The Role and Effectiveness of Community Notes in Information Verification

Bhirajaya Bhumimars, Supakarn Bunlert, Yunkai Li
Master of Information Management and System
University of California, Berkeley

Abstract

Community Notes is a crowdsourced fact-checking system that enhances online discussions by allowing users to append contextual notes to potentially misleading posts, with community voting determining their visibility. Currently implemented on Twitter/X, it has emerged as a vital tool in combating misinformation, yet it faces challenges such as manipulation risks, contributor diversity, and timeliness. This study investigates three dimensions of Community Notes: (1) **analyzing trends in Community Note data**, (2) **assessing the accuracy and effectiveness of the provided context**, and (3) **exploring its potential to improve the factuality of large language models (LLMs)**. We will use Twitter's Community Note dataset and PolitiFact's fact-checks. For accuracy testing and LLM improvement, we will focus on political topics due to their societal impact. This research offers timely insights into optimizing community-driven fact-checking for public discourse and AI applications, especially as more platforms are adopting this approach, with a pilot on YouTube in 2024 and Facebook planning to adopt similar systems in 2025.

1 Background and Motivation

The Need for Fact-Checking

Accurate information is critical for informed decision-making, especially in domains like politics, public health, and climate change. Social media platforms like Twitter/X and Facebook amplify misinformation due to anonymous posting, the viral spread of novel or false claims (Dizikes, 2018), engagement-driven algorithms, and limited verification. Traditional fact-checking by organizations like PolitiFact and Snopes relies on expert analysis but struggles with scale, speed, and potential bias. Some AI-driven tools leverage machine learning and NLP for efficiency but face issues like training data bias and contextual misunderstandings.

The Rise of Community Notes

Launched as Birdwatch by Twitter in January 2021, Community Notes empowers users to flag misinformation and add context to the post, subject to a voting process that ensures diverse perspectives before publication. Studies show it achieves high accuracy on COVID-related claims up to 97%, reducing misinformation spread (Allen et al., 2024), though delays in note visibility can limit its impact. In 2024, Meta announced a shift from a third-party fact-checking system to Community Notes for Facebook and Instagram by 2025, citing excessive errors and censorship (Duffy, 2025). This transition underscores the need to evaluate and enhance community-driven systems for transparency and scalability.

2 Research Questions and Methodology

The primary research question guiding this project is: How effective are Community Notes in enhancing information verification, and can they contribute to improving the factuality of large language models (LLMs)? This study explores three dimensions through sub-questions addressing distinct aspects of Community Notes.

2.1 Data Analysis - How effectively do Community Notes identify and address misinformation trends?

Context: Misinformation on social media often follows predictable patterns, such as rapid spread during crises or elections, as evidenced by research showing false news travels faster than truth (Dizikes, 2018). Community-driven fact-checking systems like Community Notes offer a unique approach to combating misinformation by crowdsourcing context and clarifications. Prior studies have investigated whether community-driven efforts can rival professional fact-checkers in accuracy (Borenstein et al., 2025), found that transparency in such systems increases user trust (Drolsbach et al., 2024), and highlighted how framing affects perceived effectiveness (Aruguete et al., 2023).

To situate this work within the existing literature, we will analyze the unique characteristics of Community Notes, particularly their ability to pinpoint and mitigate trending falsehoods. We will explore the factors and contextual conditions that influence their effectiveness, such as user engagement, consensus algorithms, and the nature of the content being fact-checked.

Data: Twitter's Community Note dataset (see Community Notes Guide)

Methods: Apply Latent Dirichlet Allocation (LDA) for topic modeling, analyze metadata (e.g., contributor diversity, voting patterns, timelines).

2.2 Accuracy Evaluation - How accurate and reliable are Community Notes?

Context: Accuracy is the cornerstone of any fact-checking system, and like other community-driven approaches, Community Notes face scrutiny over consistency and bias. (Kangur et al., 2024) questions the credibility of sources behind Community Notes, examining who verifies the crowd-sourced contributions and how biases may affect accuracy and trustworthiness. Studies have shown that Community Notes can achieve high accuracy in controlled contexts, such as COVID-19 misinformation (Allen et al., 2024), but they often struggle in chaotic, high-stakes scenarios like elections (Ortutay, 2024). These findings highlight the need for further investigation into how accuracy varies across different contexts and content types.

Data: Existing fact-checking databases (e.g., Politifact), cross-referenced with Twitter's Community Note dataset (see Community Notes Guide)

Methods: Use NLP techniques to calculate accuracy of information

2.3 LLM Integration - Can Community Notes enhance LLM factuality?

Context: Large language models (LLMs) often produce plausible but factually incorrect outputs, a challenge exacerbated by limitations in training data. While retrieval-augmented generation (RAG) allows models to fetch real-time web content, it does not inherently guarantee source credibility, making models susceptible to hallucinations. Research indicates that preference-based fine-tuning alone does not substantially improve factual accuracy (Ivison et al., 2024).

Integrating community-driven insights into LLMs holds significant potential for enhancing factuality. Community Notes, with their crowd-verified context, could serve as a novel data source to refine LLM outputs, ensuring that generated content aligns more closely with verified information. (Liu et al., 2024) suggest that leveraging structured fact-checking frameworks, such as large-scale knowledge graphs, enhances the efficiency of factuality assessments. Additionally, (Kumar, 2024) highlights how unified factuality frameworks provide standardized evaluation methods, offering a robust approach to measuring improvements in LLM reliability.

This study will explore how incorporating Community Notes into LLM inference pipelines can improve factual accuracy while ensuring consistency across diverse topics and contexts.

Data: Twitter's Community Note dataset (see Community Notes Guide)

Methods: Benchmark accuracy using established techniques from recent LLM evaluation frameworks, comparing LLM performance with and without Community Notes integration. One approach involves designing a verification step to the LLM query process, where LLM-generated responses are cross-referenced with structured fact-checking databases before finalizing output.

Additional Exploration

Given the available time, we may also explore the potential of Community Notes in enabling proactive misinformation detection and enhanced content moderation. Key opportunities include:

Misinformation Prediction: Leveraging Community Notes data to predict tweets likely to contain misinformation, enabling earlier intervention.

Valuable Note Prediction: Developing models to identify which notes are most likely to be rated as helpful, accelerating their visibility.

Hybrid Fact-Checking: Evaluating the effectiveness of combining Community Notes with professional fact-checking databases to assess misinformation more accurately.

3 Timeline and Team

Timeline

Week 1 : Proof of concept and Set up project environment, data pipeline; Continue literature review (Feb. 25-Mar 3)

Week 2 : Collect datasets; Perform data cleaning and preprocessing; Continue literature review(Mar 4–Mar 10)

Week 3 : Perform exploratory data analysis (EDA), topic modeling and metadata analysis; Finalize research scope and outline (Mar 11–Mar 17)

Week 4 : Accuracy evaluation; Start on LLM integration (Mar 18–Mar 24)

Week 5 : Summarize Preliminary Experimental; Prepare Midterm Report (Mar 25–Mar 31)

Week 6 : Continue with LLM integration (Apr 1–Apr 7)

Week 7 : Evaluate LLM output and perform error analysis (April 8 to April 14)

Week 8 : Evaluation, interpretation of results; Initial preparation of the report and presentation (April 15-April 21)

Week 9 : Continue with findings; Revise paper and presentation (Apr 22–Apr 26)

Week 10 : Summarize finding; Polish presentation (Apr 27–Apr 28)

Week 11 : Presentation (Apr 29–May 5)

Week 12 : Finalize all project deliverables (May 6–May 12)

Team and Responsibility

Team	Responsibility
Bhirajaya Bhumimars	Data Processing Literature Review Fact-Checking Techniques
Supakarn Bunlert	Software Development Implementation Optimization Validation and Testing
Yunkai Li	Data Collection Topic Modeling Data Analysis Methodology Research

References

Matthew R. Allen, Nimit Desai, Aiden Namazi, Eric Leas, Mark Dredze, Davey M. Smith, and John W. Ayers. 2024. Characteristics of x (formerly twitter) community notes addressing covid-19 vaccine misinformation. *JAMA*, 331(19):1670–1672.

Natalia Aruguete, Flavia Batista, Ernesto Calvo, Matías Guizzo Altube, Carlos Scartascini, and Tiago Ventura. 2023. Reducing misinformation: The role of confirmation frames in fact-checking interventions.

Nadav Borenstein, Greta Warren, Desmond Elliott, and Isabelle Augenstein. 2025. Can community notes replace professional fact-checkers? *Preprint*, arXiv:2502.14132.

Peter Dizikes. 2018. Study: On twitter, false news travels faster than true stories. *MIT News*. Available at: `https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308` (Accessed: Feb 22th, 2025).

Chiara Patricia Drolsbach, Kirill Solovev, and Nicolas Pröllochs. 2024. Community notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(7):pgae217.

Clare Duffy. 2025. Meta is getting rid of fact checkers. zuckerberg acknowledged more harmful content will appear on the platforms now. *CNN Business*. Available at: `https://www.cnn.com/2025/01/07/tech/meta-censorship-moderation/index.html` (Accessed: Feb 22th, 2025).

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Preprint*, arXiv:2406.09279.

Uku Kangur, Roshni Chakraborty, and Rajesh Sharma. 2024. Who checks the checkers? exploring source credibility in twitter's community notes. *Preprint*, arXiv:2406.12444.

Sachin Kumar. 2024. Openfactcheck: A unified framework for factuality evaluation of llms. *AP News*. Available at: `https://medium.com/@techsachin/openfactcheck-a-unified-framework-for-factuality-evaluation-of-llms-d88f2946ca94` (Accessed: Feb 22th, 2025).

Xiaoze Liu, Feijie Wu, Tianyang Xu, Zhuo Chen, Yichi Zhang, Xiaoqian Wang, and Jing Gao. 2024. Evaluating the factuality of large language models using large-scale knowledge graphs. *Preprint*, arXiv:2404.00942.

Barbara Ortutay. 2024. Report says crowd-sourced fact checks on x fail to address flood of us election misinformation. *AP News*. Available at: `https://apnews.com/article/x-musk-twitter-misinformation-ccdh-0fa4fec0f703369b93be248461e8005d` (Accessed: Feb 22th, 2025).