What Makes a Good Fact Check? Insights from X's Community Notes

Levi Goldberg

Fall 2025

Abstract

This research project examines the effectiveness of different fact-checking formats on social media through a data analysis on X's Community Notes. While Community Notes, a crowd-sourced fact-checking system, has shown promise in reducing engagement with misleading posts, the inconsistent quality of these Notes necessitates investigation into the format of successful fact-checks. Building on existing research suggesting that effective fact-checks are positively phrased, concise, and created soon after the misleading post, this study analyzes these attributes in crowd-sourced fact-checking Notes on X ($n =$ **[Number of Notes]**). These characteristics were then correlated with community-provided ratings of the Notes to assess which features most strongly predict perceived effectiveness. The analysis is broken down by the theme of the misinformation (e.g. political, entertainment, etc.) to examine the effectiveness of fact checks in different contexts.

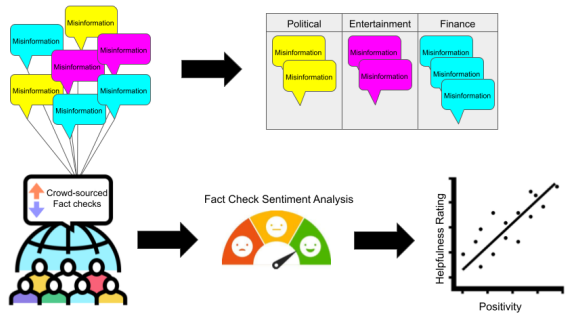

Figure 1: Graphical Abstract v.1 (still under revision)

1 Introduction

When the Community Notes program was launched on Twitter (now X) under the name of Birdwatch, its goal was to provide a transparent platform for misinformation detection and correction. If X users encountered a post that they thought contained misleading information, they could now add credibility indicators and a context providing Note to attempt to fact check the post. These Notes would then be rated by other users to determine how beneficial they were as fact checks, with the algorithm behind which Notes would be displayed alongside a misleading post being open source. Further in line with their goal of transparency, all the data collected by X regarding Community Notes are available for download on their website.

Since these Notes are provided by users, they may not be as effective at correcting misinformation as fact checks provided by professionals, but this system allows for greatly enhanced scalability and relies on "the wisdom of the crowds" - the notion that collective knowledge from a diverse group can exceed the knowledge of any individual [1]. On a platform like X with millions of daily posts, it would be impossible for professional fact checkers to review every one. The idea behind Community Notes is that a large group of individuals can review a much wider range of posts, while relying on their collective knowledge to create efficient fact checks.

Prior research in this problem space has found evidence that Community Notes are effective at detecting misinformation, but these Notes are not all equally effective at actually correcting the misinformation. Although Community Notes may provide the solution to fact checking at scale on social media, it remains unclear which types of Notes are the most effective at convincing misinformed users of the truth. Professional fact-checkers have years of experience to inform how they phrase a response to misinformation, but Community Note contributors have no such expertise. X provides their Notes contributors with some limited advice for the format of their fact checks, and I believe that a comprehensive study on the relationship between a Note's format and its effectiveness could provide much more useful, data driven insights for crowd-sourced fact checkers.

I hypothesized that the format, phrasing, and timing of a Note has a direct relationship with that Note's effectiveness at correcting misinformation, as reflected in the crowd-sourced helpfulness rating of that Note. To test this hypothesis, I analyzed the Community Notes data set, performed sentiment analysis on each Note to calculate a positivity score, and ran statistical tests to determine the influence of these factors on the success of a Note. To examine how the influence of these factors changes in different contexts, I categorized each Note by the theme of the post it was responding and tested each

category separately.

2 Related Works

2.1 Prior Research on Community Notes

The natural first step in evaluating the effectiveness of Community Notes is determining if they are able to reduce the spread of misinformation. A study by Slaughter et al. on the diffusion of misinformation on X determined that posts that received context-providing Notes had a significantly lower rate of engagement [2]. The researchers found that posts receiving a Note had an average of a 40% reduction in the number of comments and reposts. However, this reduction lessened significantly the longer between the creation of the post and the addition of the Note, suggesting that the sooner a fact check is created the more effective it will be.

This study also determined that there was a significant gap in the effectiveness of context-providing Notes on political versus non-political posts, with Notes on political posts being seen as more biased and less trustworthy. In this same vein, a paper by Drolsbach et al. found that professional fact checkers were viewed as more biased and less trustworthy than crowd-sourced fact checks [3]. Further, they determined that context-providing Notes enabled a human to better identify misinformation than simple credibility-indicators that didn't explain why the post was misleading.

The research thus far is extremely favorable towards the effectiveness of Community Notes at countering misinformation. However, it still remains to be seen if this style of countering misinformation can entirely replaced professional fact-checkers, which is precisely what Borenstein et al. attempted to determine [4]. By using an LLM to annotate the hundreds of thousands of Notes and the posts they were attached to in the Community Notes data set, they found that the highest-rated Notes and Notes responding to the most complex misinformation overwhelmingly included references to professional fact checking sources. This suggests that whether or not a professional is involved in fact checking a potentially misleading post, the research provided by these professionals is still crucial to successfully correcting misinformation. One limitation of this study is that the LLM used to annotate the dataset was unable to process X posts that contained non-text media, such as images or videos. This is a major problem for anyone working with the Community Notes data, since the scale of the dataset necessitates automating parts of the analysis.

Another flaw with Community Notes is its susceptibility to organized efforts to influence its content. This is, in part, because the algorithm behind determining which Notes would be displayed alongside a misleading post is open source, allowing groups of users to "game" the system. One possible solution to this is a new and more opaque algorithm, as posited by De et al. [5]. These researchers created a framework for AI-generated "Supernotes" that synthesize the content of existing Notes on a given post to provide a concise, accurate fact check. This system could still be gamed, but the algorithm behind Supernotes includes a simulated jury trained on the data of which Community Notes have been rated the most helpful over time to rate several different Supernote candidates and promote the one that is most likely to be rated the most helpful. De et al.'s analysis of the effectiveness of these Supernotes found that users rated them as more helpful than traditional Notes. Additionally, they determined that the most important aspect of the prompt used to generate the Supernotes was a precise description of the format it should follow. Since these Supernotes were consistently rated the most effective, this suggests that there exists a relationship between the format of a Note and its effectiveness. The primary drawback of this framework for improving Community Notes is that a Supernote cannot be generated until several notes have already been written by human users. Furthermore, another deficiency of the entire Community Notes platform is that even though it is more scalable than professional fact checking, there are simply not enough active contributors to review every single post on X.

2.2 The Format of Fact Checks

Having established that Community Notes are generally effective despite some limitations, we now explore potential improvements. Notes include quick-to-provide credibility indicators and a more time-consuming context statement. If credibility indicators alone proved effective in countering the spread of misinformation, the platform's scalability could improve. The previously discussed research from Drolsbach et al. suggested that credibility indicators on their own were not as effective as a context-providing statement, which was further confirmed in a study by Lu et al. [6]. For this project, the researchers used AI to attach credibility indicators to misleading social media posts to study how they would affect the diffusion of the post. They found that although these indicators helped some users to identify misinformation, they did little to lower the rate of engagement with or the spread of misleading information. There was also some evidence that credibility indicators are more effective at changing viewers' beliefs when there is also social influence (i.e., comments from other users stating the post is misleading) present, which further supports the notion that a combination of credibility indicators and context-providing responses is the most efficient way of countering misinformation. This conclusion was also supported by Ecker et al., who investigated whether credibility indicators on their own could backfire and cause misinformation to spread faster, which they found

no evidence for [7]. Their research also suggested that short fact checks that succinctly explained why the information is incorrect were more effective at countering misinformation than longer, more in-depth explanations.

Further regarding the format of these context statements is a paper by Burel et al. in which, rather than attaching a Note to misleading posts on X, the authors designed a bot to message the post's creator to research how different types of responses would be received by the people spreading misinformation [8]. They found that spreaders of misinformation were most likely to respond positively to being fact checked if the statement was phrased politely.

One format of responding to misinformation used by previously cited papers, including Burel et al. and Ecker et al., is a narrative fact check, in which a story is told to explain why the information is misleading. There was some belief that these may be more effective than non-narrative fact checks, due to the way a narrative format can enhance comprehension and retention. However, a different study by Ecker et al. determined that when the information provided in a narrative and non-narrative refutation has minimal differences, there is no significant difference in the refutation's effectiveness at countering misinformation [9].

Thus, the research suggests that the most effective format for a fact check is a combination of credibility indicators and a refutation of the misinformation that is provided in a timely manner, concise, and positively phrased.

3 Methods

3.1 Data Extraction

The open-source data from Community Notes is stored in Tab-Separated Value (TSV) files, each of which contains information about approximately 100,000 Notes. As of September 2025, there are twenty files of data, though more are added with the creation of new Notes. Each row in each file contains information about a specific Note, including the ID of the Note, when it was created, how helpful it was rated by other Community Notes users, and the ID of the original X post that the Note was attached to, each of which I extracted. A separate data file provides the credibility indicators that each Note contributor gave to the post they were fact checking, which were also extracted for analysis.

The Note ID and post ID can both be used to find a JSON webpage containing the text and time of creation of each Note/post. I used the `requests` library in Python to scrape through these JSON pages and extract these attributes for each Note and each post that received a Note.

3.2 Processing The Posts That Received Notes

Once the text of each post that received a Note had been gathered, it needed to be sorted through to remove any posts consisting mostly of non-text media, such as photos and videos, or non-English text. This was because my method of categorizing posts used a dictionary to associate themes with keywords and then scanned through the text of each post and labeled it with the themes corresponding to any keywords that appeared. Posts containing only photos or videos and posts in other languages could not be categorized by this method.

The theme categories I used were politics, international, entertainment, finance, science/technology, health/wellness, environment, culture, and swear words. The keywords associated with these themes were chosen to be unique to each theme, although if multiple keywords were identified, a post could be labeled as belonging to multiple categories.

Finally, the credibility indicators attached to each post receiving a Note were extracted. These indicators allow the Note contributors to label misleading posts with the reason they think the post is misleading (e.g. factual errors, out of date information, satire, etc.). Although not directly related to my hypothesis, this data could complement the post categorization data for an analysis of what kinds of posts are most frequently labeled as which kinds of misinformation.

3.3 Processing The Notes

Once all of the non-text posts and their associated Notes had been removed, the remaining Notes needed to be cleaned to remove any special characters, punctuation, and URLs to prepare them for sentiment analysis. In Python, the `unicodedata` package was used to remove special fonts, the `emoji` package was used to remove emojis, and the `re` package was used to remove URLs. Removing punctuation and standardizing the text as lowercase was done using built in Python methods. **[Explanation of sentiment analysis and how I performed it]**

Next, the timeliness of each Note needed to be calculated by comparing its time of creation with the time of creation of the post it was responding to. Another easy metric to calculate was the length of each Note.

The final Note metric that needed to be determined was the crowd-sourced helpfulness score. Ratings for each Note are provided by other X users who can tag the Note with a number of different categories, some of which label the post as helpful (e.g. helpfulInformative, helpfulGoodSources, helpfulUnbiasedLanguage, etc.) and some of which label the post as unhelpful (e.g. notHelpfulOffTopic, notHelpfulIrreleventSources, notHelpfulIncorrect, etc.). In total, there are nine labels that fall under the category of helpful and thirteen

that are unhelpful. To compute an overall score for the post, I subtracted the proportion of unhelpful ratings from the proportion of helpful ratings for each Note. For Notes that received multiple ratings, their overall score was the average of the scores for each rating. This meant that Notes that had a higher number of unhelpful ratings than helpful ones would be given a negative score, and vice versa would be given a positive score.

3.4 Evaluation

Now that all of the data had been processed, it was time to begin the analysis. By querying my compiled data with SQL, I separated the Notes based on the category of the post they were responding to. Next, I used R to calculate the correlation between each of the Note metrics (positivity score, length, and timeliness) with the Notes helpfulness score, as well as perform a principle component analysis to determine which of these metrics had the greatest affect on helpfulness score. These statistical tests were performed on the aggregate of the Note data, as well as on each individual category. Finally, I produced a number of auxiliary data visualizations in R.

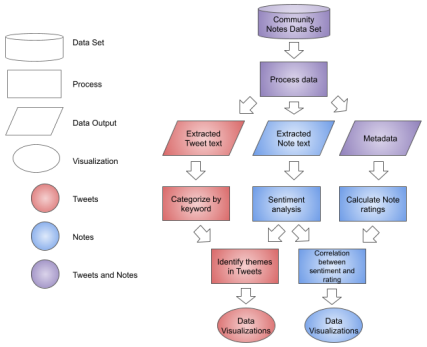

Figure 2: Data Architecture Diagram v.1 (still under revision)

4 Results

5 Limitations

The first major limitation of this research is that it did not examine any non-text posts on X. This narrowed down the analysis from [**total number of posts in dataset**] posts with [**total number of Notes in dataset**] corresponding Notes to [**final number of posts after cleaning**] with [**final number of Notes after cleaning**] corresponding Notes. These posts needed to be removed from the dataset since it would be impossible to categorize them based on keywords.

This method of keyword categorization is another limitation of my design. A post could fall within a given category, but if it doesn't contain any of the keywords associated with that category, then it won't be labeled as such. I tried to overcome this limitation by making an expansive keyword dictionary, but I had to be careful to choose only keywords that uniquely identified that category and would (almost) never appear in another context. A more comprehensive solution to this problem would be to use machine learning to train an AI model to categorize each post. This method would likely result in far fewer posts being uncategorized. Further, if the model was also trained to categorize images, it could overcome the first limitation I discussed. Although outside the scope of this research, this is a prominent direction for future work.

6 Conclusion

8 Works Cited

[1] Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* Doubleday.

[2] Slaughter, I., Peytavin, A., Ugander, J., & Saveski, M. (2025). Community notes reduce engagement with and diffusion of false information online. *Proceedings of the National Academy of Sciences*, *122*(38), e2503413122.

[3] Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS nexus*, *3*(7), pgae217.

[4] Borenstein, N., Warren, G., Elliott, D., & Augenstein, I. (2025). Can community notes replace professional fact-checkers? *arXiv preprint arXiv:2502.14132*.

[5] De, S., Bakker, M. A., Baxter, J., & Saveski, M. (2024). Supernotes: Driving consensus in crowd-sourced fact-checking. *arXiv preprint arXiv:2411.06116*.

[6] Lu, Z., Li, P., Wang, W., & Yin, M. (2022). The effects of ai-based credibility indicators on the detection and spread of misinformation under social influence. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–27.

[7] Ecker, U. K., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2020). The effectiveness of short-format refutational fact-checks. *British journal of psychology*, *111*(1), 36–54.

[8] Burel, G., Tavakoli, M., & Alani, H. (2024). Exploring the impact of automated correction of misinformation in social media. *AI Magazine*, *45*(2), 227–245.

[9] Ecker, U. K., Butler, L. H., & Hamby, A. (2020). You don't have to tell a story! a registered report testing the effectiveness of narrative versus non-narrative misinformation corrections. *Cognitive Research: Principles and Implications*, *5*, 1–26.