UNIVERSITY OF LEEDS   University of Sheffield   UNIVERSITY of York

Truth with a Twist: The Rhetoric of Persuasion in Professional vs. Community-Authored Fact-Checks

Olesya Razuvayevskaya
o.razuvayevskaya@sheffield.ac.uk
The University of Sheffield
Sheffield, United Kingdom

Kalina Bontcheva
k.bontcheva@sheffield.ac.uk
The University of Sheffield
Sheffield, United Kingdom

Abstract

This study presents the first large-scale comparison of persuasion techniques present in crowd- versus professionally-written debunks. Using extensive datasets from Community Notes (CNs), EUvsDisinfo, and the Database of Known Fakes (DBKF), we quantify the prevalence and types of persuasion techniques across these fact-checking ecosystems. Contrary to prior hypothesis that community-produced debunks rely more heavily on subjective or persuasive wording, we find no evidence that CNs contain a higher average number of persuasion techniques than professional fact-checks. We additionally identify systematic rhetorical differences between CNs and professional debunking efforts, reflecting differences in institutional norms and topical coverage. Finally, we examine how the crowd evaluates persuasive language in CNs and show that, although notes with more persuasive elements receive slightly higher overall helpfulness ratings, crowd raters are effective at penalising the use of particular problematic rhetorical means.

CCS Concepts

• **Information systems** → **Crowdsourcing**; **Web applications**; **Social networks**; **Answer ranking**; **Trust**; **Reputation systems**.

Keywords

community notes, content moderation, misinformation

1 Introduction and Background

Community Notes (CNs) on X (formerly known as Twitter Birdwatch) is a crowd-driven content moderation mechanism for countering potentially misleading information by providing contextual clarifications, corrections, or additional sources [5, 23]. This moderation process is self-sustaining: contributors both author notes and collectively evaluate their quality by voting on their helpfulness. Bridging algorithms subsequently determine which notes achieve sufficient cross-ideological agreement to be displayed publicly [2, 15, 25]. Prior research highlights the effectiveness of CNs in reducing the spread and impact of misinformation [3, 7, 19], particularly for posts containing factual inaccuracies [21]. Such interventions frequently lead to post retractions by the original authors [3, 9, 17], although they may also elicit increased user outrage or polarization in subsequent discussions [4].

In comparison to professional fact-checking efforts, CNs offer broader coverage–both in the volume of misinformation addressed and in the diversity of topics requiring regional, cultural, or domain-specific expertise [1]. Beyond these dissemination effects, users tend to express greater trust in corrective interventions when they are accompanied by CNs that explicitly articulate why a post is misleading, in contrast to opaque "false information" labels issued by experts [8]. Additionally, by engaging a substantially larger and more demographically diverse contributor base than traditional fact-checking organizations [1, 2], CNs promote a more democratic and transparent model of content governance. This inclusivity exposes readers to a wider range of perspectives and sources, particularly on controversial or geographically salient topics, reinforcing CNs' potential as a scalable, community-centered approach to social-media accountability [1, 2].

Despite these strengths, there is no consensus among researchers on whether crowdsourced fact-checks are more susceptible to biases than professional ones [1]. Some researchers argue that the bridging algorithm—which requires cross-ideological agreement for a note publication—reduces one-sidedness of CNs relative to professional fact-checks [10]. Others emphasize the system's vulnerabilities to coordination and manipulation. For example, in the context of U.S. politics, posts by Republicans have been shown to receive more notes, both overall and published, than posts by Democrats [18]. Moreover, Draws et al. demonstrate that crowd workers' assessments can diverge from expert-assigned ground-truth labels depending on their political orientations and broader cognitive or ideological dispositions [6]. They further find that crowd workers are more likely than experts to regard misleading claims as truthful [6]. Within this broader debate, some scholars argue that professional fact-checks are generally more objective in style, whereas CNs may employ persuasive or manipulative linguistic strategies, such as appeals to authority or emotionally charged claims [1, 11].

Notwithstanding these concerns, no study to date has carried out a systematic comparison between professional and crowdsourced debunks with respect to their use of persuasive language. This gap motivates the present work, which aims to quantify the extent to which Community Notes adhere to their core values[1] of avoiding personal opinions, insults, and manipulative behavior. Accordingly, we pose the following research questions: (a) Is the CNs' self-governing nature able to identify and penalize the use of persuasion techniques in community notes? (b) Do community-written debunks exhibit a higher degree of persuasive language than professionally-authored ones?

Our code and supplementary material[2] are made openly available to facilitate reproducibility.

---

[1]https://communitynotes.x.com/guide/en/contributing/values
[2]https://github.com/LesyaR/CN_Persuasion_Analysis

2 Datasets and Methodology

This study utilises three datasets containing debunk texts: (1) Community Notes (CN), representing crowd-authored debunks, and two datasets representing debunks written by verification professionals, (2) EUvsDisinfo and (3) the Database of Known Fakes (DBKF). For the CN dataset, we extracted the full, daily updated set of notes on August 10, 2025, yielding 2,016,841 debunks. The EUvsDisinfo dataset consists of professionally-written debunks addressing pro-Kremlin disinformation [12]. Our experiments use only the debunk articles rather than the associated misinformation ones. Since a single article may debunk multiple misinformation claims, only unique articles are retained, resulting in 8,522 debunks in English. The DBKF database[3] contains a large, diverse collection of 681,346 debunks published by fact-checking organisations worldwide, which makes it more directly comparable in scale to CNs. Unlike the EUvsDisinfo debunks which are English only, the DBKF and CNs datasets are highly multilingual, containing debunks in 56 and 72 languages respectively. The language distributions for CNs and DBKF are provided in the supplementary material.

Persuasion techniques in debunk texts are detected using the state-of-the-art multilingual classifier developed by Wu et al. [26]. To increase confidence in individual predictions, the classification threshold is set to 0.8. For consistency across datasets and due to the limitations of transformer-based models, all texts were truncated to a maximum length of 512 tokens. The full list of techniques, along with their definitions, is provided in the supplementary material.

RQ1. Is the CNs' self-governing nature able to identify and penalize the use of persuasion techniques in community notes? To address this question, we examined two aspects. First, using the complete set of note ratings from the CN website, we assessed whether the *degree of note persuasiveness* is correlated with the *degree of note helpfulness*. We operationalised the degree of note persuasiveness as the number of persuasion techniques identified in a given note. To estimate the helpfulness degree, we considered individual user ratings that mark the note as being *Not Helpful*, *Somewhat Helpful*, or *Helpful*. We then mapped each note rating to a numeric value: *Not Helpful*->1, *Somewhat Helpful*->2, or *Helpful*->3. For notes with multiple ratings, we calculated the average rating value. The association between the degree of persuasiveness and helpfulness level is measured using the Spearman correlation coefficient [24]. For each persuasion technique, we additionally compared the distribution of helpfulness scores between notes that use the technique and those that do not, using the Mann–Whitney U test [20]. To assess the *direction* and *magnitude* of the effect, we computed Cliff's delta (δ), a non-parametric effect size metric [14].

RQ2. Do community- written debunks exhibit a higher degree of persuasive language than professionally-authored ones? To compare CN with the two professional fact-checking datasets, we first removed CN notes that classify posts as "Not Misinformation", ensuring comparability with EUvsDisinfo and DBKF, which publish debunks only for misinformation claims. We then performed two analyses. First, we tested whether the CNs dataset contains more persuasion techniques per note on average compared to professionally-written debunks, EUvsDisinfo and DBKF. Given the binary nature of the data and large sample sizes, we used

[3]https://dbkf.ontotext.com/

the non-parametric Mann-Whitney U test to compare distributions between datasets [20].

We next performed a per-technique comparison by formulating the following directional hypotheses for each technique:

- H_1: The prevalence of a given technique is higher in CNs than in DBKF.
- H_2: The prevalence of a given technique is higher in CNs than in EUvsDisinfo.

For each comparison, the H_0 hypothesis assumes that CNs have equal or lower prevalence for each technique. Because the data for each technique are binary (presence/absence), we employed a one-sided *proportions z-test* to test the directional hypotheses [22].

3 Results

RQ1. Is the CNs' self-governing nature able to identify and penalize the use of persuasion techniques in community notes? We first evaluated whether the degree of persuasiveness of notes was correlated with their average helpfulness rating. Our results show the correlation to be statistically significant, but the effect size is extremely small ($\rho \approx 0.039$, p-value$< 10^{-4}$). Importantly, the direction of correlation is *positive*, indicating that, contrary to our hypothesis, notes containing more persuasion techniques are, on average, rated *more* helpful by the crowd.

Table 1 reports the p-values and Cliff's deltas estimated based on Mann–Whitney U statistic for each technique. Positive δ values indicate that the technique is associated with *higher* helpfulness, while negative values indicate the *lower* helpfulness association and near-zero numbers indicate small effect, often minimal in practice. As can be seen, several techniques exhibit small but meaningful directional effects. The largest positive effects were observed for *Consequential Oversimplification* ($\delta = 0.255$), *Causal Oversimplification* ($\delta = 0.115$), *Appeal to Hypocrisy* ($\delta = 0.114$), and *Appeal to Time* ($\delta = 0.1053$), indicating that notes containing these techniques tend to receive *higher* helpfulness ratings. Conversely, techniques such as *Flag Waving* ($\delta = -0.142$), *Guilt by Association* ($\delta = -0.088$), and *Straw Man* ($\delta = -0.068$) were associated with *lower* helpfulness ratings. Most remaining techniques showed very small effects ($|\delta| < 0.05$), suggesting limited practical impact on helpfulness despite highly significant p-values driven by large sample size. Two techniques, *Appeal to Authority* and *False Dilemma / No Choice*, were not statistically significant and exhibited negligible effect sizes, indicating no detectable relationship with helpfulness. For *Appeal to Authority*, this can be attributed to its general association with trustworthy narration and objective presentation [13]. Overall, whereas the degree of persuasiveness has a minimal directional effect, individual techniques do exhibit systematic positive or negative associations with perceived helpfulness.

RQ2. Do community- written debunks exhibit a higher degree of persuasive language than professionally-authored ones? We first tested whether the CNs dataset contains more persuasion techniques on average compared to the EUvsDS and DBKF datasets using the Mann-Whitney U test. The results for both comparisons were insignificant, with p-value=1.00. This suggests that contrary to prior research (Section 1), it is not the case that the argumentation of crowd-written fact-checks is more persuasive in nature as compared to professionally-written fact-checks.

Table 1: Directional effects of individual persuasion techniques on helpfulness (Mann–Whitney U tests).

Technique	p	δ
Appeal to Authority	0.064	-0.005
Appeal to Fear-Prejudice	3×10^{-5}	-0.010
Appeal to Hypocrisy	$< 1 \times 10^{-300}$	0.114
Appeal to Popularity	1×10^{-80}	0.066
Appeal to Time	1×10^{-9}	0.105
Appeal to Values	2×10^{-64}	0.051
Causal Oversimplification	7×10^{-267}	0.115
Consequential Oversimplification	$< 1 \times 10^{-300}$	0.255
Conversation Killer	3×10^{-35}	-0.025
Doubt	1×10^{-57}	0.017
Exaggeration-Minimisation	8×10^{-222}	0.061
False Dilemma-No Choice	0.227	0.005
Flag Waving	$< 1 \times 10^{-300}$	-0.142
Guilt by Association	3×10^{-111}	-0.088
Loaded Language	$< 1 \times 10^{-300}$	0.041
Name Calling-Labeling	1×10^{-102}	0.027
Obfuscation-Vagueness	4×10^{-185}	0.085
Questioning the Reputation	8×10^{-20}	-0.023
Red Herring	1×10^{-17}	0.042
Repetition	$< 1 \times 10^{-300}$	0.079
Slogans	8×10^{-15}	-0.032
Straw Man	1×10^{-11}	-0.068
Whataboutism	1×10^{-44}	0.094

Table 2 summarizes the results of the one-sided z-tests for per-technique presence comparison between the datasets. Positive z-values indicate higher prevalence in CNs, and p-values below 0.05 indicate statistically significant differences. The results indicate that the prevalence of persuasion techniques in CNs (underlined values) differs markedly depending on the comparison dataset. Compared to DBKF, only a few techniques (*Conversation Killer*, *Repetition*, *False Dilemma-No Choice*, and *Whataboutism*) are significantly more prevalent in CNs. For most other techniques, prevalence is lower in CNs, contrary to our initial hypothesis. Compared to EUvsDisinfo, much more techniques (*Appeal to Hypocrisy*, *Appeal to Popularity*, *Appeal to Values*, *Consequential Oversimplification*, *Conversation Killer*, *Doubt*, *Obfuscation-Vagueness-Confusion*, *Questioning the Reputation*, *Red Herring*, *Straw Man*, and *Whataboutism*) are significantly more prevalent in CNs. This can be attributed to the stricter format and the domain specificity of the EUvsDisinfo debunks.

4 Discussion and Conclusions

This paper reported the first large-scale examination of the widely held assumption that crowdsourced debunks tend to employ more persuasive wording than professional ones. Contrary to concerns that the decentralised, volunteer-driven nature of CNs might make them less objective, we found no evidence that CNs employ a greater number of persuasion techniques than professional fact-checks.

It should be noted, however, that our per-technique comparisons reveal a more complex picture. While CNs are less persuasive than the DBKF fact-checks along most dimensions, they show higher prevalence for several techniques when compared to EUvsDisinfo.

This highlights the importance of domain and organisational context: EUvsDisinfo debunks are produced within a narrow geopolitical niche with very stringent editorial oversight, whereas CNs address a far broader thematic and cultural space. The differences therefore underline that rhetorical variation is shaped not only by authorship (crowd vs. professional) but also by institutional aims, editorial norms, and the type of misinformation being countered.

The analysis of rater behaviour further extends existing knowledge of CNs' self-regulating mechanisms. The overall degree of persuasiveness in a note exhibits only a negligible relationship with its average helpfulness degree, and—contrary to our hypothesis—the association is slightly positive. However, examining the presence of individual techniques reveals that crowd raters are particularly sensitive to penalise certain types of persuasive rhetoric, especially patriotically charged claims (*Flag-Waving*); statements that exhibit, biases against individuals through association with specific groups (ethnic, racial, political, ideological, etc.; *Guilt by Association*); and the notes that purposefully misrepresent the original post by addressing an argument that was not made by the post author (*Straw Man*). The latter crowd reaction aligns with the broader prevalence of NNN flagging in Community Notes, in which users identify notes as unnecessary or irrelevant to the post [16]. In contrast, the use of other techniques appears to persuade the raters and increase their trust in CNs. This suggests that CN contributors are sensitive to specific problematic rhetorical strategies even if they do not penalise their cumulative presence.

5 Acknowledgments

References

[1] Isabelle Augenstein, Michiel Bakker, Tanmoy Chakraborty, David Corney, Emilio Ferrara, Iryna Gurevych, Scott Hale, Eduard Hovy, Heng Ji, Irene Larraz, et al. 2025. Community Moderation and the New Epistemology of Fact Checking on Social Media. *arXiv preprint arXiv:2505.20067* (2025).

[2] Paul Bouchaud and Pedro Ramaciotti. 2025. Algorithmic resolution of crowd-sourced moderation on X in polarized settings across countries. doi:10.48550/arXiv.2506.15168 arXiv:2506.15168 [cs].

[3] Yuwei Chuai, Moritz Pilarski, Thomas Renault, David Restrepo-Amariles, Aurore Troussel-Clément, Gabriele Lenzini, and Nicolas Pröllochs. 2024. Community-based fact-checking reduces the spread of misleading posts on social media. doi:10.48550/arXiv.2409.08781 arXiv:2409.08781 [cs].

[4] Yuwei Chuai, Anastasia Sergeeva, Gabriele Lenzini, and Nicolas Pröllochs. 2025. Community Fact-Checks Trigger Moral Outrage in Replies to Misleading Posts on Social Media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23. doi:10.1145/3706598.3713909 arXiv:2409.08829 [cs].

[5] Emillie de Keulenaar and Ivan Kisjes. 2025. Content moderation from Twitter to X: policies, enforcement and community notes. doi:10.17632/D9GGKZP7BC.2

[6] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. 2022. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2114–2124. doi:10.1145/3531146.3534629

[7] Chiara Drolsbach and Nicolas Pröllochs. 2023. Diffusion of Community Fact-Checked Misinformation on Twitter. doi:10.48550/arXiv.2205.13673 arXiv:2205.13673 [cs].

[8] Chiara Patricia Drolsbach, Kirill Solovev, and Nicolas Pröllochs. 2024. Community notes increase trust in fact-checking on social media. *PNAS Nexus* 3, 7 (June 2024), pgae217. doi:10.1093/pnasnexus/pgae217

[9] Yang Gao, Maggie Zhang, and Huaxia Rui. 2024. Can Crowdchecking Curb Misinformation? Evidence from Community Notes. doi:10.2139/ssrn.4992470

Table 2: Per-technique one-sided z-tests comparing prevalence in *CNs* vs *DBKF* and *EUvsDisinfo*.

Technique	CNs vs DBKF		CNs vs EUvsDisinfo	
	Z-statistic	p-value	Z-statistic	p-value
Appeal to Authority	-124.10	1.000	-44.70	1.000
Appeal to Fear-Prejudice	-160.98	1.000	-86.84	1.000
Appeal to Hypocrisy	-128.00	1.000	16.04	< 0.001
Appeal to Popularity	-73.97	1.000	5.18	< 0.001
Appeal to Time	-50.45	1.000	2.20	0.014
Appeal to Values	-107.68	1.000	12.85	< 0.001
Causal Oversimplification	-18.30	1.000	-48.28	1.000
Consequential Oversimplification	-48.95	1.000	6.22	< 0.001
Conversation Killer	45.36	< 0.001	19.08	< 0.001
Doubt	-187.06	1.000	3.97	0.00004
Exaggeration-Minimisation	-120.47	1.000	-17.01	1.000
False Dilemma-No Choice	22.85	< 0.001	-0.82	0.79
Flag Waving	-88.85	1.000	-60.16	1.000
Guilt by Association	-111.78	1.000	-61.53	1.000
Loaded Language	-84.67	1.000	-18.40	1.000
Name Calling-Labeling	-205.16	1.000	-79.70	1.000
Obfuscation-Vagueness-Confusion	-5.62	1.000	12.04	< 0.001
Questioning the Reputation	-267.89	1.000	15.41	< 0.001
Red Herring	-76.88	1.000	4.84	< 0.001
Repetition	137.26	< 0.001	-21.62	1.000
Slogans	-244.23	1.000	-1.05	0.85
Straw Man	-77.73	1.000	2.14	0.016
Whataboutism	6.98	< 0.001	4.86	< 0.001

[10] Uku Kangur, Roshni Chakraborty, and Rajesh Sharma. 2024. Who Checks the Checkers? Exploring Source Credibility in Twitter's Community Notes. doi:10.48550/arXiv.2406.12444 arXiv:2406.12444 [cs].

[11] Sarawut Kankham and Jian-Ren Hou. 2025. Community Notes vs. Related Articles: Assessing Real-World Integrated Counter-Rumor Features in Response to Different Rumor Types on Social Media. *International Journal of Human–Computer Interaction* 41, 12 (2025), 7711–7725. arXiv:https://doi.org/10.1080/10447318.2024.2400389 doi:10.1080/10447318.2024.2400389

[12] João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. EUvsDisinfo: a dataset for multilingual detection of pro-Kremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 5380–5384.

[13] João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2025. Weakly supervised veracity classification with LLM-predicted credibility signals. *EPJ Data Science* 14, 1 (2025), 16.

[14] Guillermo Macbeth, Eugenia Razumiejczyk, and Rubén Daniel Ledesma. 2011. Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations. *Universitas Psychologica* 10, 2 (2011), 545–555.

[15] Aviv Ovadya and Luke Thorburn. 2023. Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance. doi:10.48550/arXiv.2301.09976 arXiv:2301.09976 [cs].

[16] Olesya Razuvayevskaya, Adel Tayebi, Ulrikke Dybdal Sørensen, Kalina Bontcheva, and Richard Rogers. 2025. Timeliness, Consensus, and Composition of the Crowd: Community Notes on X. *arXiv preprint arXiv:2510.12559* (2025).

[17] Thomas Renault, David Restrepo Amariles, and Aurore Troussel. 2024. Collaboratively adding context to social media posts reduces the sharing of false news. doi:10.48550/arXiv.2404.02803 arXiv:2404.02803 [econ].

[18] Thomas Renault, Mohsen Mosleh, and David G. Rand. 2025. Republicans are flagged more often than Democrats for sharing misinformation on X's Community Notes. *Proceedings of the National Academy of Sciences* 122, 25 (2025), e2502053122. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2502053122 doi:10.1073/pnas.2502053122

[19] Isaac Slaughter, Axel Peytavin, Johan Ugander, and Martin Saveski. 2025. Community Notes Moderate Engagement With and Diffusion of False Information Online. doi:10.48550/arXiv.2502.13322 arXiv:2502.13322 [cs].

[20] Kruskal-Wallis Test. 2000. Mann-Whitney Test. *group* 1, 1 (2000), 1–1.

[21] Chenlong Wang and Pablo Lucas. 2024. Efficiency of Community-Based Content Moderation Mechanisms: A Discussion Focused on Birdwatch. *Group Decision and Negotiation* 33, 3 (June 2024), 673–709. doi:10.1007/s10726-024-09881-1

[22] Rachel Webb. 2017. *Mostly Harmless Statistics*. LibreTexts. https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Mostly_Harmless_Statistics

[23] Valerie Wirtschafter and Sharanya Majumder. 2023. Future Challenges for Online, Crowdsourced Content Moderation: Evidence from Twitter's Community Notes. *Journal of Online Trust and Safety* 2, 1 (Sep. 2023). doi:10.54501/jots.v2i1.139

[24] Clark Wissler. 1905. The Spearman correlation formula. *Science* 22, 558 (1905), 309–311.

[25] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. 2022. Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation. doi:10.48550/arXiv.2210.15723 arXiv:2210.15723 [cs].

[26] Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João A. Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (Eds.). Association for Computational Linguistics, Toronto, Canada, 1995–2008. doi:10.18653/v1/2023.semeval-1.275