fact-checks from third-party fact-checking organizations may be limited.

In order to address the drawbacks of the expert verification approach, recent research has proposed to employ non-expert fact-checkers in the crowd to verify social media content [2, 3, 6, 11, 14, 20, 27]. The rationale is that the "wisdom of crowds" (i. e., the aggregated assessments of non-expert fact-checkers) could result in an accuracy that is similar to that of experts [12].

Compared to the expert verification approach, harnessing the crowd for fact-checking would enable large numbers of fact-checks that could be carried out at higher frequency and lower cost [2, 27]. Furthermore, crowd-based fact-checking has the potential to remedy the problem of distrust in expert fact-checkers [2]. Recent experimental studies indeed yielded promising results – suggesting that even relatively small crowds achieve an accuracy comparable to that of experts when fact-checking social media content [6, 11, 27].

While community-based fact-checking systems might be able to produce accurate fact-checks at scale, an understanding of how (mis-)information diffuses through social networks is still in its infancy. Prior works have analyzed the spread of rumors that have been fact-checked by third-party fact-checking organizations [13, 31, 40, 45]. For instance, several studies have compared characteristics of resharing cascades (e. g., how often a social media post is shared) across true vs. false rumors, finding that falsehood is more viral than the truth [31, 40, 45]. However, third-party fact-checking organizations tend to fact-check rumors on topics that are deemed to be of interest to a broad public and/or particularly concerning from the perspective of experts, while other misinformation remains unnoticed. In contrast, community fact-checked posts represent social media content that has been deemed worth fact-checking by actual social media users. Analyzing their diffusion would shed new light on the question of whether misinformation is more viral than the truth – or rather a result of sample selection. However, we are not aware of any previous research analyzing the diffusion of crowd fact-checked posts on social media. Moreover, little is known about which social media posts are picked up in community-based fact-checking and how the spread varies across different types of misinformation (e. g., factual errors, missing context). Answering these questions is the goal of this study.

Research questions: In this work, we empirically analyze the diffusion of misleading vs. not misleading social media posts that have been fact-checked by the crowd. Specifically, we address the following research questions:

- **(RQ1)** How do community fact-checked posts spread on social media? Are misleading posts more viral than not misleading posts?
- **(RQ2)** Are there differences in virality across different sub-types of community fact-checked misinformation (e. g., factual errors, missing context, manipulated media)?
- **(RQ3)** How do the fact-checking targets differ between community fact-checkers and expert fact-checkers?
- **(RQ4)** To what extent are (real-world) community-created fact-checks perceived as reliable?

Data & methodology: We collect a comprehensive dataset consisting of community-created fact-checks from Twitter's Birdwatch platform. We then map the fact-checks to the fact-checked tweet using Twitter's historical API. This allows us to calculate the size of the resharing cascades (i. e., the number of retweets) in order to measure the virality of the fact-checked post. Subsequently, we implement an empirical regression model and link the fact-checking label to the number of retweets. We further control for the sentiment of the post and the social influence of its author (e. g., number of followers, account age, etc.). We then perform hypothesis testing to analyze whether posts categorized as being misleading are more viral than not misleading posts.

Contributions: This study is the first to analyze the spread of crowd fact-checked misinformation on social media. We show that crowd fact-checked misleading posts are *less* viral than not misleading posts. Specifically, misleading posts are estimated to receive 36.85 % fewer retweets than not misleading posts. Notably, this finding differs from earlier work [45], which has analyzed the diffusion of misinformation that has been fact-checked by third-party fact-checking organizations. We find that a partial explanation may lie in differences in the fact-checking targets: our findings suggest that community fact-checkers tend to fact-check posts from influential user accounts with

many followers, while expert fact-checks tend to target rumors that are shared by less influential accounts. Our results further imply that there are significant differences in virality across different sub-types of misinformation (e. g., factual errors, missing context, manipulated media).

As an additional contribution, we conduct a user study to assess the perceived reliability of (real-world) community-created fact-checks. Here, we observe that users agree with a large share of community-created fact-checks, whereas only a relatively small share is perceived as being purposely deceptive (e.g., due to motivated reasoning).

2 BACKGROUND

2.1 Misinformation on Social Media

Over the last decade, the importance of social media (e. g., Twitter, Facebook) as an information platform for large parts of society has been subject to considerable growth [17, 28]. On social media, any user can share information with his/her follower base [38]. Compared to traditional media, there is little control authority or oversight regarding the contents. For this reason, social media is highly vulnerable to the spread of misinformation. In fact, previous research suggests that social media platforms have become primary enablers of misinformation [e. g., 17]. Online exposure to misinformation can affect how opinions are formed and causes detrimental societal effects [e. g., 1, 9]. The latter has been repeatedly observed, especially during elections [e. g., 1, 5] and crisis situations [e. g., 22, 23, 26, 40, 41].

A key feature of modern social media platforms is that users can also share others' content to increase its reach (e. g., "retweeting" on Twitter). This can result in misinformation cascades going "viral." While previous research has mainly focused on characteristics and (negative) consequences of misinformation on social media, studies analyzing differences in the virality across misleading vs. not misleading posts are relatively scant. Existing works in this direction have analyzed the diffusion of posts that have been fact-checked by third-party fact-checking organizations [13, 31, 33, 40, 45]. These studies found that misinformation diffuses significantly more virally than the truth. We are not aware of any previous study analyzing the spread of misleading vs. not misleading social media posts that have been fact-checked by the crowd.

2.2 Fact-Checking on Social Media

Reliable fact-checking strategies are a crucial necessity to limit the spread of misinformation on social media. Currently, there are two predominant strategies. First, expert assessment in the form of human experts can check the veracity of content; e. g., via third-party fact-checking platforms (e. g., snopes.com, politifact.com, factcheck.org). Second, machine learning models can be trained to automatically classify misinformation [18, 34]. For this purpose, content-based features (e. g., text, images, video), context-based features (e. g., time, location), or propagation patterns (i. e., how misinformation circulates among users) can be used. However, both methods suffer from several drawbacks. While experts classify misinformation fairly accurately, this strategy is difficult to scale due to the limited number of available humans experts [19, 27]. Besides, a large proportion of social media users do not trust the independence of expert fact-checkers [29]. In contrast, machine learning-based approaches are straightforward to scale, but typically show comparatively low accuracy [47].

Given the trade-off between scalability and accuracy of existing approaches, recent works have proposed to outsource fact-checking of social media content to non-expert fact-checkers in the crowd [2, 3, 6, 11, 14, 20, 27]. The rationale is that the "wisdom of crowds" (i. e., the aggregated assessments of non-expert fact-checkers) could result in an accuracy that is comparable to that of experts [12, 46]. The ability of crowds to ensure relatively trustworthy and high-quality accumulation of

knowledge has been observed in various other online settings, such as on platforms like Wikipedia and Stack Overflow [e. g., 10, 15, 24]. Applying a crowd-based approach to fact-check social media posts might have several benefits [27]. First, compared to expert assessments, significantly larger quantities of posts could be fact-checked. Second, trust issues with expert fact-checkers could, at least partially, be mitigated. Experimental studies suggest that, while the assessment of individuals might be noisy and ineffective [46], the crowd can be quite accurate in identifying misleading social media content. Here the assessment of even relatively small crowds has been found to be comparable to those of experts [6, 11, 27]. Despite challenges with politically motivated reasoning [4, 30], recent research further shows that users, to a large extent, perceive community-created fact-checks for social media posts as being informative and helpful [30].

3 DATA

3.1 Data Source: Community Fact-Checked Tweets from Birdwatch

We analyze the spread of social media posts that have been community fact-checked on Twitter's Birdwatch pilot [44]. On January 23, 2021, Twitter launched Birdwatch as a new approach to address misinformation on their platform [44]. The goal is to fact-check social media content by harnessing the "wisdom of crowds." Birdwatch allows users to identify tweets they believe are misleading or not misleading and write notes that provide context to the tweet (so-called "Birdwatch notes"). Users can fact-check *any* tweet they come across on Twitter – directly when browsing Twitter (see examples in Fig. 1). Community fact-checking on Birdwatch comprises (1) checkbox questions that allow users to state whether a tweet might or might not be misleading (*Fact-Checking Label*); (2) an open text field (max 280 characters) where users can explain their judgment (*Text Explanation*), and (3) checkbox questions in which users can characterize the tweet and select reasons *why* they perceive the tweet as being misleading (*Misinformation Type*). For the latter, Birdwatch users can select one (or multiple) of the following answer options: (i) "Factual Error," (ii) "Missing Important Context," (iii) "Unverified Claim as Fact," (iv) "Outdated Information," (v) "Manipulated Media," (vi) "Satire," and (vii) "Other."

After a Birdwatch note is submitted, the fact-check is publicly available for other users to read. Birdwatch also features a rating system, which allows users to rate the helpfulness of the community-created fact-checks. These ratings are supposed to help identify which notes are most helpful and raise their visibility. Specifically, Birdwatch notes are shown directly on the fact-checked tweet if (i) the tweet is classified as misleading and (ii) it is rated by the community to be particularly helpful (see Fig. 1).

Importantly, the data for this study originates from Birdwatch's pilot phase in the U. S. During this pilot phase, interested users were required to actively sign up to join Birdwatch. Any Twitter user could apply to become a Birdwatch contributor. Users that had signed up on Birdwatch could see Birdwatch notes directly when browsing Twitter next to the fact-checked tweet. Non-participating users could access Birdwatch notes via a separate Birdwatch website (birdwatch.twitter.com). In early 2022, Birdwatch had approximately 3250 contributors, compared to 41.5 million daily active Twitter users in the U. S. [42]. Hence, during Birdwatch's pilot phase, community fact-checks from Birdwatch were practically not visible to the vast majority of social media users and, thus, were unlikely to directly influence the diffusion of the fact-checked tweets.[1] The Birdwatch pilot phase thus provides a unique opportunity to study the spread of community fact-checked posts with little confounding factors.

---

[1]In early October 2022 (i. e., after our observation period), Twitter started to expand the Birdwatch program, allowing more Twitter users to view fact-checks directly on Twitter. Furthermore, Twitter rebranded Birdwatch to "Community Notes."

(a) Misleading · (b) Not misleading

Fig. 1. Examples of community fact-checked tweets. Only Birdwatch notes for misleading tweets are eligible to be directly shown on tweets. However, during our study period, community fact-checks from Birdwatch were practically not visible to the vast majority of social media users (i. e., only to pilot participants) and, thus, were unlikely to directly influence the diffusion of the fact-checked tweets. (a) Example of a tweet classified as misleading (*Fact-Checking Label*) and the *Text Explanation* of the corresponding Birdwatch note. The contributor selected "Factual Error," "Manipulated Media," and "Satire" as reasons for his/her classification (*Misinformation Type*). (b) Example of a tweet classified as not misleading.

3.2 Data Collection

We downloaded *all* Birdwatch notes between the introduction of the feature on January 23, 2021, and the end of February 2022 from the Birdwatch website, i. e., for an observation period of more than one year. The dataset contains a total number of 20 218 Birdwatch notes (i. e., community-created fact-checks) from 3 257 different contributors. We used the Twitter historical API to map the *tweetID* referenced in each Birdwatch note to the source tweet. This approach allowed us to collect the following information about each source tweet and the account of its authors: (i) the number of retweets, (ii) the number of followers, (iii) the number of followees, (iv) the account age, and (v) whether the user has been verified by Twitter.

Notably, multiple Birdwatch users can write Birdwatch notes for the same tweet. Therefore, the data sometimes includes multiple fact-checks for the same post. The average number of Birdwatch notes per tweet is 1.33, with few tweets having many notes and most tweets having few. Only 18.79% of the fact-checked tweets received more than one Birdwatch note. To avoid distortions due to multiple fact-checked tweets, we focus our analysis on the temporally first fact-check after the tweet has been posted. This filtering step resulted in a dataset consisting of 15 256 unique fact-checks (for 15 256 unique tweets). As part of our robustness checks, we also tested alternative approaches for handling multiple fact-checks (e. g., using Birdwatch's rating system, majority vote). Here we obtained qualitatively identical results.

3.3 Variable Description

Our dataset contains variables from two sources: (i) variables that are provided by the community-created fact-checks (i. e., the Birdwatch notes); and (ii) variables that represent information about the source tweet (e. g., the social influence of the author of the fact-checked tweet).

Fact-checks: The Birdwatch notes provide us with the following variables:

- *Misleading:* A binary indicator of whether a tweet has been reported as being misleading by the author of the Birdwatch note (= 1; otherwise = 0).
- *Delay:* A numeric variable measuring the number of days elapsed between the posting date of the source tweet and the fact-check.
- *Misinformation Type:* Seven dummy variables indicating reasons why a tweet has been reported as being misleading ("Factual Error," "Missing Important Context," "Unverified Claim as Fact," "Outdated Information," "Manipulated Media," "Satire," and "Other").

Source tweet: We used the Twitter historical API to map the *tweetID* referenced in each Birdwatch note to the source tweet and collected the following information about each source tweet:

- *Retweet Count:* A numeric variable denoting the number of retweets a single tweet receives on Twitter. The retweet count is a common measure for the virality of a resharing cascade [e. g., 8, 40].
- *Followers:* The number of followers, i. e., the number of accounts that follow the author of the source tweet on Twitter.
- *Followees:* The number of followees, i. e., the number of accounts whom the author of the source tweet follows on Twitter.
- *Account Age:* The age of the author of the source tweet's account (in years).
- *Verified:* A binary dummy indicating whether the account of the source tweet has been officially verified by Twitter (= 1; otherwise = 0).
- *Sentiment:* We calculate a sentiment score measuring the positivity/negativity of the source tweet. Here we use a dictionary-based approach analogous to earlier research [e. g., 7, 16, 35, 37, 45]. We first remove stopwords, punctuation, special characters (e. g., hashtags), and URLs in each source tweet. Subsequently, we employ the NRC lexicon [21], which categorizes English words into positive and negative words. Following previous work [e. g., 35, 39], the sentiment scores are then measured by calculating the difference between positive and negative words relative to the tweet length. For our sentiment analysis, we use the default implementation of the `sentimentr` package (with the built-in NRC lexicon) that also accounts for negations and valence shifters (see [36] for details).

4 EMPIRICAL ANALYSIS

4.1 Diffusion of Misleading vs. Not Misleading Posts (RQ1)

We now empirically analyze the diffusion of misleading vs. not misleading posts that have been fact-checked on Twitter's Birdwatch platform. For this purpose, we first compare summary statistics. Note, however, that summary statistics should be interpreted with caution as the virality of social media posts strongly depends on the social influence of the author. To account for such confounding effects, we subsequently implement an empirical regression model with control variables that links the fact-checking label to the number of retweets. We then perform hypothesis testing to analyze whether posts categorized as being misleading are more viral than not misleading posts.

Summary statistics: Birdwatch users are vastly more likely to report misleading tweets than not misleading tweets. Out of 15 256 community fact-checked tweets, 14 384 (94.28 %) are classified as misleading and 872 (5.72 %) are classified as not misleading. In total, the fact-checked tweets in our dataset have been retweeted 29.45 million times. However, the retweet volume is higher for not misleading tweets than for misleading tweets. Specifically, the average retweets count amounts to 2 478 for not misleading tweets and to 1 478 for misleading tweets. A two-sided *t*-test confirms that the difference in means are statistically significant ($p < 0.01$). Misleading vs.

not misleading tweets also exhibit considerable heterogeneity with regards to sentiment and the social influence of the author. The sentiment tends to be significantly more positive in not misleading tweets (mean sentiment of 0.022) than in misleading tweets (mean sentiment of −0.004). Misleading tweets are posted by users that have, on average, 41.17 % fewer followers. Also here, two-sided t-tests confirm that the difference in means are statistically significant ($p < 0.01$). We find only small differences in means for the variables *Followees*, *Account Age* and, *Verified*, which are not statistically significant at common significance thresholds. Fig. 2 further visualizes the complementary cumulative distribution functions (CCDFs). Kolmogorov-Smirnov (KS) tests show that, with the exception of *Account Age*, the differences in the distributions between misleading and not misleading tweets are statistically significant ($p < 0.01$).

Fig. 2. Complementary cumulative distribution functions (CCDFs) for (a) *Retweet Count*, (b) *Sentiment*, (c) *Followers*, (d) *Followees*, (e) *Account Age*, and (f) *Delay*.

Regression model: We implement explanatory regression analysis to better understand the diffusion of misleading vs. not misleading crowd fact-checked posts. In contrast to summary statistics, this allows us to estimate effect sizes after controlling for confounding effects. The dependent variable in our regression analysis is given by $RetweetCount_i$, that is, the number of retweets for a fact-checked tweet *i*. The retweet count is a non-negative count variable, and its variance is larger than the mean. To adjust for overdispersion, we draw upon a negative binomial regression to model the retweets count [31, 40]. The key explanatory variable is $Misleading_i$, i. e., whether the tweet has been classified as misleading by Birdwatch users (i. e., = 1 if true, otherwise = 0). Additionally, we include the elapsed time between the publication of the tweet and the fact-check ($Delay_i$). Furthermore, we must control for the social influence of the source tweet and its author. Therefore, we adjust for variables known to affect the retweet rate [8, 32, 40, 43, 45], which includes the number of followers ($Followers_i$) and followees ($Followees_i$), the account age ($AccountAge_i$), and whether the account was verified by Twitter ($Verified_i$). In addition, we control for the sentiment of the source tweet ($Sentiment_i$). The resulting model is

$$\log(\mathrm{E}(RetweetCount_i \mid {}^*)) = \beta_0 + \beta_1\, Misleading_i \tag{1}$$

$$+ \beta_2\ Delay_i + \beta_3\ Sentiment_i + \beta_4\ Followers_i$$
$$+ \beta_5\ Followees_i + \beta_6\ AccountAge_i + \beta_7\ Verified_i + u_i,$$

with intercept β_0 and month-year fixed effects u_i to adjust for differences in the start date and age of the resharing cascades. For the sake of interpretability, we z-standardize all continuous variables. This allows us to compare the effects of regression coefficients on the dependent variable measured in standard deviations. Note that since we apply a negative binomial regression, the interpretation of the effect sizes requires an exponential transformation of the coefficients.

Coefficient estimates: The coefficient estimates for the regression model are reported in Fig. 3. We find that misleading tweets are significantly less viral than not misleading tweets. Specifically, the coefficient for $Misleading$ is -0.456 ($p < 0.01$), which implies that misleading tweets are expected to receive $e^{-0.459} - 1 \approx 36.62\,\%$ fewer retweets. Furthermore, we observe that the coefficient estimate for $Delay$ is small in magnitude and not statistically significant at common significance threshold. This implies that differences in the fact-checking speed are not significantly associated with differences in virality of crowd fact-checked posts.

Fig. 3. Coefficient estimates for negative binomial regression with the retweet count as dependent variable. Model (a) includes all variables given by the source tweet (orange). Model (b) additionally includes variables concerning the fact-check (green). The vertical bars represent 99 % confidence intervals. Month-year fixed effects are included.

Concordant with the literature [40, 43, 45], we observe statistically significant estimates for the variables characterizing the social influence of the author of the source tweet. The number of followers has a large positive effect on the number of retweets (coef: 0.267; $p < 0.01$), while the number of followees has a smaller positive effect (coef: 0.076; $p < 0.01$). A higher account age decreases the expected number of retweets (coef: -0.216; $p < 0.01$), while posts from verified accounts are expected to receive more retweets (coef: 0.783; $p < 0.01$). Similar to earlier work [31], we also find that more positive sentiment is associated with more retweets (coef: 0.111; $p < 0.01$).

4.2 Diffusion of Different Types of Misinformation (RQ2)

If fact-checkers on Birdwatch have classified a tweet as being misleading, they additionally need to answer checkbox questions on the reasons *why* they perceive it as such. As aforementioned, Birdwatch users can select one (or multiple) of the following answer options: (i) "Factual Error," (ii) "Missing Important Context," (iii) "Unverified Claim as Fact," (iv) "Outdated Information," (v) "Manipulated Media," (vi) "Satire," and (i) "Other." Fig. 4 shows that the vast majority of tweets have

been categorized as misleading because of factual errors (62.13 %), missing context (61.38 %), or because they treat unverified claims as fact (49.99 %). The other categories are relatively rare.

Fig. 4. Barplot showing the number of tweets per checkbox answer option in response to the question "Why do you believe this tweet may be misleading?"

We repeat our regression analysis with dummy variables referring to the different types of misleading posts as provided by Birdwatch contributors. This allows us to examine differences in the virality across different types of misinformation. The coefficient estimates in Fig. 5 show that misleading tweets are less viral than not misleading tweets if they belong to the misinformation sub-types "Factual Error" (coef: -0.251; $p < 0.01$), "Missing Important Context" (coef: -0.127; $p < 0.01$), "Unverified Claim as Fact" (coef: -0.300; $p < 0.01$) and, "Other" (coef: -0.221; $p < 0.01$). In contrast, tweets belonging to the misinformation sub-types "Manipulated Media" (coef: 0.461; $p < 0.01$), and "Satire" (coef: 0.411; $p < 0.01$) receive more retweets. These results suggest that there are significant differences in virality across different sub-types of misinformation. The coefficient estimates for the other variables do not differ qualitatively from the previously performed regressions.

Fig. 5. Coefficient estimates for negative binomial regression with the retweet count as dependent variable. Here, dummy variables referring to different sub-types of misinformation are included. Model (a) includes all posts (green), whereas Model (b) only includes the subset of posts classified as misleading (orange). The reference type in Model (a) are tweets classified as "not misleading," whereas the reference type in Model (b) are misleading tweets that have not been assigned to a subtype. The vertical bars represent 99 % confidence intervals. Month-year fixed effects are included.

4.3 Comparison to Expert-Based Fact-Checking (RQ3)

In contrast to the work by Vosoughi et al. (2018), which found that expert fact-checked falsehood on Twitter is *more* viral than the truth, our analysis suggests that crowd fact-checked tweets perceived as misleading are *less* viral than those perceived as not misleading. A possible explanation for this finding lies in the sample selection, i. e., third-party fact-checking organizations vs. Birdwatch contributors might fact-check social media posts published by different account types.

To shed light on this question, Fig. 6 compares the mean values of different user characteristics of the authors of misleading and not misleading crowd fact-checked posts to those of authors true and false rumors in the dataset of expert fact-checked posts from Vosoughi et al. (2018). Compared to expert fact-checked tweets, we find that user accounts of authors of crowd fact-checked posts have, on average, $\approx$ 40 times more followers, 41.65 % more followees, and approximately twice the account age. Moreover, while 49.21 % percent of the accounts of authors of crowd fact-checked posts are verified by Twitter, this is only the case for 2.00 % of the authors of expert fact-checked posts. Two-sided t-tests confirm that each difference in means is statistically significant ($p < 0.01$). These findings suggest that social media users contributing to crowd-based fact-checking tend to fact-check posts from larger accounts with greater social influence, while expert fact-checks tend to target rumors that are shared by smaller accounts.

Fig. 6. Comparison of characteristics (mean values) of authors of crowd fact-checked and expert fact-checked tweets for (a) the number of followers, (b) the number of followees, (c) the account age and, (d) the verified status. We compare the authors of misleading and not misleading crowd fact-checked posts on Birdwatch to those of true and false rumors in the dataset of expert fact-checked posts from Vosoughi et al. (2018).

We further observe that, for expert fact-checked posts, falsehood tends to originate from accounts with relatively more followers, while we observe the opposite pattern for crowd fact-checked posts (see Fig. 6). Specifically, we find that authors of crowd fact-checked posts perceived as misleading have 67.71 % more followers than accounts of posts perceived as not misleading ($p < 0.01$). In contrast, authors of falsehood in expert fact-checked posts have 34.04 % less followers than authors of the true tweets ($p < 0.01$). This suggests that fact-checks from Birdwatch contributors are more likely to endorse/emphasize the accuracy of not misleading tweets authored by influential users with a wide reach. Opposite to this, expert fact-checked tweets authored by influential accounts are more likely to convey false information. Since author characteristics are inherently linked to the virality of posts (e. g., users with a wider reach can generate more retweets), the observed

differences in fact-checking targets provide a (partial) explanation for the overall higher virality of not misleading posts in the case of Birdwatch.

4.4 Perceived Reliability of Community-Created Fact-Checks (RQ4)

In order to assess the perceived reliability of the community-based fact-checks from Birdwatch, we conducted a user study on the online survey platform Prolific (www.prolific.com). We recruited n = 7 participants, four women and three men, who were on average 35 years old. All participants were based in the U. S., and English native speakers. All but one participant indicated that they are familiar with Twitter and regularly share content on social media. Participants were presented with a randomized sample of 300 tweets (150 not misleading and 150 misleading) and the corresponding fact-checks from Birdwatch (fact-checking label and text explanation). Note that we purposely presented the participants with both the source tweet and the fact-check (instead of only the source tweet). In the absence of a ground truth (which might require expert assessment), we were interested in the *perceived* reliability of the fact-checks rather than testing how much one crowd agrees with another. As such, for each tweet, participants were asked for their assessment on (i) the extent to which they agree with the fact-checking label, and (ii) whether they perceive the fact-check as purposely deceptive (e. g., because of motivated reasoning, manipulation attempts, etc.). The participants answered both questions on a 5-point Likert scale, ranging from 1 ("strongly disagree") to 5 ("strongly agree").

Fig. 7 visualizes the distribution of the median votes for the individual tweets across all response options. We first evaluate the extent to which the participants agree with the fact-checks from Birdwatch. We find that the participants at least somewhat agree with 73.33 % of the community-created fact-checks performed by Birdwatch users. Interestingly, the agreement is lower for tweets categorized as misleading (72.00 %) than for tweets classified as not misleading (74.67 %).

We find a consistent pattern for the second question item: the median ratings of the seven participants suggest that only a relatively small share of 7.00 % of fact-checks are perceived as purposely deceptive. Notable, we again observe considerable differences across fact-checks across fact-checks reporting misleading vs. not misleading tweets. Specifically, fact-checks reporting misleading tweets are more likely to be perceived as being purposely deceptive (9.33 %) than fact-checks reporting not misleading tweets (4.67 %).

Fig. 7. User study evaluating the perceived reliability of community-based fact-checks from Birdwatch. $n = 7$ participants were recruited via Prolific. Here we report the median responses to the questions (a) "Do you agree with the fact-checking label?" and (b) "Do you feel that the fact-check is purposely deceptive?"

The participants showed statistically significant inter-rater agreements. Kendall's W was 0.43 ($p < 0.01$) for the first question item (agreement with the fact-checking label); and 0.32 ($p < 0.01$) for the second question item (purposely deceptive fact-checks).

In sum, the results of our user study suggest that the vast majority of community-created fact-checks are perceived as being reliable. This supports the results of previous experimental works, which suggest that the risk of users purposely trying to "game the system" is tolerable [e. g., 2]. Even though inaccurate fact-checks and misuse of the platform cannot be prevented completely, community-based fact-checking should be seen as one tool (as part of a larger toolset) that may help to combat the spread of misinformation on social media [11, 14].

4.5   Robustness Checks

We conducted an extensive set of checks that yielded consistent findings: (1) we controlled for outliers in the dependent variables; (2) we ran separate regressions for misleading vs. not misleading posts; (3) we calculated variance inflation factors for all independent variables and found that all remain below the critical threshold of four; (4) we repeated our analysis with user-specific random effects; (5) we incorporated quadratic effects; (6) we included interaction terms between user-specific variables and the fact-checking label; (7) we evaluated alternative approaches to handle multiple fact-checks for the same tweet (e. g., majority vote, no filtering of multiple fact-checks). In all of these checks, our findings are supported. Detailed results are reported in the supplementary materials.

5   DISCUSSION

Summary of findings: This study is the first to examine the diffusion of misleading vs. not misleading posts on social media that have been fact-checked by the crowd. Our key findings are as follows: (i) community fact-checked misleading tweets receive 36.85 % fewer retweets than not misleading tweets (RQ1). (ii) There are significant differences in virality across different sub-types of misinformation (RQ2). Specifically, we find that misleading tweets are less viral than not misleading tweets across almost all sub-types of misinformation, except for (the relatively rare categories) satire, manipulated media, and outdated information. (iii) The fact-checking targets significantly differ between community fact-checkers and expert fact-checkers (RQ3). In particular, the crowd tends to fact-check posts from accounts with greater social influence (e. g., high-follower accounts).

As an additional contribution, we conducted a user study to assess the perceived reliability of (real-world) community-created fact-checks (RQ4). We find that users agree with a relatively high share (73.33 %) of community-created fact-checks, whereas only a relatively small share (7.00 %) is perceived as being purposely deceptive (e. g., due to manipulation attempts). These results corroborate previous findings of experimental studies, which suggested that crowds can achieve a high level of accuracy when fact-checking social media content [e. g., 2].

Research implications: In contrast to previous research examining the spread of misinformation that has been fact-checked by third-party organizations [31, 40, 45], we find that community fact-checked misleading posts receive fewer retweets than not misleading posts. The diverging results may be a consequence of differences in the sample selection. While third-party organizations tend to fact-check posts on topics experts believe are of broad public interest and/or particularly concerning to society, community fact-checked posts comprise posts that have been deemed to be worth fact-checking by actual social media users.

Our analysis suggests that crowd vs. experts focus on different targets when fact-checking social media content. We find that community fact-checkers tend to fact-check posts from larger accounts with high social influence, while expert fact-checks tend to target rumors shared by smaller accounts. Furthermore, community fact-checkers are relatively more likely to endorse/emphasize the accuracy of not misleading posts authored by influential users (i. e., users with a wide reach). This pattern is opposite to expert fact-checking where posts authored by influential accounts are relatively more likely to convey misinformation. Since author characteristics are inherently linked

to the virality of posts (e. g., users with a wider reach can generate more retweets), the observed differences in fact-checking targets provide a (partial) explanation for the higher virality of not misleading community fact-checked posts. Note, however, that author characteristics are unlikely to be the only reason. In our explanatory regression analysis, we find the pattern that community fact-checked posts are more viral to persist – even after controlling for the social influence of the author. This suggests that there might be additional differences between experts and the crowd in how fact-checking targets are selected (see *Limitations and future research*).

Importantly, while our study complements earlier work studying the diffusion of expert fact-checked posts, we do not claim that the selection by the crowd is more representative for the population of misinformation on social media as a whole. Rather, our results imply that the crowd focuses on different targets when fact-checking social media content and that sample selection plays a key role when studying misinformation diffusion. Compiling a representative sample of *all* misinformation circulating on social media presents an important – yet difficult – challenge for future research.

Practical implications: From a practical perspective, policy initiatives around the world oblige social media platforms to develop countermeasures against misinformation. Community-based fact-checking opens new avenues to increase the scalability and speed of fact-checking of social media content. Furthermore, the community-based approach has the potential to to overcome trust issues associated with expert-created fact-checks [3]. The observed differences in the selection of fact-checking targets between community and expert fact-checkers suggest that both approaches might well complement each other. Here, community-created fact-checking may help to identify misinformation that is actually of interest to actual social media users – and which may go unnoticed on third-party fact-checking organizations. The results of our user study further suggest that the vast majority of community-created fact-checks are perceived as being reliable. Although misuse of the platform cannot be prevented completely, previous research suggests that many issues with bad actors can effectively be addressed using sophisticated ranking mechanisms (e. g., helpfulness ratings), incentivizing high-quality fact-checks (e. g., blocking malicious contributors) or additional community-based content moderation efforts [11, 14]. In sum, community-based fact-checking systems (as part of a larger toolset) allow social media platforms for improved coverage and may help to combat misinformation on social media more effectively.

Limitations and future research: Our work has a number of limitations, which provide promising opportunities for future research. First, similar to related studies [e. g., 40, 45], we do not make causal claims. Future work should thus seek to validate our results in controlled experiments. Second, our user study evaluates the *perceived* reliability of community-based fact-checks. While earlier experimental studies have already shown that crowds can achieve a high level of accuracy when fact-checking social media content (e.g., Allen et al. 2021), it is necessary to further investigate the performance of the crowd in the field (e. g., via expert assessments of Birdwatch notes). Also, more research is necessary to better understand the role of manipulation attempts, and the conditions under which the wisdom of crowds can be unlocked for fact-checking. Third, our study shows that the fact-checking targets in community vs. expert fact-checks differ in terms of their author characteristics (e. g., number of followers). Future research should complement this analysis with a fine-grained study of additional characteristics of the fact-checked posts. For instance, it is a promising extension to employ topic modeling to study how the virality varies across topics (e. g., politics, health, entertainment, etc.) and other misinformation characteristics (e. g., novelty, believability). Fourth, our results are limited to Twitter's Birdwatch pilot. As such, the restricted set of Birdwatch contributors might not be representative for the overall user base on Twitter. Fifth, the community-created fact-checks in our study were not visible to the vast majority of Twitter users (i. e., only to pilot participants), whereas Twitter's goal is that Birdwatch will be

available to everyone on Twitter. Future research may expand the current investigation by studying how (community-based) fact-checking *labels* influence users' sharing behavior on social media.

6   CONCLUSION

The spread of misinformation on social media is a pressing societal problem that platforms, policy-makers, and researchers continue to grapple with. As a countermeasure, recent research proposed to build on crowd wisdom to fact-check social media content. In this study, we empirically analyzed the spread of posts that have been fact-checked by the crowd on Twitter's Birdwatch platform. Different from earlier studies that have analyzed the spread of misinformation fact-checked by third-party organizations, we find that crowd fact-checked misleading posts are less viral than not misleading posts. Our results also suggest that there are significant differences in virality across different sub-types of misinformation (e. g., factual errors, missing context, satire). Altogether, our findings offer insights into how misleading vs. not misleading posts spread and highlight the crucial role of sample selection when studying misinformation on social media.

7   ETHICS STATEMENT

This research did not involve interventions with human subjects, and, thus, no approval from the Institutional Review Board was required by the authors' institutions.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–236.

[2] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7, 36 (2021), eabf4393.

[3] Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances* 6, 14 (2020). eaay3539.

[4] Jennifer Allen, Cameron Martel, and David G Rand. 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*.

[5] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[6] 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. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[7] Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2022. Finding Qs: Profiling QAnon supporters on Parler. *arXiv* 2205.08834 (2022).

[8] William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *PNAS* 114, 28 (2017), 7313–7318.

[9] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *PNAS* 113, 3 (2016), 554–559.

[10] Indika Dissanayake, Sridhar Nerur, Rahul Singh, and Yang Lee. 2019. Medical Crowdsourcing: Harnessing the "Wisdom of the Crowd" to Solve Medical Mysteries. *Journal of the Association for Information Systems* 20, 11 (2019), 4.

[11] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *CHI*.

[12] Vincenz Frey and Arnout van de Rijt. 2021. Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science* 67, 7 (2021), 4273–4286.

[13] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *ICWSM*.

[14] William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A Tucker. 2021. Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety* 1, 1 (2021), 1–36.

[15] Yue Han, Pinar Ozturk, and Jeffrey V. Nickerson. 2021. Leveraging the wisdom of crowd to address societal challenges: A revisit to the knowledge reuse process for innovation through analytics. *Journal of the Association for Information Systems* forthcoming (2021).

[16] Johannes Jakubik, Michael Vössing, Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2022. Online Emotions During the Storming of the US Capitol: Evidence from the Social Media Network Parler. *arXiv* 2204.04245 (2022).

[17] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[18] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

[19] Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. 2022. True or false: Studying the work practices of professional fact-checkers. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–44.

[20] 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 *International Conference on Big Data*.

[21] Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[22] Onook Oh, Manish Agrawal, and H. Raghav Rao. 2013. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly* 37, 2 (2013), 407–426.

[23] Onook Oh, Kyounghee Hazel Kwon, and H. Raghav Rao. 2010. An exploration of social media in extreme events: Rumor theory and Twitter during the Haiti earthquake 2010. In *International Conference on Information Systems (ICIS)*.

[24] Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2014. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology* 65, 12 (2014), 2381–2403.

[25] Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* 66, 11 (2020), 4944–4957.

[26] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science* 31, 7 (2020), 770–780.

[27] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *PNAS* 116, 7 (2019), 2521–2526.

[28] Pew Research Center. 2016. News use across social media platforms 2016. https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

[29] Poynter. 2019. Most Republicans don't trust fact-checkers, and most Americans don't trust the media. https://www.poynter.org/ifcn/2019/most-republicans-dont-trust-fact-checkers-and-most-americans-dont-trust-the-media/.

[30] Nicolas Pröllochs. 2022. Community-based fact-checking on Twitter's Birdwatch platform. In *ICWSM*.

[31] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. 2021. Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports* 11 (2021). 22721.

[32] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. 2021. Emotions in online rumor diffusion. *EPJ Data Science* 10, 1 (2021). 51.

[33] Nicolas Pröllochs and Stefan Feuerriegel. 2022. Mechanisms of true and false rumor sharing in social media: Collective intelligence or herd behavior? *arXiv* 2207.03020 (2022).

[34] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *EMNLP*.

[35] Eugenia Ha Rim Rho and Melissa Mazmanian. 2020. Political hashtags & the lost art of democratic discourse. In *CHI*.

[36] Tyler W. Rinker. 2019. *sentimentr: Calculate Text Polarity Sentiment*. Buffalo, New York. http://github.com/trinker/sentimentr version 2.7.1.

[37] Claire Robertson, Nicolas Pröllochs, Kaoru Schwarzenegger, Phillip Parnamets, Jay J. Van Bavel, and Stefan Feuerriegel. 2022. Negativity drives online news consumption. https://figshare.com/articles/journal_contribution/Negativity_drives_online_news_consumption_Registered_Report_Stage_1_Protocol_/19657452.

[38] Jesse Shore, Jiye Baek, and Chrysanthos Dellarocas. 2018. Network structure and patterns of information diversity on Twitter. *MIS Quarterly* 42, 3 (2018), 849–972.

[39] Kirill Solovev and Nicolas Pröllochs. 2022. Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. In *WWW*.

[40] Kirill Solovev and Nicolas Pröllochs. 2022. Moral emotions shape the virality of COVID-19 misinformation on social media. In *WWW*.

[41] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M. Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. In *iConference*.

[42] Statista. 2022. Number of monetizable daily active Twitter users (mDAU) in the United States from 1st quarter 2017 to 2nd
       quarter 2022.   https://www.statista.com/statistics/970911/monetizable-daily-active-twitter-users-in-the-united-states/

[43] Stefan Stieglitz and Linh Dang-Xuan. 2013. Emotions and information diffusion in social media: Sentiment of microblogs
       and sharing behavior. *Journal of Management Information Systems* 29, 4 (2013), 217–248.

[44] Twitter. 2021. Introducing Birdwatch, a Community-Based Approach to Misinformation. https://blog.twitter.com/en_
       us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

[45] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018),
       1146–1151.

[46] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence
       for a collective intelligence factor in the performance of human groups. *Science* 330, 6004 (2010), 686–688.

[47] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition,
       manipulation, and detection. *SIGKDD Explorations Newsletter* 21, 2 (2019), 80–90.

Supplementary Materials

A REGRESSION RESULTS WITHOUT OUTLIERS

To assess the robustness of our analysis regarding outliers, we remove tweets with the top 1% highest values for the retweet count. The results are presented Table 1. All results are robust and confirm our previous findings.

Table 1. Regression Results Without Outliers

Dependent Variable: Number of Retweets ($RetweetCount$)			
	Source Tweet	*Fact-Checking Label*	*Misinformation Types*
	Model 1	Model 2	Model 3
Misleading		−0.281***	
		(0.067)	
Factual Error			−0.233***
			(0.033)
Missing Important Context			−0.093***
			(0.032)
Unverified Claim As Fact			−0.273***
			(0.032)
Outdated Information			0.086*
			(0.052)
Satire			0.076
			(0.075)
Manipulated Media			0.479***
			(0.078)
Other			−0.203***
			(0.069)
Delay	−0.018	−0.017	−0.028
	(0.018)	(0.018)	(0.018)
Sentiment	0.076***	0.074***	0.070***
	(0.015)	(0.015)	(0.015)
Followers	0.214***	0.212***	0.213***
	(0.020)	(0.020)	(0.020)
Followees	0.133***	0.134***	0.144***
	(0.016)	(0.016)	(0.016)
Account age	−0.237***	−0.238***	−0.238***
	(0.016)	(0.016)	(0.016)
Verified	1.063***	1.070***	1.114***
	(0.033)	(0.033)	(0.034)
Intercept	7.125***	7.354***	7.335***
	(0.087)	(0.103)	(0.090)
Fixed effects (month-year)	Yes	Yes	Yes
AIC	217 196	218 106	216 960
Observations	15 103	15 103	15 103

Significance levels: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; standard errors in parentheses
Note: Negative binomial regression explains the number of retweets of the fact-checked tweet.
Month-year fixed effects are included.

B SEPARATE REGRESSIONS FOR MISLEADING AND NOT MISLEADING TWEETS

We run separate for regressions for the subsets of misleading and not misleading tweets. The results remain robust (see Table 2).

Table 2. Regression Results for Subsets of Misleading and Not Misleading Tweets

Dependent Variable: Number of Retweets ($RetweetCount$)		
	Subset: *Misleading*	Subset: *Not Misleading*
	Model 1	Model 2
Delay	−0.041**	−0.286***
	(0.019)	(0.073)
Sentiment	0.115***	−0.016
	(0.016)	(0.070)
Followers	0.273***	0.270***
	(0.020)	(0.057)
Followees	0.084***	0.027
	(0.017)	(0.047)
Account age	−0.224***	−0.063
	(0.017)	(0.077)
Verified	0.803***	0.504***
	(0.035)	(0.158)
Intercept	7.751***	8.416***
	(0.098)	(0.234)
Fixed effects (month-year)	Yes	Yes
AIC	209 875	13 287
Observations	14 384	872

Significance levels: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; standard errors in parentheses
Note: Negative binomial regression explains the number of retweets of the fact-checked tweet.
Month-year fixed effects are included.

C VARIANCE INFLATION FACTORS

We calculated variance inflation factors for all explanatory variables in our regression models for RQ1 and RQ2 (Table 3). The VIFs are substantially below the critical threshold of four. This indicates that multicollinearity is not an issue in our analysis.

Table 3. Variance Inflation Factors for Regression Models

	RQ1	RQ2
Misleading	1.019	
Delay	1.004	1.008
Sentiment	1.011	1.016
Followers	1.069	1.072
Followees	1.001	1.003
Account Age	1.155	1.177
Verified	1.193	1.251
Factual Error		1.093
Missing Important Context		1.082
Unverified Claim As Fact		1.129
Outdated Information		1.045
Satire		1.049
Manipulated Media		1.053
Other		1.021

D ANALYSIS WITH USER-SPECIFIC RANDOM EFFECTS

Fact-checks on Birdwatch are performed by many different contributors. To account for this, we include random effects for the individual Birdwatch contributors into our regression model. The regression results are reported in Table 4. All results are robust and confirm our previous findings.

Table 4. Regression Results With User-Specific Random Effects

Dependent Variable: Number of Retweets ($RetweetCount$)			
	Source Tweet	*Fact-Checking Label*	*Misinformation Types*
	Model 1	Model 2	Model 3
Misleading		−0.456***	
		(0.068)	
Factual Error			−0.251***
			(0.034)
Missing Important Context			−0.127***
			(0.033)
Unverified Claim As Fact			−0.300***
			(0.033)
Outdated Information			−0.061
			(0.054)
Satire			0.411***
			(0.077)
Manipulated Media			0.462***
			(0.080)
Other			−0.222***
			(0.071)
Delay		−0.048***	−0.054***
		(0.018)	(0.018)
Sentiment	0.068***	0.068***	0.061***
	(0.014)	(0.014)	(0.014)
Followers	0.271***	0.267***	0.268***
	(0.019)	(0.019)	(0.019)
Followees	0.074***	0.076***	0.086***
	(0.016)	(0.016)	(0.016)
Account age	−0.213***	−0.216***	−0.227***
	(0.017)	(0.017)	(0.017)
Verified	0.774***	0.783***	0.866***
	(0.034)	(0.034)	(0.035)
Intercept	7.922***	8.266***	8.122***
	(0.089)	(0.105)	(0.092)
Fixed effects (month-year)	Yes	Yes	Yes
Random effects (user)	Yes	Yes	Yes
AIC	223 233	223 182	222 893
Observations	15 256	15 256	15 256

Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; standard errors in parentheses
Note: Negative binomial regression explains the number of retweets of the fact-checked tweet.
Month-year fixed effects are included.

E QUADRATIC EFFECTS AND INTERACTION TERMS

As a robustness check, we include quadratic effects and interaction terms between the fact-checking label and the source tweet variables into our regression analysis. The results remain robust and support our findings (see Table 5).

Table 5. Regression Results With Quadratic Effects and Interaction Terms

Dependent Variable: Number of Retweets ($RetweetCount$)		
	Quadratic Effects	Interaction Terms
	Model 1	Model 2
Misleading	−0.479***	−0.659***
	(0.068)	(0.106)
Delay	0.086**	−0.242***
	(0.042)	(0.070)
Delay^2	−0.010***	
	(0.003)	
Sentiment	0.136***	0.020
	(0.016)	(0.066)
Sentiment^2	0.011	
	(0.007)	
Followers	0.548***	0.249***
	(0.045)	(0.054)
Followers^2	−0.029***	
	(0.006)	
Followees	0.115***	0.025
	(0.024)	(0.045)
Followees^2	−0.003*	
	(0.002)	
Account age	−0.416***	−0.058
	(0.022)	(0.073)
Account age^2	−0.341***	
	(0.021)	
Verified	0.739***	0.389***
	(0.035)	(0.150)
Misleading × Delay		0.201***
		(0.072)
Misleading × Sentiment		0.096
		(0.068)
Misleading × Followers		0.024
		(0.058)
Misleading × Followees		0.059
		(0.048)
Misleading × Account age		−0.166**
		(0.075)
Misleading × Verified		0.416***
		(0.155)
Intercept	8.641***	8.436***
	(0.107)	(0.130)
Fixed effects (month-year)	Yes	Yes
AIC	222 906	223 179
Observations	15 256	15 256

Significance levels: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$; standard errors in parentheses
Note: Negative binomial regression explains the number of retweets of the fact-checked tweet.
Month-year fixed effects are included.

F   ALTERNATIVE HANDLING OF MULTIPLE FACT-CHECKS

Our main analysis focuses on the temporally first fact-check after the tweet has been posted. As a robustness check, we evaluate whether our results are robust to alternative handling of multiple fact-checks. We repeated our analysis with the following variants: (i) we determined the fact-checking label via majority vote; (ii) we use Birdwatch's rating mechanism (see [44] for details) to identify the fact-check with which most users agree; (iii) we consider all fact-checks without any filtering.

The regression results are presented in Table 6. In all cases, we find qualitatively identical results that support our previous findings.

Table 6. Regression Results With Alternative Handling of Multiple Fact-Checks

Dependent Variable: Number of Retweets ($RetweetCount$)						
	(i) Majority Vote		*(ii) Highest Agreement*		*(iii) All Fact-Checks*	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Misleading		−0.630***		−0.728***		−0.849***
		(0.068)		(0.057)		(0.043)
Delay		0.018		0.011		−0.004
		(0.017)		(0.018)		(0.013)
Sentiment	0.115***	0.112***	0.112***	0.109***	0.046***	0.064***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.013)	(0.013)
Followees	0.087***	0.092***	0.074***	0.078***	0.021	0.033**
	(0.017)	(0.017)	(0.016)	(0.016)	(0.013)	(0.013)
Followers	0.284***	0.273***	0.271***	0.257***	0.319***	0.316***
	(0.020)	(0.020)	(0.019)	(0.019)	(0.014)	(0.014)
Account age	−0.223***	−0.223***	−0.213***	−0.214***	−0.190***	−0.192***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.014)	(0.014)
Verified	0.789***	0.796***	0.774***	0.763***	0.717***	0.731***
	(0.035)	(0.035)	(0.034)	(0.034)	(0.029)	(0.029)
(Intercept)	7.879***	8.375***	7.922***	8.477***	8.821***	9.559***
	(0.091)	(0.107)	(0.089)	(0.100)	(0.070)	(0.079)
Fixed effects (month-year)	Yes	Yes	Yes	Yes	Yes	Yes
AIC	211 311	211 213	223 233	223 039	317 871	317 409
Observations	14 619	14 619	15 256	15 256	20 218	20 218

Significance levels: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$; standard errors in parentheses

Note: Negative binomial regression explains the number of retweets of the fact-checked tweet.

Month-year fixed effects are included.