Literature Review: Crowdsourced Misinformation Verification: Community Notes

Nathaniel Hurst

January 20, 2025

1 Introduction

As Meta announces its progressive switch in favor of a community notes model similar to that of Twitter/X, instead of traditional fact checkers, ample opportunity has arisen to apply misinformation research for the good of these platforms and their users. The community notes model introduces a new and interesting approach to fact-checking. By outsourcing to the community, social media platforms offload responsibility for fact checking and reduce payroll. But is this approach effective for the company and the community? We will explore this question through a literature review of some recent misinformation papers and an analysis of the community notes model.

2 Why not Professionals?

One study by Drolsbach, Solovev, & Pröllochs, (2024) showed that while people have a high amount of trust in professional fact checkers, community notes are often preferred because they are able to give more information and context rather than just a misinformation flag. This allows the community notes model to have a highly specialized approach to each instance of misinformation, increasing the value and trust in fact checking. This study also suggests that community notes have a positive effect when it comes to misinformation along political lines. It shows that a community notes model is surprisingly effective when it comes to people identifying and accepting misinformation which aligns with their political beliefs. Especially when dealing with partisan people, replacing an expert flag with a community note significantly increases trust levels. One expert (Jay Van Bavel, 2025) also believes that when it comes to political information, professional fact checkers are less likely to be trusted, further lending merit to the community notes model.

A study by Pilarski et al. (2024) assessed the differences in fact checking methods in how much misinformation they cover and what posts tend to attract more attention, mainly focusing on the distinction between snoping (linking to professional fact checking websites) and the community notes model. This study shows that these two methods are both effective, but tend to cover widely different areas of misinformation. For example the community notes model tends to focus most of its attention on highly followed accounts and posts. Another key observation from this study is that the community notes model is often slower at fact checking than snoping. This is a critical observation as when posts on social media have short half lives, every second that misinformation stays not fact checked

is a second that more and more people are exposed to and engaging with it. One more thing to note is that this study found that when snopers and community notes models overlapped they tended to agree with each other on the accuracy of certain statements. This is important as it shows that at least to an extent the community notes model is doing a good job in terms of accuracy.

3 Community Notes

The obvious advantages of community notes is the increases in trust and outsourcing of labor, but how do they actually affect the social media platform? One study suggests that community notes did not actually impact the amount of misinformation that was being engaged with (Chuai et al., 2024). This article shows evidence that while the implementation of community notes increased the amount of fact-checking going on, especially when it came to verified users with large followings, the community notes model did not significantly reduce how much users engaged with misinformation through likes, retweets, etc. It is unknown whether this was a result of the community notes model, or the algorithm which governs the community notes. This paper however does suggest that the system used by twitter in which a certain critical mass of ratings is needed to publish a community note is not effective in keeping up with the rapid spread of misinformation; it just takes too long to have the community posted, by then the misinformation has spread too far.

Another study by Renault et al. (2024) showed that the community notes model reduced the spread of posts which are labeled with additional context. Adding this additional context through the community notes model drastically reduced retweets, and had a less, but still significant, effect on comments and quotes. However it also points out that the community notes system is inherently flawed, as the average time for a note to be published is after the tweet has gotten to 80% of its total reach. This highlights problems with the current algorithm in the form of timeliness.

4 Crowdsourced Content Moderation

Another data-driven analysis of community notes yielded valuable insights into the effectiveness of the community notes model to fight misinformation in the context of crowdsourcing (Gao, Zhang, & Rui, 2024). This study shows evidence that crowd-checking is an effective method to encourage retractions of misinformation. For example a community note under a misinformation post increases the likelihood as well as timeliness of retraction of the post. Additionally this study showed that retraction had more to do with observed effects (those who actually interacted with the misinformation) rather than presumed effects (those who just saw the post). These insights are vital to understanding the appeal of a community-based model.

5 Methodology

Most studies to do with the spread of misinformation and the community notes model center around analyzing data from Twitter/X as this is the platform where the model was pioneered. After Elon Musk bought X there has been significant restrictions on data

gathering, most notably X stopped offering access to their API for academic researchers. However data on the community notes model is mostly public and available for download, this includes information about the rating system and information on each note published in a given time window. One more thing to note is that the algorithm which governs the community notes publishing and rating system is also public on GitHub.

6 Conclusion

There are still many questions related to the community notes model, ranging from questions about its effectiveness to questions about implementation. There is also minimal talk about the community notes algorithm itself, is it effective, can it be changed to improve, and so on. Research about the effectiveness of community notes across topics is also scarce. Overall it is an exciting time for the community notes model and misinformation research, and with Meta's recent shift towards this model there is ample opportunity for improvement in these verification systems.

References

[1] Drolsbach, C. P., Solovev, K., and Pröllochs, N., "Community Notes increase trust in fact-checking on social media," *PNAS Nexus*, vol. 3, no. 7, July 2024. Available at: `https://doi.org/10.1093/pnasnexus/pgae217`. Published: May 31, 2024. Accessed: January 19, 2025.

[2] Adam, D. (2025). Does Fact-Checking Work? Here's What the Science Says. *Nature Magazine*.

[3] Pilarski, M., Solovev, K. O., & Pröllochs, N. (2024). Community Notes vs. Snoping: How the Crowd Selects Fact-Checking Targets on Social Media. Proceedings of the International AAAI Conference on Web and Social Media, 18(1), 1262-1275.

[4] Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2), 1-52. https://doi.org/10.1145/3686967

[5] Renault, T., Restrepo Amariles, D., Troussel, A. (2024). Collaboratively adding context to social media posts reduces the sharing of false news. *arXiv:2404.02803 [econ.GN]*. https://doi.org/10.48550/arXiv.2404.02803

[6] Gao, Yang and Zhang, Maggie and Rui, Huaxia, Can Crowdchecking Curb Misinformation? Evidence from Community Notes (October 17, 2024). Available at SSRN: https://ssrn.com/abstract=4992470 or http://dx.doi.org/10.2139/ssrn.4992470