Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Community Notes vs. Snoping: How the Crowd Selects Fact-Checking Targets on Social Media

Moritz Pilarski, Kirill Olegovich Solovev, Nicolas Pröllochs

JLU Giessen, Germany
moritz.pilarski@wi.jlug.de, kirill.solovev@wi.jlug.de, nicolas.proellochs@wi.jlug.de

Abstract

Deploying links to professional fact-checking websites (so-called "snoping") is a common misinformation intervention technique that can be used by social media users to refute misleading claims made by others. However, the real-world effect of snoping may be limited as it suffers from low visibility and distrust towards professional fact-checkers. As a remedy, X (formerly known as Twitter) recently launched its community-based fact-checking system "Community Notes" on which fact-checks are carried out by actual X users and directly shown on the fact-checked posts. Yet, an understanding of how fact-checking via Community Notes differs from regular snoping is largely absent. In this study, we empirically analyze differences in how contributors to Community Notes and Snopers select their targets when fact-checking social media posts. For this purpose, we collect and holistically analyze two unique datasets from X: (a) 25,912 community-created fact-checks from X's Community Notes platform, and (b) 52,505 "snopes" that debunk posts via fact-checking replies that link to professional fact-checking websites. We find that Notes contributors and Snopers focus on different targets when fact-checking social media content. For instance, Notes contributors tend to fact-check posts from larger accounts with higher social influence and are relatively less likely to emphasize the accuracy of non-misleading posts. Fact-checking targets of Notes contributors and Snopers rarely overlap; however, those overlapping exhibit a high level of agreement in the fact-checking assessment. Moreover, we demonstrate that Snopers fact-check social media posts at a higher speed. Altogether, our findings imply that different fact-checking approaches – carried out on the same social media platform – can result in vastly different social media posts getting fact-checked. This has important implications for future research on misinformation, which should not rely on a single fact-checking approach when compiling misinformation datasets. From a practical perspective, our findings imply that different fact-checking approaches complement each other and may help social media providers to optimize strategies to combat misinformation on their platforms.

Introduction

Social media has shifted the quality control for content from trained journalists towards regular users (Kim and Dennis 2019). The inevitable lack of oversight makes social media platforms (e. g., X, Facebook) vulnerable to misinformation (Shao et al. 2016; Pew Research Center 2016; Kim and Dennis 2019). If misinformation becomes viral, it can have detrimental consequences on how opinions are formed and on the offline world (Allcott and Gentzkow 2017; Moore, Dahlke, and Hancock 2023; Bakshy, Messing, and Adamic 2015; Oh, Agrawal, and Rao 2013; Gallotti et al. 2020; Geissler et al. 2023; Jakubik et al. 2023; Bär, Pröllochs, and Feuerriegel 2023). In order to identify and eventually curb the spread of misinformation, third-party fact-checking organizations (e. g., snopes.com, politifact.com) regularly fact-check social media rumors (Vosoughi, Roy, and Aral 2018). These fact-checking assessments are supposed to help users to identify misleading content (Shao et al. 2016). Yet, a major challenge is that fact-checks from third-party fact-checking organizations suffer from low visibility as their websites are rarely visited (Robertson, Mourão, and Thorson 2020; Opgenhaffen 2022). Users are oftentimes not aware of these fact-checks when consuming potentially misleading content on social media. Hence, the real-world effect of third-party fact-checks in curbing the spreading of misinformation on social media is limited (Opgenhaffen 2022).

A popular intervention to raise the visibility of third-party fact-checks on social media is conversational fact-checking – also known as "snoping" (Hannak et al. 2014). Here, users independently refute misleading claims in posts by replying with a link to a third-party fact-check debunking the rumor (see Fig. 1a). This approach builds on the premise that linking to a fact-check directly in the place where the misinformation is circulating can make the fact-check more visible to users who would otherwise not actively seek out for fact-checks (Opgenhaffen 2022). While snoping has the potential to make users more aware of third-party fact-checks, its effectiveness may still be limited for multiple reasons: (i) fact-checks in replies to posts may easily be overlooked and are oftentimes simply ignored by users (Hannak et al. 2014); (ii) Snopers have been observed to focus on specific targets (e. g., members of outgroups) and snoping may be a performative rather than deliberative act (e. g., to gain social status; Hannak et al. 2014). (iii) A large proportion of social media users distrust professional fact-checkers (Pew Research Center 2019). Hence, even when users become aware of a snoped post, the impact of the fact-check may be limited due to a lack of trust (Brandtzaeg and Følstad 2017).

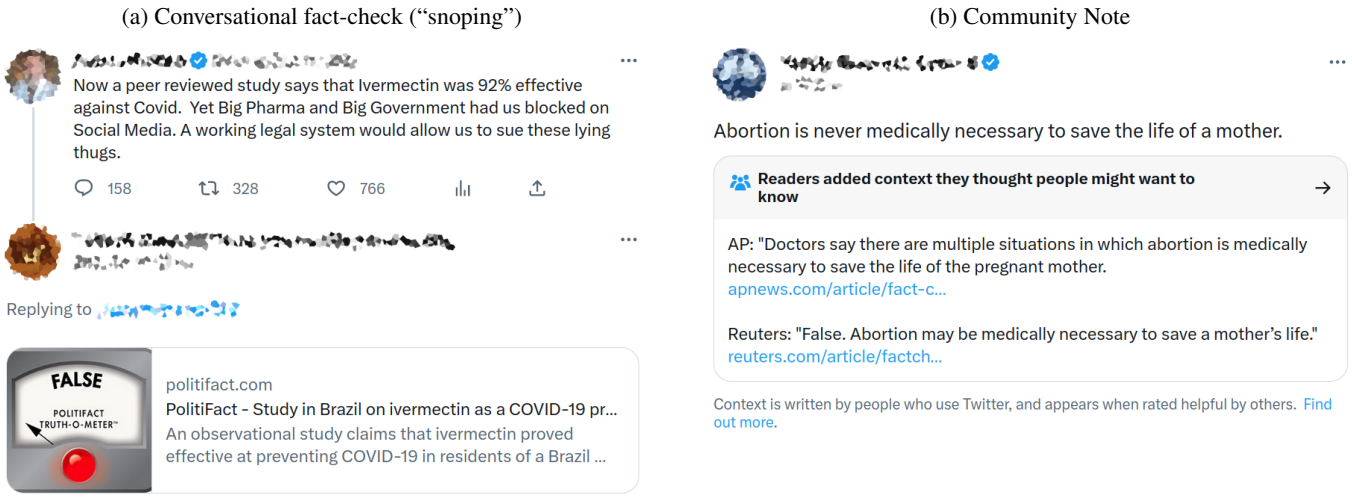

Figure 1: (a) Example of a “snoped” post on X with a reply linking to a fact-check from a third-party fact-checking organization. (b) Example of a Community Note on X.

As a remedy, X recently launched its community-based fact-checking system “Community Notes,” formerly known as “Birdwatch” (Twitter 2021; Pröllochs 2022). This X feature allows users to identify posts they believe are misleading or not misleading and write (textual) notes that provide context to the post. Users can add Community Notes to *any* post they come across on X. Compared to conversational fact-checks, Community Notes promise increased visibility as they can appear directly on the fact-checked posts (see example in Fig. 1b). Furthermore, Community Notes are carried out anonymously and may address the trust problem with professional fact-checkers. Recent research yielded promising results – suggesting that Community Notes can achieve high accuracy in fact-checking social media posts (Wojcik et al. 2022). However, an understanding of how fact-checking on a dedicated community fact-checking system (such as Community Notes) differs from conversational fact-checking (i. e., snoping) is absent. In particular, little is known regarding how (and how fast) Notes contributors and Snopers select their fact-checking targets and the extent to which both features complement each other.

Research Goal: In this work, we empirically analyze how contributors to Community Notes and Snopers select their fact-checking targets on X. Specifically, we address the following research questions:

- **(RQ1)** *How do the fact-checking targets of Notes contributors and Snopers differ in terms of author, content, and engagement characteristics?*
- **(RQ2)** *Do Community Notes reach social media users faster than conversational fact-checks?*
- **(RQ3)** *How do Community Notes and Snoping complement each other?*

Data & Methodology: To address our research questions, we collected two unique datasets from X: (a) 25,912 community-created fact-checks from X’s Community Notes platform; and (b) 52,505 snopes that debunk posts using fact-checking replies linking to professional fact-checking websites. We extract a wide variety of author characteristics (e. g., followers), content characteristics (e. g., topics), and engagement characteristics (e. g., virality) from the fact-checked posts. This allows us to holistically analyze how Snopers and Notes contributors select their targets for fact-checking. Furthermore, we implement a regression analysis to study differences in the fact-checking speed and evaluate the extent to which the fact-checking assessments of Snopers and Notes contributors agree.

Contributions: We find that Notes contributors and Snopers focus on different targets when fact-checking content on X. For instance, Notes contributors tend to fact-check posts from larger accounts with higher social influence and are relatively less likely to emphasize the accuracy of not misleading posts. Fact-checking targets of Snopers and Notes contributors rarely overlap; however, those overlapping exhibit a high level of agreement in the fact-checking assessment. Moreover, we demonstrate that Snopers fact-check posts at a higher speed. In sum, our findings imply that different fact-checking approaches – carried out on the same platform – can result in vastly different posts getting fact-checked. This has important implications for future research on misinformation, which should not rely on a single fact-checking approach when compiling misinformation datasets. From a practical perspective, our findings imply that different fact-checking approaches complement each other and may help social media providers to optimize strategies to combat misinformation on their platforms.

Background

Misinformation on Social Media

Compared to most traditional mass media outlets, social media platforms have lower standards for content moderation. As user-generated content can often be disseminated without undergoing any significant third-party filtering, fact-checking, or editorial scrutiny (Allcott and Gentzkow 2017), so-

cial media is much more vulnerable to the spread of misinformation (Shao et al. 2016; Pew Research Center 2016; Kim and Dennis 2019; Lutz et al. 2023). Several studies suggest that misleading information on social media tends to spread further, faster, deeper, and more widely than not misleading information (Vosoughi, Roy, and Aral 2018; Solovev and Pröllochs 2022; Pröllochs, Bär, and Feuerriegel 2021a; Pröllochs and Feuerriegel 2023). Misinformation is considered a threat to democracy and society, as it can contribute to a wide range of issues including, but not limited to, increased political polarization, threats to public safety, and erosion of trust in institutions (Lazer et al. 2018; Bär, Pröllochs, and Feuerriegel 2023). Given these potential harms, there have been increasing calls to social media providers to take action and address the spread of misinformation on their platforms (Donovan 2020; Feuerriegel et al. 2023).

The most widespread approach to fact-checking is to have professional fact-checkers verify claims. While this expert fact-checking approach has been shown to be effective in numerous studies (a comprehensive review can be found in Walter et al. 2020), it still has several critical limitations. Due to the time-consuming nature of thoroughly investigating claims (often many hours or even days for a single claim; Guo, Schlichtkrull, and Vlachos 2022) and the limited number of fact-checkers available, many misleading stories never get tagged. Limited resources often force fact-checkers to prioritize content that is blatantly false or deliberately misleading over content that is more nuanced or complex (Pennycook and Rand 2019). As a result, they may overlook biased or misleading coverage of events, incomplete information, or the use of misleading statistics. Furthermore, many U.S. citizens distrust professional fact-checkers. According to a study conducted by the Pew Research Center (2019), a majority of Republican partisans (70 %) and half of all U.S. adults believe that fact-checkers are biased and that their corrections cannot be trusted. Also, professional fact-checks oftentimes have very limited reach. Besides some collaborations with social media platform providers on specific topics, fact-checking organizations mainly communicate the results of their fact-checks through their websites. According to a study (Robertson, Mourão, and Thorson 2020), in 2017, over half of all U.S. adults had never visited any fact-checking website.

Conversational Fact-Checking ("Snoping")

A popular strategy employed by social media users to combat misleading statements is to link professional fact-checking articles from third-party fact-checking organizations (e. g., snopes.com, politifact.com) in their replies to the original message. This conversational approach to fact-checking – commonly referred to as "snoping" (e. g., Hannak et al. 2014; Friggeri et al. 2014) – (partially) addresses the issue of the limited reach of third-party fact-checking websites by increasing the visibility of articles in the contexts of the respective fact-checked statements. Researchers have utilized data on conversational fact-checks to investigate various phenomena surrounding the spread of misinformation on social media platforms like X/Twitter (e. g., Hannak et al. 2014; Margolin, Hannak, and Weber 2018; Vosoughi, Roy, and Aral 2018; Mosleh et al. 2022), Facebook (e. g., Friggeri et al. 2014), or Reddit (e. g., Bond and Garrett 2023). For instance, Friggeri et al. (2014) study the effect of snoping on the propagation of rumors on Facebook. The authors find that snopes on individual reshares of rumors increase the probability of those reshares being deleted.

Only a few works have studied how users engage in snoping and their motifs. For example, Hannak et al. (2014) and Margolin, Hannak, and Weber (2018) focus on the effect that social relations between Snopers and Snopees have on the recognition of corrections on X/Twitter. They find that while only a small share of all snopes is made by friends (i. e., mutually following users), those are especially likely to get the Snopee's attention. They attribute this to the circumstance that individuals feel a greater obligation to respond scientifically and be more open to facts that challenge their original positions when interacting with their friends. Another work has studied the role of linguistic and engagement features (Ma et al. 2023). The authors find that misinformative posts expressing negative emotion and impoliteness are more likely to receive countering replies. Additionally, they observe that countered post tend to have a higher proportion of reply engagement compared to like, reshare, and quote engagement. Furthermore, research has analyzed the network of follower-relations among Snopers and Snopees. Hannak et al. (2014) find that the network exhibits strong polarization between two large densely connected communities that roughly reflect the political camps forming along U.S. party lines. At the same time, most of the snope-relations are spanning between those communities, which suggests that snoping is commonly directed outwards as criticism of individuals that Snopers otherwise are not interested in. Hence, snoping, in many cases, should be seen as a performative rather than a deliberative act, in which Snopers are displaying their political affiliation (Hannak et al. 2014). Since a large share of people in the U.S. dislikes and distrusts users from the opposing party – a phenomenon typically subsumed under the term "affective polarization" (Iyengar et al. 2019) – many snopes may go unheard.

Community-Based Fact-Checking Systems

As snoping essentially passes on judgments made by professional fact-checkers, it faces many of the same challenges and limitations as professional fact-checking. For example, users cannot snope claims that have not yet been verified by professional fact-checkers. Furthermore, snoping suffers from low visibility and distrust towards fact-checking organizations. A possible remedy to those problems is crowdsourcing the fact-checking process. This would have the advantage that an abundance of users willing to participate in content moderation would grant the effort almost unlimited resources (Allen et al. 2021; Pennycook and Rand 2019). Additionally, trust issues with professional fact-checkers could be mitigated. Despite those promises, there are, however, reasons to also be concerned about crowd judgments. For example, unlike professional fact-checkers, the crowd typically lacks specific training and systematic practices for evaluating the veracity of stories (Graves 2017).

In recent years, a growing body of research (Micallef

et al. 2020; Bhuiyan et al. 2020; Pennycook and Rand 2019; Epstein, Pennycook, and Rand 2020; Allen et al. 2020, 2021; Godel et al. 2021; Drolsbach and Pröllochs 2023a,b; Pröllochs 2022) has focused on community-based fact-checking systems that leverage the "wisdom of crowds" (Surowiecki 2005). These systems rely on the principle that, while individual users' fact-checks may be prone to bias or inaccuracies, high levels of accuracy can be attained through the collective judgments of politically diverse groups (a summary of related literature can be found in Martel et al. 2023). For example, Allen et al. (2021) compare the correlation between the average ratings of differently sized crowds of laypeople and three professional fact-checkers to the correlation between the individual fact-checkers' ratings. Whereas the laypeople were merely presented the headline and lede of articles, the fact-checkers were thoroughly researching them. As they keep increasing the crowd size, they stop finding a significant difference between the correlations of ratings at a crowd size of about eight people (similar results in Bhuiyan et al. 2020; Resnick et al. 2021).

Informed by those promising research findings, X recently introduced its community-based fact-checking system "Community Notes" (formerly known as "Birdwatch"). This new feature provides users with the ability to fact-check any post they come across by creating so-called Community Notes. Community Notes consist of a categorization of whether a post might or might not be misleading and an open text field (max. 280 characters) that allows contributors to explain their decision and include links to relevant sources. After a note is created, other users can rate its helpfulness and, if the note reaches a certain level of helpfulness, it is displayed prominently beneath the original post (see Fig. 1b). Until recently, the Community Notes feature was in pilot phase and only available to registered participants in the U.S. The pilot phase started on January 25, 2021, and ended on October 6, 2022. As of December 11, 2022, registration for Community Notes is open to users worldwide, and helpful notes are visible to everyone on X.

Given the recency of the platform, research on Community Notes is scant. Early works suggest that politically motivated reasoning might pose challenges in community-based fact-checking (Allen, Martel, and Rand 2022; Pröllochs 2022). For instance, Note contributors tend to focus their fact-checking efforts on content posted by individuals with whom they hold opposing political views (Allen, Martel, and Rand 2022). Notwithstanding, fact-checks on Community Notes have been found to be perceived as informative and helpful by the vast majority of users (Pröllochs 2022). Saeed et al. (2022) additionally highlight the important role played by Notes contributors in refuting false claims that have already been fact-checked by professional journalists but continue to circulate on X/Twitter nonetheless.

Furthermore, recent studies indicate that community fact-checked misleading posts are less viral than not misleading posts (Drolsbach and Pröllochs 2023a; Chuai et al. 2023) and that displaying notes may reduce users' propensity to share misleading posts (Wojcik et al. 2022). In a user study conducted directly on X/Twitter (Wojcik et al. 2022), users were randomly assigned to either view post annotations or no annotations. The results showed that those who were exposed to annotations on posts were 25 % to 34 % less likely to like or reshare them compared to the control group.

Data Sources

Dataset I: Community Notes

Fact-Checks: Community Notes is a community-based fact-checking system that allows registered users to fact-check statements made on X. Users can fact-check *any* post they come across on X – directly when browsing the platform. We obtained the data on Community Notes from the complete database dumps that are published by X on a weekly basis.[1] From this dataset, we used the notes' publication dates, veracity judgments (i. e., whether the post is categorized as misleading or not misleading), as well as the free-text explanations (max. 280 characters) that are used by contributors to explain their judgments. In our study, we consider all fact-checks that were created during Community Note's pilot phase in the U. S., which started on January 26, 2021 and ended on October 5, 2022.

Fact-Checked Posts: We used X's post lookup API endpoint to collect all fact-checked posts, i. e., posts that have received a Community Note. Furthermore, we collected various information about the authors of the fact-checked posts (e. g., number of followers). We excluded all posts that were not classified as written in English by X's language detection algorithm as well as all posts by the user `@CommunityNotes` since those were officially recommended for testing purposes. Notably, multiple contributors can write Community Notes for the same post. Therefore, the data sometimes includes multiple fact-checks for the same post. In our data, 22.0 % of the fact-checked posts received more than one Community Note. Our final dataset encompasses a total of 25,912 Community Notes, contributed by 4,288 unique (pseudonymous) contributors, covering a total of 18,805 distinct posts (Dataset I). All of our data was collected in February 2023. Any content that was deleted before that time is not included in our analysis.

We performed basic text preprocessing on the fact-checked posts by removing user-mentions (`@screen-name`) from the beginnings of the posts' texts[2], removing URLs, and parsing HTML-characters (e. g., `&` → &).

Dataset II: Conversational Fact-Checks ("Snopes")

Fact-Checks: Our approach to collecting conversational fact-checks, i. e., snopes, was guided by best practices from earlier research (Vosoughi, Roy, and Aral 2018). We focused on three reputable fact-checking websites that thoroughly investigate social media rumors, namely, snopes.com, politifact.com, and truthorfiction.com. We scraped all fact-checks and their corresponding veracity judgments published on any of these websites (a total of 44,086 articles). The fact-checking organizations have different ways of labeling the veracity of a story. For example, politifact.com

[1] https://twitter.com/i/birdwatch/download-data

[2] User mentions at the beginning of a post typically refer to the structures of the reply-trees in which the posts are embedded.

articles are given a "Pants on Fire" rating for misleading stories, whereas snopes.com assigns a "false" label. Analogous to earlier work (Vosoughi, Roy, and Aral 2018; Solovev and Pröllochs 2022), we normalized the veracity labels across the different sites by mapping them to a score of 1 to 5. All stories with a score of 1 or 2 were categorized as "misleading," whereas stories with a score of 4 or 5 were categorized as "not misleading" (e. g., "Pants on Fire!" $\rightarrow$ *misleading*).[3]

Fact-Checked Posts: We used X's full-archive search API endpoint to collect all reply posts featuring a link to any of the previously scraped fact-checking articles. To ensure comparability with Dataset I, we considered only replies that were posted between January 25, 2021 and October 6, 2022 (i. e., during Community Note's pilot phase). As mentioned earlier, all the data we collected is from late February 2023. Any content deleted prior to that date is not accounted for in our analysis. Of those posts, we excluded all that were not classified as written in English by X's language detection algorithm. To ensure that replies featuring links to fact-checking articles are actual fact-checks of statements made in their respective parent posts, we compared the semantic contents of the fact-checking articles' assessed claims with the texts of the posts to which they were given as replies. Given that 18 % of the fact-checked posts have images attached, we first employed optical character recognition to extract the textual content from those images[4]. After applying the same preprocessing steps as before, we generated document embeddings for all fact-checked posts' texts (including the ones retrieved from the images) and all fact-checking articles' assessed claims using the pre-trained TwHIN-BERT language model (Zhang et al. 2022). Finally, we calculated cosine-similarities between the embedding-vectors of all observed pairs of text and discarded those with a similarity-value below 0.75. This resulted in a final dataset comprising 52,505 conversational fact-checks contributed by 34,188 unique authors, covering a total of 45,368 unique posts (Dataset II).

User Study: We evaluated the performance of our method for excluding unrelated pairs of snopes and posts with a user study. To this end, we employed two trained research assistants (hourly wage: ≈$14) that were tasked with rating the semantic similarity of posts with the corresponding fact-checked claim. For this, participants had to answer the question "How related is this post to the fact-checked claim?" on a 5-point Likert scale ranging from "Completely Unrelated" to "Completely Related." We observed a relatively high Kendall's coefficient of concordance of $W = 0.738$ ($p < 0.001$), and 81.4 % of the pairs classified as related by our model were adjudged to be at least "somewhat related" by the human raters. This implies that our method identifies snoped posts on X with high accuracy.

[3]For the sake of simplicity and comparability, we omitted stories with a score of 3, i. e., stories with a "mixed" veracity (10.5 % of all conversational fact-checks). Including those stories yields qualitatively identical results in our later analysis.

[4]We preprocessed the images with ImageMagick (2023), performed optical character recognition with the Tesseract engine (Smith 2007), and performed several postprocessing steps based on DBSCAN clustering to identify coherent lines of text.

Empirical Analysis

Target Selection (RQ1)

To answer **RQ1**, we analyzed how the fact-checking targets of Notes contributors and Snopers differ in terms of their account, content, and engagement characteristics.

Account Characteristics: Fig. 2 plots the kernel density estimates as well as mean and quartile values for the distributions of the fact-checked users' numbers of followers. We found that Notes contributors tend to annotate posts authored by users with much higher popularity and reach. The mean number of followers for notes is almost five times higher than it is for snopes ($\text{mean}_{\text{notes}}$ = 3,385,810; $\text{mean}_{\text{snopes}}$ = 679,684; [KS-test: $D = 0.409$; $p < 0.001$]).

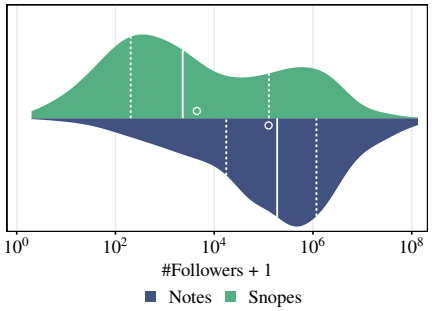

Figure 2: Split violin plot comparing the distributions of follower counts among authors of fact-checked posts. Shown are kernel density estimates (colored areas), mean values (white circles), and quartile values (white lines).

Fig. 3 presents the distribution of fact-checks across posts authored by users whose account authenticity has been verified by X, potentially indicating greater perceived credibility in their statements. Notably, a significantly larger proportion of Notes contributors (62.4 %) compared to Snopers (26.8 %) focus their fact-checking efforts on posts by verified users (χ^2-test: X^2 = 9,269; $p < 0.001$).

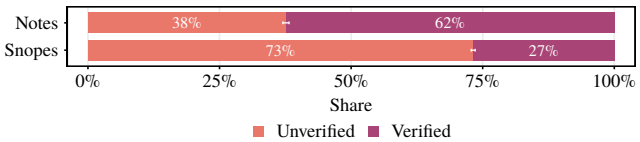

Figure 3: Proportions of fact-checking targets that are verified users and their 95 % confidence intervals (error bars).

We further analyzed additional account characteristics such as the users' followee counts and account ages. Here, we found comparably smaller differences. On average, Notes contributors are slightly more likely to focus on users with higher followee counts and older user accounts (see Supplementary Materials for details).

Content Characteristics: We determined the number of word tokens (#Words) and calculated sentiment scores based on the NRC Word-Emotion Association Lexicon (EmoLex; Mohammad and Turney 2010) for all fact-checked posts.

The number of word tokens (i. e., the post length) potentially indicates the extent of detail within the fact-checked claims, while the expressed sentiment might affect the readers' emotional reactions towards the statements. For our sentiment analysis, we used the default implementation of the `sentimentr` R package (with the built-in NRC lexicon) that also accounts for negations and valence shifters (see Rinker 2019 for details), analogous to previous research (e. g., Robertson et al. 2023; Pröllochs, Bär, and Feuerriegel 2021b). Fig. 4 visualizes the corresponding distributions. There is a slightly higher share of notes than of snopes on relatively short posts ($\text{mean}_{\text{notes}}$ = 26.7; $\text{mean}_{\text{snopes}}$ = 28.4 [KS-test: $D = 0.105$; $p < 0.001$]). However, there is no significant difference in the mean sentiment scores ($\text{mean}_{\text{notes}}$ = 0.004; $\text{mean}_{\text{snopes}}$ = 0.006; [t-test: $t = -1.210$, $p = 0.228$]). Overall, the observed differences regarding the length and sentiment of the fact-checked posts are rather small.[5]

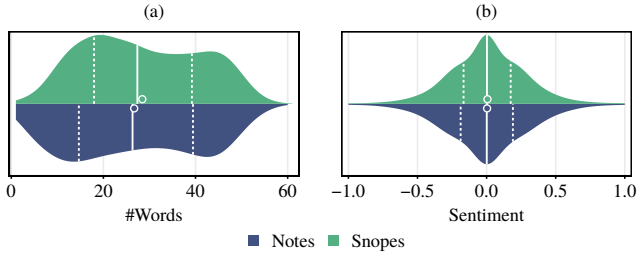

Figure 4: Split violin plot comparing the distributions of the (a) length and (b) sentiment of the fact-checked posts. Shown are kernel density estimates (colored areas), mean values (white circles), and quartile values (white lines).

Next, we conducted topic modeling to explore potential differences in the topics that Notes contributors and Snopers focus on. Our rationale was that different topics may imply distinct groups of authors and target audiences, and, thus, may draw different types of fact-checkers.To this end, we employed supervised machine learning to categorize the fact-checked posts from our dataset into eight predefined topics: *Business; Disasters; Entertainment; Health; Politics; Science; War; Other*. These topics have been identified based on a manual assessment of the fact-checked posts in our dataset and the selection of topics in previous works (e. g., Vosoughi, Roy, and Aral 2018). To create training data, we employed a trained research assistant to assign topic labels (multiple selection possible) to a random subset of 7,500 posts. We then used this labeled data to train a deep neural network classifier that predicts whether a post belongs to each topic. The input data for the training of the classifier was a vector representation of the labeled posts and the topic labels. To create vector representations of posts, we used the pre-trained TwHIN-BERT language model (Zhang et al. 2022). In our deep neural network classifier, we treated the task of predicting topic labels for (vector representations of) posts as a multi-label problem considering that one post may belong to multiple topics. All hyperparameters were tuned using 10-fold cross-validation. Our classifier achieved a relatively high micro-averaged F_1 score of 0.75 and an accuracy of 0.93 on out-of-sample posts.

The shares of fact-checks on posts per topic are displayed in Fig. 5. Note that since the fact-checked posts can have multiple topic labels, those shares do not sum up to 100 %. There are significant differences in the distributions of the fact-checked posts' topics between Community Notes and snopes (χ^2-test: $X^2 = 2{,}164$; $p < 0.001$). In particular, Community Notes are relatively more prevalent on posts about *Disasters*, *Entertainment*, *Health*, and *Other* topics. In contrast, snopes are relatively more common on posts about *Business*, *Politics*, and *Science*.

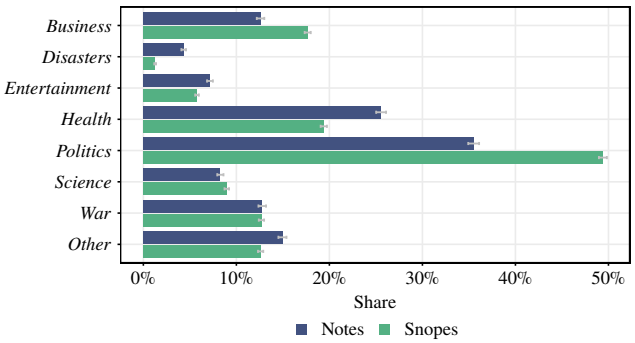

Figure 5: Proportions of fact-checks on posts with different topics (colored bars) and their 95 % confidence intervals (gray error bars).

Furthermore, we examined differences in veracity judgments between Snopers and Notes contributors (see Fig. 6). Our analysis revealed that a majority of Snopers and Notes contributors adjudge the claims made in their targeted posts as misleading. Specifically, for notes, the proportion of misleading verdicts (86.8 %) is 6.6 times higher than that of not misleading verdicts (13.2 %). Snopers classify 3.2 times more of the fact-checked posts as misleading (76.2 %) than as not misleading (23.8 %). Overall, snopes exhibit a relatively higher share of not misleading verdicts compared to Community Notes [χ^2-test: $X^2 = 1{,}215$; $p < 0.001$].

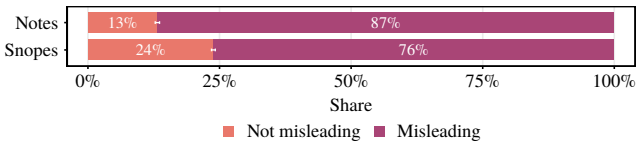

Figure 6: Proportions of fact-checks' verdicts (colored bars) and their 95 % confidence intervals (white error bars).

We also examined whether fact-checkers prefer fact-checking conversation starting posts or reply posts (see Fig. 7). Understanding this difference is important as conversation starting posts usually have a higher visibility than replies (Hannak et al. 2014). Our analysis revealed that the

[5]We additionally analyzed discrete emotions (e. g., anger, fear) in the Supplementary Materials. Again, the observed differences between Notes contributors and Snopers are small.

proportion of Notes addressing conversation starting posts (86 %) is nearly twice as high as the corresponding proportion for Snopes (44 %; χ^2-test: X^2 = 12,223; $p < 0.001$).

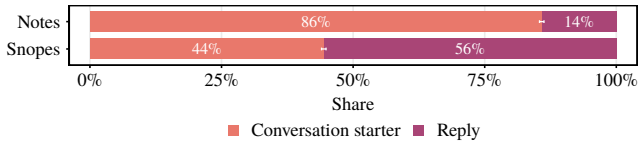

Figure 7: Proportions of fact-checks on posts that are either conversation starters or replies (colored bars) and their 95 % confidence intervals (white error bars).

Engagement Characteristics: Fig. 8 depicts the distributions of the fact-checked posts' engagement metrics for Community Notes and snopes. We observed much higher values for Community Notes across all dimensions. Notes contributors, on average, fact-check posts with almost five times more likes ($\text{mean}_{\text{notes}}$ = 28,519; $\text{mean}_{\text{snopes}}$ = 6,089; [KS-test: D = 0.415; $p < 0.001$]), nearly four times more reshares ($\text{mean}_{\text{notes}}$ = 4,816; $\text{mean}_{\text{snopes}}$ = 1,265; [KS-test: D = 0.425; $p < 0.001$]), roughly five times more replies ($\text{mean}_{\text{notes}}$ = 3,157; $\text{mean}_{\text{snopes}}$ = 638; [KS-test: D = 0.397; $p < 0.001$]), and close to 8 times more quotes than Snopers ($\text{mean}_{\text{notes}}$ = 1,504; $\text{mean}_{\text{snopes}}$ = 200; [KS-test: D = 0.441; $p < 0.001$]). These results suggest that Notes contributors are more likely to fact-check highly "viral" posts, whereas Snopers tend to focus on more "regular" posts.

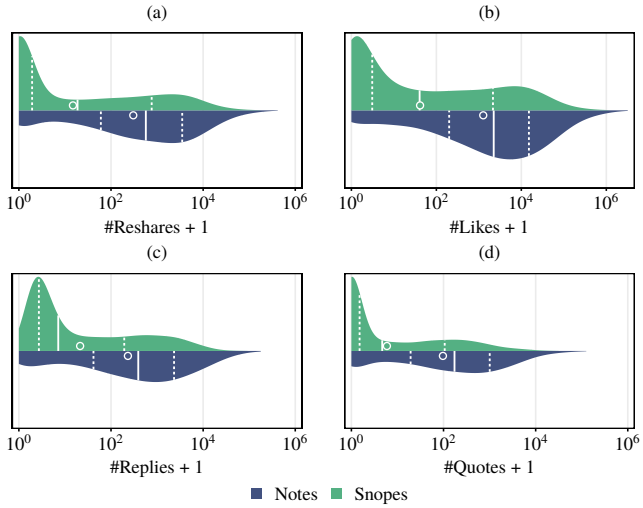

Figure 8: Split violin plot comparing the distributions of the fact-checked posts' different engagement metrics, namely, (a) the number of reshares, (b) the number of likes, (c) the number of replies, (d) the number of quotes. Shown are kernel density estimates (colored areas), mean values (white circles), and quartile values (white lines).

Fact-Checking Speed (RQ2)

To answer **RQ2**, we analyzed the lengths of the timespans between the posting dates of the original posts and the fact-checks. For this purpose, we first compared summary statistics. Subsequently, we implemented an explanatory regression model to analyze which post features are linked to a higher fact-checking speed.

Summary Statistics: Fig. 9 shows the distributions of fact-check delays (in days). The lengths of the timespans between the publication dates of fact-checks and their respective parent posts tend to be longer for Community Notes than for snopes. It takes Notes contributors, on average, more than twice as long as Snopers to publish their fact-checks ($\text{mean}_{\text{notes}}$ = 10.8 days; $\text{mean}_{\text{snopes}}$ = 4.9 days; [KS-test: D = 0.241; $p < 0.001$]).

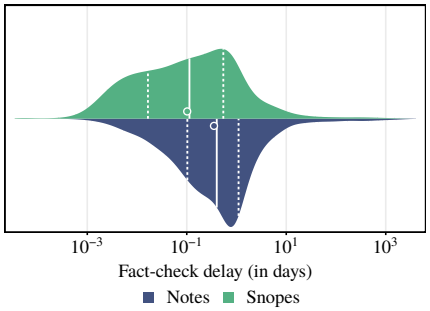

Figure 9: Split violin plot comparing the distributions of the fact-checking delays, i. e., the lengths of the timespans between the posting dates of the original posts and the fact-checks (in days). Shown are kernel density estimates (colored areas), mean values (white circles), and quartile values (white lines).

Regression Analysis: To further examine the differences in fact-checking delays, we performed an explanatory regression analysis. The dependent variable is the fact-checking delay (in days), i. e., the timespans between the posting dates of the original posts and the fact-checks. The explanatory variables comprise the author, content, and engagement[6] characteristics of the fact-checked post that were presented in the previous analyses. We also included monthly fixed effects to control for differences in the fact-checking date. In our model, the fact-checking delays are first log-transformed and then modeled via a normal distribution. This modeling approach is consistent with previous research assuming a log-normal distribution of response times (e. g., Pröllochs, Bär, and Feuerriegel 2021a,b) and allowed us to estimate the model using ordinary least squares (OLS). We z-standardized all continuous explanatory variables in order to facilitate interpretability.

As detailed in the previous section, significant differences exist across nearly all the examined attributes' distributions between Community Notes and Snopes. In order to reduce the possibility of confounding biases in regression outcomes, we implemented propensity score matching. The propensity scores were calculated using logistic regression, followed by nearest neighbor matching with calipers

[6]The engagement characteristics (e. g., #Reshares, #Likes) are highly correlated. To circumvent possible multicollinearity issues, we restricted our model to #Reshares and #Replies.

set at 0.1 standard deviations of the propensity scores' distribution. This process culminated in a dataset encompassing 19,545 observations from each group. The outcome was a substantial reduction in standardized mean differences across all variables to levels below 0.05, with an average relative reduction of those differences by 67.40 %. In the following, we present the regression results for the propensity-matched dataset. The results of the regression conducted on the unmatched (i. e., full) dataset can be found in the Supplementary Materials (the results are qualitatively identical).

Coefficient Estimates: Fig. 10 illustrates the regression coefficients and their corresponding 95 % confidence intervals. Looking at only the Community Notes dataset (depicted as the blue model in Fig. 10), we found that posts authored by individuals with higher social influence undergo fact-checking at an accelerated pace. A one standard deviation increase in the number of followers corresponds to an $e^{-0.036} \approx 3.50\,\%$ reduction in the time taken for fact-checking (coef. = −0.036, $p = 0.042$). Additionally, we observed that posts originating from older accounts receive faster fact-checking. A one standard deviation increase in account age correlates with a 6.34 % decrease in fact-checking delays (coef. = −0.066, $p < 0.001$). When considering post characteristics, we found that longer posts (coef. = 0.100, $p < 0.001$) and those with a positive sentiment (coef. = 0.066, $p < 0.001$) tend to undergo slower fact-checking. Conversation starters experience a substantial 85.03 % increase in fact-checking time (coef. = 0.615, $p < 0.001$), whereas misleading posts exhibit a 10.63 % decrease in fact-checking time (coef. = −0.112, $p = 0.007$). In terms of topics, posts discussing *Science* face 16.84 % slower fact-checking times (coef. = 0.156, $p = 0.005$). Conversely, posts related to *War* and *Business* experience 9.63 % (coef. = −0.101, $p = 0.038$) and 15.68 % (coef. = −0.171, $p < 0.001$) faster fact-checking times, respectively. posts involving *Politics* are subjected to the fastest fact-checking, displaying an estimated 22.96 % reduction in the time before fact-checking (coef. = −0.261, $p < 0.001$). Finally, we examined several engagement metrics, as indicated by the number of reshares and replies. A one standard deviation increase in the number of reshares leads to an 8.66 % increase in fact-checking time (coef. = 0.083, $p < 0.001$), while the coefficient associated with the reply count does not achieve statistical significance within common thresholds.

Next, we compared the estimates with the regression results for Snopers (see the green model in Fig. 10). We again observed that posts from individuals with higher social influence tend to undergo faster fact-checking. In particular, a one standard deviation increase in the number of followers corresponds to approximately a 15.39 % reduction in fact-checking time (coef. = −0.167, $p < 0.001$). Posts from verified users are estimated to undergo 20.92 % faster fact-checking (coef. = −0.235, $p < 0.001$). A one standard deviation increase in the number of followees has a slight positive effect and is associated with a 5.46 % increase in fact-checking time (coef. = 0.053, $p = 0.001$). Similar to Notes contributors, we find that longer posts (coef. = 0.059, $p < 0.001$) and posts with positive sentiment (coef. = 0.035, $p = 0.030$) tend to undergo fact-checking at a slower pace. Conversation starting posts exhibit a significantly higher fact-checking delay, with a remarkable 289.61 % increase in delays compared to replies (coef. = 1.360, $p < 0.001$). Different from the Notes contributors model, misleading posts tend to receive 57.87 % slower fact-checking by Snopers (coef. = 0.475, $p < 0.001$). Examining topic effects reveals 56.76 % slower (coef. = 0.450, $p < 0.001$) fact-checks for posts related to *Disasters*. Similar to the model for Note contributors, *Business*-related posts correspond to a 10.77 % faster fact-checking (coef. = −0.114, $p = 0.020$). The shortest delay between a post and its corresponding fact-check for Snopers is associated with *War* and *Politics*, with posts on these topics corresponding to a 22.56 % (coef. = −0.256, $p < 0.001$) and 17.34 % (coef. = −0.190, $p < 0.001$) reduction in time before fact-checking, respectively. The engagement metrics show similar effects as for the Notes contributors. A one standard deviation increase in the number of reshares is associated with a 39.76 % longer delay until fact-checking (coef. = 0.335, $p < 0.001$), and a one standard deviation increase in the number of replies corresponds to a 34.66 % longer fact-checking delay (coef. = 0.298, $p < 0.001$).

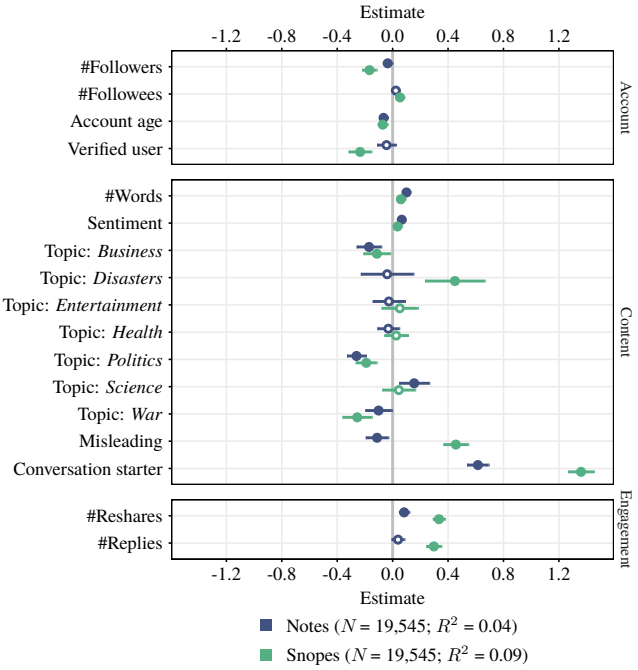

Figure 10: Coefficient estimates (circles) and their 95 % confidence intervals (bars) based on the propensity-matched datasets. The dependent variable is the fact-checking delay, i. e., the lengths of the timespans between the posting dates of the original posts and the fact-checks. Intercepts and monthly fixed effects are included. Coefficient estimates that are statistically significant ($p < 0.05$) are shown with filled circles.

In summary, we observed a clear similarity in the way both groups tend to fact-check high-status individuals on X relatively faster, and stark differences across different topics. While political content is fact-checked quickly by both

groups, Snopers exhibit a relatively longer delay in verifying the accuracy of posts related to *Disasters*. In contrast, Notes contributors take more time assessing the veracity of posts concerning *Science*. Interestingly, we observed different signs for the coefficients of the veracity label in the two models. Specifically, we found that posts considered as misleading are fact-checked slightly faster by Notes contributors, while they are fact-checked slower by Snopers in comparison to posts considered as not misleading. Out of all explanatory variables in our models, the conversation starting status is associated with the highest difference in fact-checking delays for both, notes and snopes. A plausible explanation for this finding is that old replies have lower visibility than old conversation starting posts and, thus, are less likely to get fact-checked at a later date.

Robustness Checks: We conducted a wide variety of checks to validate the robustness of our analysis. First, we carried out standard diagnostic tests to validate the fulfillment of key OLS assumptions. This encompassed a range of checks, including confirming that all variance inflation factors were well below the critical threshold of 4 and verifying the normality of the residuals. Second, there is a possibility of a bidirectional relationship where engagement not only determines the delay in fact-checking but also vice versa. To alleviate such endogeneity concerns, we repeated our analysis and excluded the engagement metrics from the regressions models. The results are qualitatively identical with no significant alterations in the magnitudes, signs, or significance values of the other coefficients.

Overlap and Agreement (RQ3)

Next, we explored the extent to which the two fact-checking approaches complement each other. For this purpose, we examined the overlap and the within-/between-group agreement of the fact-checking assessments of Notes contributors and Snopers (**RQ3**).

Overlap: To analyze the overlap between contributors to Community Notes and Snopers, we mapped the post IDs of the fact-checked posts in Dataset I to those in Dataset II. We found that 28.7 % (18,224) of all fact-checked posts are exclusively fact-checked by Notes contributors, 70.4 % (44,787) are exclusively fact-checked by Snopers, and merely 0.9 % (581) are fact-checked by both groups. Overall, this implies that the fact-checking targets of Snopers and Notes contributors rarely overlap.

Within-Group Agreement: Community fact-checkers sometimes create multiple fact-checks for the same post. This allows us to study the within-group agreement of the fact-checking verdicts (i. e., whether the post is categorized as misleading or not misleading). Among the posts with any Community Notes, 22.0 % (4,131) have multiple notes associated with them. In contrast, among the posts with snopes, only 6.3 % (2,854) have multiple snopes associated with them. This suggests that Notes contributors tend to concentrate their efforts on a narrower set of targets, while Snopers exhibit a broader coverage. Fig. 11a shows the distributions of the shares of fact-checks agreeing with the respective majority verdicts. Among the posts with Community Notes, the average share of agreement with the majority verdict is 83.1 %. On the other hand, posts with snopes show a higher average agreement of 97.9 % (KS-test: $D = 0.355$; $p < 0.001$). This discrepancy may be attributed to the fact that snopes rely on verdicts from professional fact-checking organizations, which typically exhibit a very high level of agreement (Vosoughi, Roy, and Aral 2018).

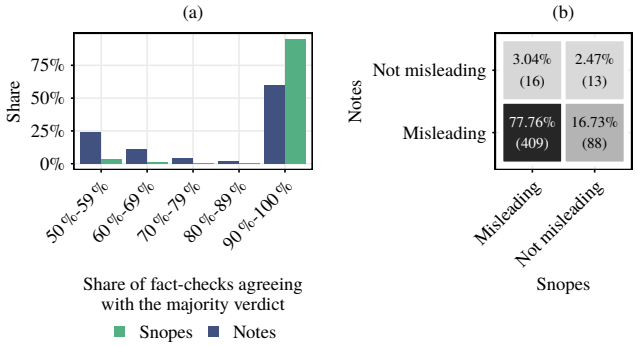

Figure 11: (a) Shares of fact-checks per group that agree with the majority verdict for posts with multiple fact-checks (i. e., within-group agreement). (b) Agreement of majority verdicts between Notes contributors and Snopers for posts that have been fact-checked by both groups (i. e., between-group agreement).

Between-Group Agreement: Fig. 11b shows the agreement between the majority verdicts of Notes contributors and Snopers for all posts that have been fact-checked by both groups. The overall agreement share between Notes contributors and Snopers is high at 80.2 %. Notably, we observed a much higher between-group agreement for posts considered misleading compared to posts considered not misleading. However, the findings for the latter should be interpreted with caution due to the limited number of posts (117) with overlapping fact-checks and a not misleading majority verdict by either Notes contributors or Snopers.

Discussion

Relevance: There are widespread concerns that misinformation on social media is damaging societies and democratic institutions (Calo et al. 2021; Grinberg et al. 2019; Lazer et al. 2018; Donovan 2020; Feuerriegel et al. 2023). Hence, policy initiatives around the world urge social media platforms to limit its spread. A crucial prerequisite to curb the spread of misinformation on social media is its accurate identification (Pennycook and Rand 2019). Community-based fact-checking has the potential to partially overcome the drawbacks of alternative approaches to fact-checking, e. g., in terms of speed, volume, and trust (Allen et al. 2020). While earlier studies suggest that crowds might be able to accurately assess the veracity of social media content (Bhuiyan et al. 2020; Epstein, Pennycook, and Rand 2020; Pennycook and Rand 2019), an understanding of how community fact-checkers select their targets for fact-checking is still largely absent. Here, we contribute to research into misinformation and fact-checking by characterizing how con-

tributors to Community Notes and Snopers select their targets when fact-checking posts on the social media platform X.

Summary of Findings: Our key findings are as follows: (i) The targets of Notes contributors and Snopers significantly differ in terms of their author, content, and engagement characteristics. For instance, Notes contributors tend to fact-check posts from larger accounts with higher social influence and are relatively less likely to endorse/emphasize the accuracy of not misleading posts (*RQ1*). (ii) Compared to Notes contributors, Snopers fact-check posts at a higher speed (*RQ2*). (iii) The fact-checking targets of Notes contributors and Snopers rarely overlap; however, the overlapping fact-checks exhibit a high level of agreement in their assessments (RQ3).

Implications: Our analysis implies that Notes contributors and Snopers focus on different targets when fact-checking X content. A possible reason is that these user groups have different motivations and goals when fact-checking posts. In previous research, Snopers have already been observed to frequently focus on specific targets such as, for example, outgroup members (e. g., to gain social status). As such, their motivation to fact-check social media posts may be – at least partially – performative rather than deliberative (Hannak et al. 2014). Furthermore, both approaches vary in terms of the effort required to fact-check posts. Writing a full-fledged community fact-check arguably requires more time and expertise. Hence, snoping may draw groups of fact-checkers that are less willing to invest the necessary efforts to write a full-fledged community fact-check and/or that select posts that are faster (or easier) to fact-check. In line with this notion, we also find that Snopers fact-check posts at a higher speed. In sum, our findings imply that different fact-checking approaches – carried out on the same social media platform – can result in vastly different social media posts getting fact-checked.

These findings have important implications for future research studying misinformation on social media. Previous research has predominantly identified misinformation based on the presence of replies linking to fact-checks from third-party fact-checking organizations – i. e., based on snoping. For instance, many works have studied the diffusion patterns of misleading vs. not misleading ("snoped") posts, finding that misinformation is more viral than the truth (e. g., Vosoughi, Roy, and Aral 2018; Solovev and Pröllochs 2022; Friggeri et al. 2014). However, our analysis suggests that such an identification strategy may impede the generalizability of the findings. While we do not claim that the selection of users contributing to a dedicated community-based fact-checking system is more representative for the population of misinformation on social media as a whole, our results still imply that Notes contributors and Snopers focus on different targets when fact-checking social media content. Due to differences in user bases and content dynamics, earlier findings obtained for snoped posts might not apply to posts that have been fact-checked on community-based fact-checking systems such as Community Notes. Future research should be aware that sample selection plays a key role when studying misinformation and attempt to compile datasets that do not rely on a single fact-checking approach. In particular, compiling a representative sample of *all* misinformation circulating on social media presents an important – yet difficult – challenge for future research.

From a practical perspective, our work has important implications for social media platforms, which can utilize our results to optimize community-based fact-checking systems and strategies to combat misinformation. The observed differences in the selection of fact-checking targets suggest that both approaches might complement each other well. Actively encouraging fact-checking of social media content via both snoping and dedicated community-based fact-checking systems (such as Community Notes) could lead to improved coverage and may help to combat misinformation on social media more effectively. Alternatively, platforms could integrate snopes on their platforms (e. g., by highlighting fact-checks in reply threads) or even actively encourage users that have snoped a social media post to write a full-fledged community fact-check. Platforms could further combine both approaches with machine learning, in order to enhance early warning systems for misinformation. In sum, by considering both snopes and Community Notes, future work might develop more effective strategies for reducing the proliferation of misinformation.

Limitations and Future Research: Our work has several limitations, which provide promising opportunities for future research. First, due to the observational nature of our work, we report associations and refrain from making causal claims. Second, more research is necessary to better understand which groups of users engage in community-based fact-checking and differences in the expertise of these groups. Third, X may have removed some particularly egregious misinformation through content moderation efforts (X 2023). However, related work suggests that the number of deleted posts is relatively small and unlikely to change the main findings in observational misinformation studies (Solovev and Pröllochs 2022). Fourth, our inferences are limited to community-based fact-checking on X and the pilot phase of the Community Notes feature. Community-based fact-checking on X may evolve to a different steady-state due to a growing/more experienced user base and changes in functionality. Future work may analyze whether the observed patterns are generalizable to posts from other fact-checking systems and platforms. Lastly, more research is necessary to better understand the role of manipulation attempts, (political) biases, performative vs. deliberative motivations, and the conditions under which the wisdom of crowds can be unlocked for fact-checking.

Conclusion

The spread of misinformation on social media is a pressing societal problem that researchers and practitioners 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 how community fact-checkers select their targets on social media. For this purpose, we compared the characteristics of social media posts that have been community fact-checked on X's Community Notes platform with social media posts that

have been snoped. Our analysis implies that Notes contributors and Snopers focus on different targets when fact-checking social media content and that both approaches might complement each other well. These findings have important implications for social media providers, which can use our results to optimize community-based fact-checking systems and strategies to combat misinformation on their platforms.

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. All analyses are based on publicly available data. To respect privacy, we explicitly do not publish usernames in our paper and only report aggregate results. We declare no competing interests.

Acknowledgments

This work was supported by a research grant from the German Research Foundation (DFG grant 492310022).

References

Allcott, H.; and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2): 211–236.

Allen, J.; Arechar, A. A.; Pennycook, G.; and Rand, D. G. 2021. Scaling up Fact-Checking Using the Wisdom of Crowds. *Science Advances*, 7(36): eabf4393.

Allen, J.; Howland, B.; Mobius, M.; Rothschild, D.; and Watts, D. J. 2020. Evaluating the Fake News Problem at the Scale of the Information Ecosystem. *Science Advances*, 6(14): eaay3539.

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*.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science*, 348(6239): 1130–1132.

Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. New Threats to Society From Free-Speech Social Media Platforms. *Communications of the ACM*, 66(10): 37–40.

Bhuiyan, M. M.; Zhang, A. X.; Sehat, C. M.; and Mitra, T. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. In *CSCW*.

Bond, R. M.; and Garrett, R. K. 2023. Engagement With Fact-Checked Posts on Reddit. *PNAS Nexus*, 2(3): pgad018.

Brandtzaeg, P. B.; and Følstad, A. 2017. Trust and Distrust in Online Fact-Checking Services. *Communications of the ACM*, 60(9): 65–71.

Calo, R.; Coward, C.; Spiro, E. S.; Starbird, K.; and West, J. D. 2021. How Do You Solve a Problem Like Misinformation? *Science Advances*, 7(50): eabn0481.

Chuai, Y.; Tian, H.; Pröllochs, N.; and Lenzini, G. 2023. The Roll-Out of Community Notes Did Not Reduce Engagement With Misinformation on Twitter. arXiv:2307.07960.

Donovan, J. 2020. Social-Media Companies Must Flatten the Curve of Misinformation. *Nature*.

Drolsbach, C.; and Pröllochs, N. 2023a. Diffusion of Community Fact-Checked Misinformation on Twitter. In *CSCW*.

Drolsbach, C. P.; and Pröllochs, N. 2023b. Believability and Harmfulness Shape the Virality of Misleading Social Media Posts. In *WWW*.

Epstein, Z.; Pennycook, G.; and Rand, D. 2020. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *CHI*.

Feuerriegel, S.; DiResta, R.; Goldstein, J. A.; Kumar, S.; Lorenz-Spreen, P.; Tomz, M.; and Pröllochs, N. 2023. Research Can Help to Tackle AI-Generated Disinformation. *Nature Human Behaviour*, 7: 1818–1821.

Friggeri, A.; Adamic, L. A.; Eckles, D.; and Cheng, J. 2014. Rumor Cascades. In *ICWSM*.

Gallotti, R.; Valle, F.; Castaldo, N.; Sacco, P.; and De Domenico, M. 2020. Assessing the Risks of 'Infodemics' in Response to COVID-19 Epidemics. *Nature Human Behaviour*, 4(12): 1285–1293.

Geissler, D.; Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. Russian Propaganda on Social Media During the 2022 Invasion of Ukraine. *EPJ Data Science*, 12(1): 1–20.

Godel, W.; Sanderson, Z.; Aslett, K.; Nagler, J.; Bonneau, R.; Persily, N.; and Tucker, J. A. 2021. Moderating With the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*, 1(1): 1–36.

Graves, L. 2017. Anatomy of a Fact Check: Objective Practice and the Contested Epistemology of Fact Checking. *Communication, Culture & Critique*, 10(3): 518–537.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425): 374–378.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

Hannak, A.; Margolin, D.; Keegan, B.; and Weber, I. 2014. Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations. In *ICWSM*.

ImageMagick Studio LLC. 2023. ImageMagick. https://imagemagick.org. Version: 7.0.10.

Iyengar, S.; Lelkes, Y.; Levendusky, M.; Malhotra, N.; and Westwood, S. J. 2019. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22(1): 129–146.

Jakubik, J.; Vössing, M.; Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. Online Emotions During the Storming of the US Capitol: Evidence from the Social Media Network Parler. In *ICWSM*.

Kim, A.; and Dennis, A. R. 2019. Says Who? The Effects of Presentation Format and Source Rating On Fake News in Social Media. *MIS Quarterly*, 43(3): 1025–1039.

Lazer, D. M. J.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; and Zittrain, J. L. 2018. The Science of Fake News. *Science*, 359(6380): 1094–1096.

Lutz, B.; Adam, M. T. P.; Feuerriegel, S.; Pröllochs, N.; and Neumann, D. 2023. Affective Information Processing of Fake News: Evidence from NeuroIS. *European Journal of Information Systems*, Forthcoming.

Ma, Y.; He, B.; Subrahmanian, N.; and Kumar, S. 2023. Characterizing and Predicting Social Correction on Twitter. In *WebSci*.

Margolin, D. B.; Hannak, A.; and Weber, I. 2018. Political Fact-Checking on Twitter: When Do Corrections Have an Effect? *Political Communication*, 35(2): 196–219.

Martel, C.; Allen, J. N. L.; Pennycook, G.; and Rand, D. 2023. Crowds Can Effectively Identify Misinformation at Scale. *Perspectives on Psychological Science*, 1–12.

Micallef, N.; He, B.; Kumar, S.; Ahamad, M.; and Memon, N. 2020. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. In *IEEE BigData*.

Mohammad, S.; and Turney, P. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *NAACL-HLT*.

Moore, R. C.; Dahlke, R.; and Hancock, J. T. 2023. Exposure to Untrustworthy Websites in the 2020 US Election. *Nature Human Behaviour*, 7: 1096–1105.

Mosleh, M.; Martel, C.; Eckles, D.; and Rand, D. 2022. Promoting Engagement With Social Fact-Checks Online. *OSF Preprints*.

Oh, O.; Agrawal, M.; and Rao, H. R. 2013. Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises. *MIS Quarterly*, 37(2): 407–426.

Opgenhaffen, M. 2022. Fact-Checking Interventions on Social Media Using Cartoon Figures: Lessons Learned from "the Tooties". *Digital Journalism*, 10(5): 888–911.

Pennycook, G.; and Rand, D. G. 2019. Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality. *PNAS*, 116(7): 2521–2526.

Pew Research Center. 2016. Many Americans Believe Fake News is Sowing Confusion. https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/. Accessed: 2023-11-23.

Pew Research Center. 2019. Republicans Far More Likely Than Democrats to Say Fact-Checkers Tend to Favor One Side. https://pewrsr.ch/2Fz9e22. Accessed: 2023-11-23.

Pröllochs, N. 2022. Community-Based Fact-Checking on Twitter's Birdwatch Platform. In *ICWSM*.

Pröllochs, N.; Bär, D.; and Feuerriegel, S. 2021a. Emotions Explain Differences in the Diffusion of True vs. False Social Media Rumors. *Scientific Reports*, 11: 22721.

Pröllochs, N.; Bär, D.; and Feuerriegel, S. 2021b. Emotions in Online Rumor Diffusion. *EPJ Data Science*, 10(1): 51.

Pröllochs, N.; and Feuerriegel, S. 2023. Mechanisms of True and False Rumor Sharing in Social Media: Collective Intelligence or Herd Behavior? In *CSCW*.

Resnick, P.; Alfayez, A.; Im, J.; and Gilbert, E. 2021. Informed Crowds Can Effectively Identify Misinformation. arXiv:2108.07898.

Rinker, T. W. 2019. *sentimentr: Calculate Text Polarity Sentiment*. Buffalo, New York. Version 2.7.1.

Robertson, C.; Pröllochs, N.; Schwarzenegger, K.; Parnamets, P.; Van Bavel, J. J.; and Feuerriegel, S. 2023. Negativity Drives Online News Consumption. *Nature Human Behaviour*, 7(5): 812–822.

Robertson, C. T.; Mourão, R. R.; and Thorson, E. 2020. Who Uses Fact-Checking Sites? The Impact of Demographics, Political Antecedents, and Media Use on Fact-Checking Site Awareness, Attitudes, and Behavior. *The International Journal of Press/Politics*, 25(2): 217–237.

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 *CIKM*.

Shao, C.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2016. Hoaxy: A Platform for Tracking Online Misinformation. In *WWW Companion*.

Smith, R. 2007. An Overview of the Tesseract OCR Engine. In *ICDAR*.

Solovev, K.; and Pröllochs, N. 2022. Moral Emotions Shape the Virality of COVID-19 Misinformation on Social Media. In *WWW*.

Surowiecki, J. 2005. *The Wisdom of Crowds*. New York, NY, USA: Knopf Doubleday Publishing Group.

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. Accessed: 2023-11-23.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The Spread of True and False News Online. *Science*, 359(6380): 1146–1151.

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.

Wojcik, S.; Hilgard, S.; Judd, N.; Mocanu, D.; Ragain, S.; Hunzaker, M. B. F.; Coleman, K.; and Baxter, J. 2022. Birdwatch: Crowd Wisdom and Bridging Algorithms Can Inform Understanding and Reduce the Spread of Misinformation. arXiv:2210.15723.

X. 2023. X is Commited to the Open Exchange of Information. https://transparency.twitter.com/en.html. Accessed: 2023-11-23.

Zhang, X.; Malkov, Y.; Florez, O.; Park, S.; McWilliams, B.; Han, J.; and El-Kishky, A. 2022. TwHIN-BERT: A Socially-Enriched Pre-Trained Language Model for Multilingual Tweet Representations. arXiv:2209.07562.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? No, all datasets are open source and publicly available
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Supplementary Materials

Additional Account Characteristics

Fig. S1a illustrates the distributions of the number of followees, i. e., the number other users that the fact-checked users are following. Notes contributors tend to focus on users with higher followee counts ($\text{mean}_{\text{notes}}$ = 5,771; $\text{mean}_{\text{snopes}}$ = 5,412; [KS-test: $D = 0.095$; $p < 0.001$]). Additionally, Fig. S1b displays the distributions of the fact-checked users' account ages (in days). It becomes apparent that authors of Community Notes tend to target posts by users with older accounts compared to Snopers ($\text{mean}_{\text{notes}}$ = 3,171; $\text{mean}_{\text{snopes}}$ = 2,567; [KS-test: $D = 0.160$; $p < 0.001$]).

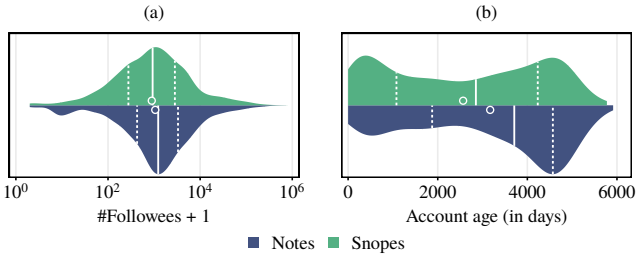

Figure S1: Split violin plot comparing the distributions of the (a) the numbers of followees, and (b) the account ages (in days) of the fact-checked posts' authors. Shown are kernel density estimates (colored areas), mean values (white circles), and quartile values (white lines).

Discrete Emotions

Analogous to previous research (Vosoughi, Roy, and Aral 2018), we used the NRC emotion lexicon (EmoLex; Mohammad and Turney 2010) to measure eight basic emotions in the fact-checked posts, namely, anticipation, surprise, anger, fear, trust, disgust, joy, and sadness. For all posts, the content was tokenized and the frequency of dictionary terms per basic emotion was counted, resulting in an eight-dimensional emotion score.

Fig. S2 shows the fact-checked posts' mean emotion scores. There is barely any difference between Community Notes and snopes. Nonetheless, KS-tests on the individual dimensions are still all significant at common statistical significance thresholds (each $p < 0.05$).

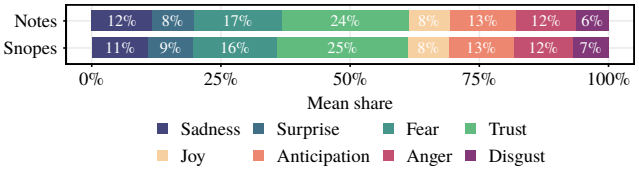

Figure S2: Discrete emotions in fact-checked posts.

Regression Results Without Propensity Score Matching

Fig. S3 depicts the regression coefficients along with their associated 95 % confidence intervals for the models without propensity score matching. The findings closely mirror those derived from the model based on the propensity-matched datasets (refer to Fig. 10).

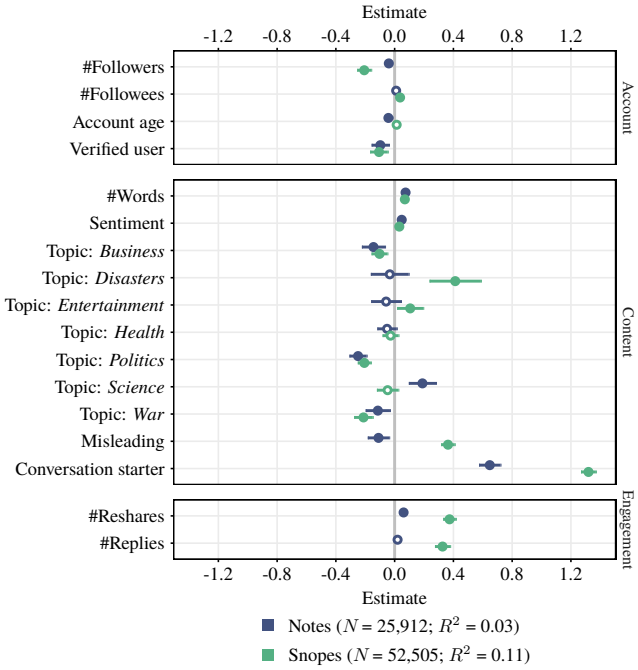

Figure S3: Coefficient estimates (circles) and their 95 % confidence intervals (bars) for the models based on the full datasets (i. e., without propensity score matching). The dependent variables are the fact-checking delays, i. e., the lengths of the timespans between the posting dates of the original posts and the fact-checks. Intercepts and monthly fixed effects are included. Coefficient estimates that are statistically significant ($p < 0.05$) are shown with filled circles.