Trust, Topics, and Echo Chambers: Analyzing Community Notes on X

Team members info:

Omer Barlev, omer.barlev@mail.huji.ac.il, barlevo

Yair Ben Eliyahu, yair.ben@mail.huji.ac.il, yairbe

Avihu Brown, avihu.brown@mail.huji.ac.il, ab07121993

Yair Ben Menachem (Miluim), yair.benmenachem@mail.huji.ac.il, Yairbenmena

Problem description:

Community Notes on X (formerly Twitter) are a crowdsourced tool for adding context to potentially misleading posts. However, not all notes are judged equally helpful, and little is known about what determines whether a note succeeds in being accepted.

Our project aims to address these questions. Specifically, we ask:

What characteristics distinguish Helpful from Not Helpful notes? Do certain features, such as trustworthy sources or missing context, play a decisive role in note acceptance? Which topics dominate the fact-checking ecosystem, and how do real-world events influence note creation? Do contributors organize into distinct communities, and if so, do these groups reflect healthy diversity or echo-chamber polarization? To what extent does user activity follow the typical “long-tail” pattern, what does this imply about influence and potential manipulation (e.g., bots)?

By formulating these questions, we position our study within a broader concern: whether Community Notes serve as a genuine corrective to misinformation or risk reproducing the same divisions that already shape online discourse.

Data

For this project we will use the publicly available Community Notes dataset provided by X (formerly Twitter). The data is released in five separate collections: Notes, Ratings, Note Status History, User Enrolment, and Note Requests:

- **Notes:** Contains a table representing all notes. 900 MB in size and 1.9 million records.
- **Ratings:** Contains a table representing all ratings. 25.6 GB and 13 million records.
- **Note Status History:** Contains a table with metadata about notes including what statuses they received and when. 515 MB and 2 million records.
- **User Enrolment:** Contains a table with metadata about each user's enrolment state 120 MB and 1.1 million records.
- **Note Requests:** Contains a table representing all requests for a Community Note, which can be submitted by any account on X – 600MB and 6.9 million records.

Data format - The datasets come in tsv files. Detailed format of the fields of each dataset can be found in the link attached above.

Introduction - How a Community Note is added

[1]A Community Note starts when a note author adds context to a post, often with explanations and sources (Figure 1). Other contributors then rate the note as *Helpful* or *Not Helpful*. Instead of simple majority voting, an algorithm requires agreement across people with different viewpoints - a "bridging" rule that ensures broad trust. Once this threshold is met, the note becomes visible under the post; continued ratings can later remove it if it loses support.

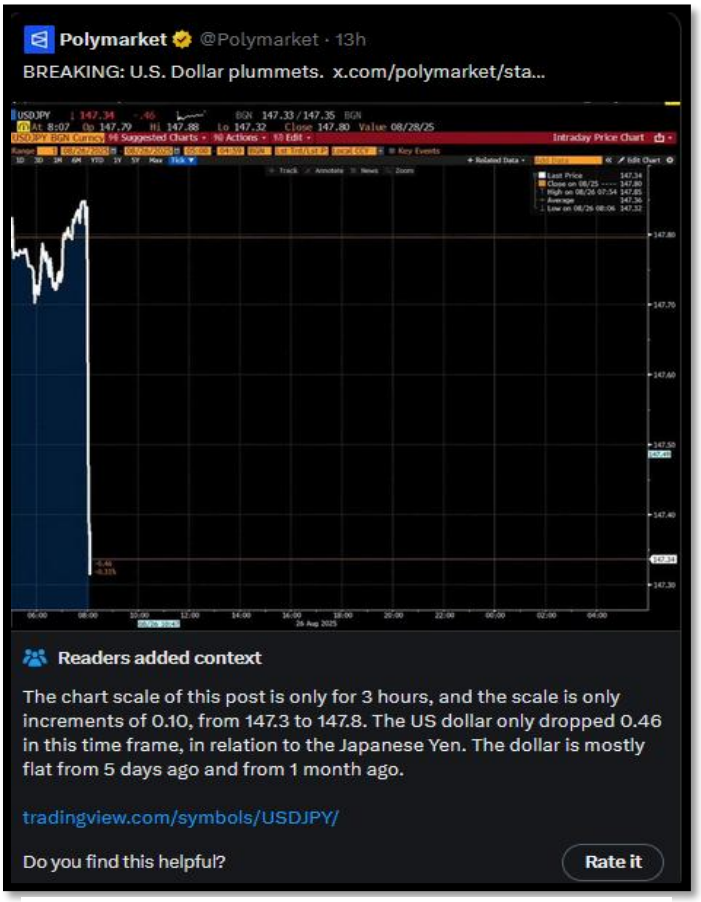

Figure 2 – Example of a Community Note

[1] https://communitynotes.x.com/guide/en/about/introduction

Introduction - EDA and feature analysis

Before diving into models, we wanted to get a feel for what Community Notes actually look like. The first surprise (Figure 2) : out of over 1.2 million notes, only about 12.5% ever reach a decision - and within that small pool, just 9% are crowned Helpful while 3.5% are dismissed as Not Helpful. In other words, the vast majority of notes float in limbo, never gathering enough attention to matter.

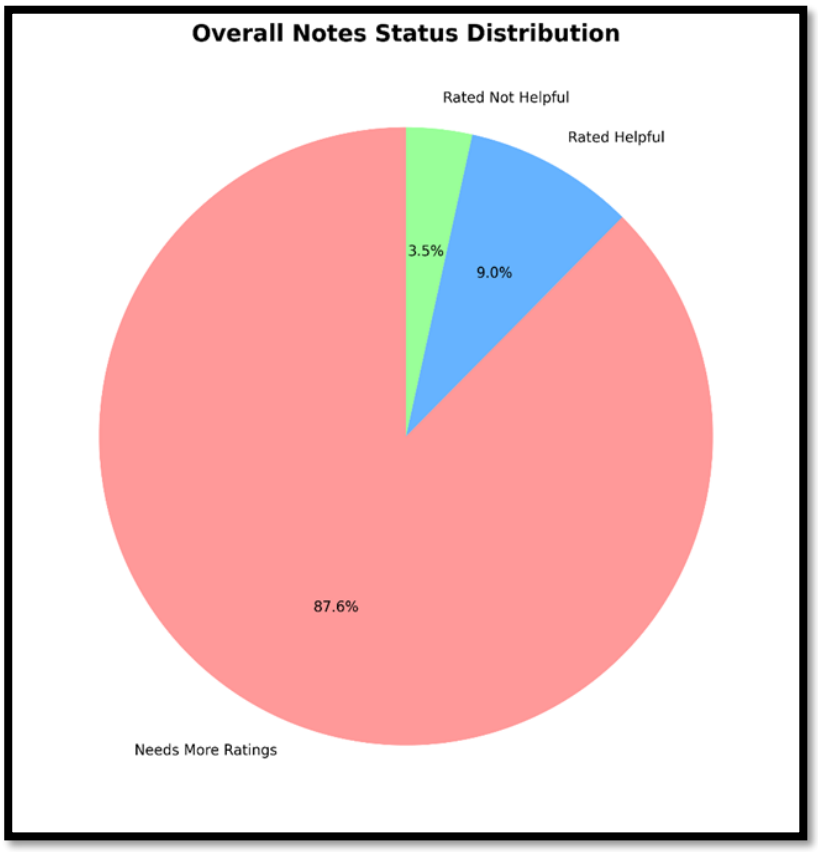

Figure 2 – Most notes status is waiting for more ratings

Looking closer at the features authors attach to their notes (Figure 3), the landscape is far from balanced. An overwhelming 85% proudly declare they cite trustworthy sources. Nearly half claim to add missing context, and over 40% argue they correct a factual error. Meanwhile, other flags like satire or manipulated media barely register. It's almost as if everyone wants to be seen as the sober voice of reason - but of course, not all notes are judged that way.

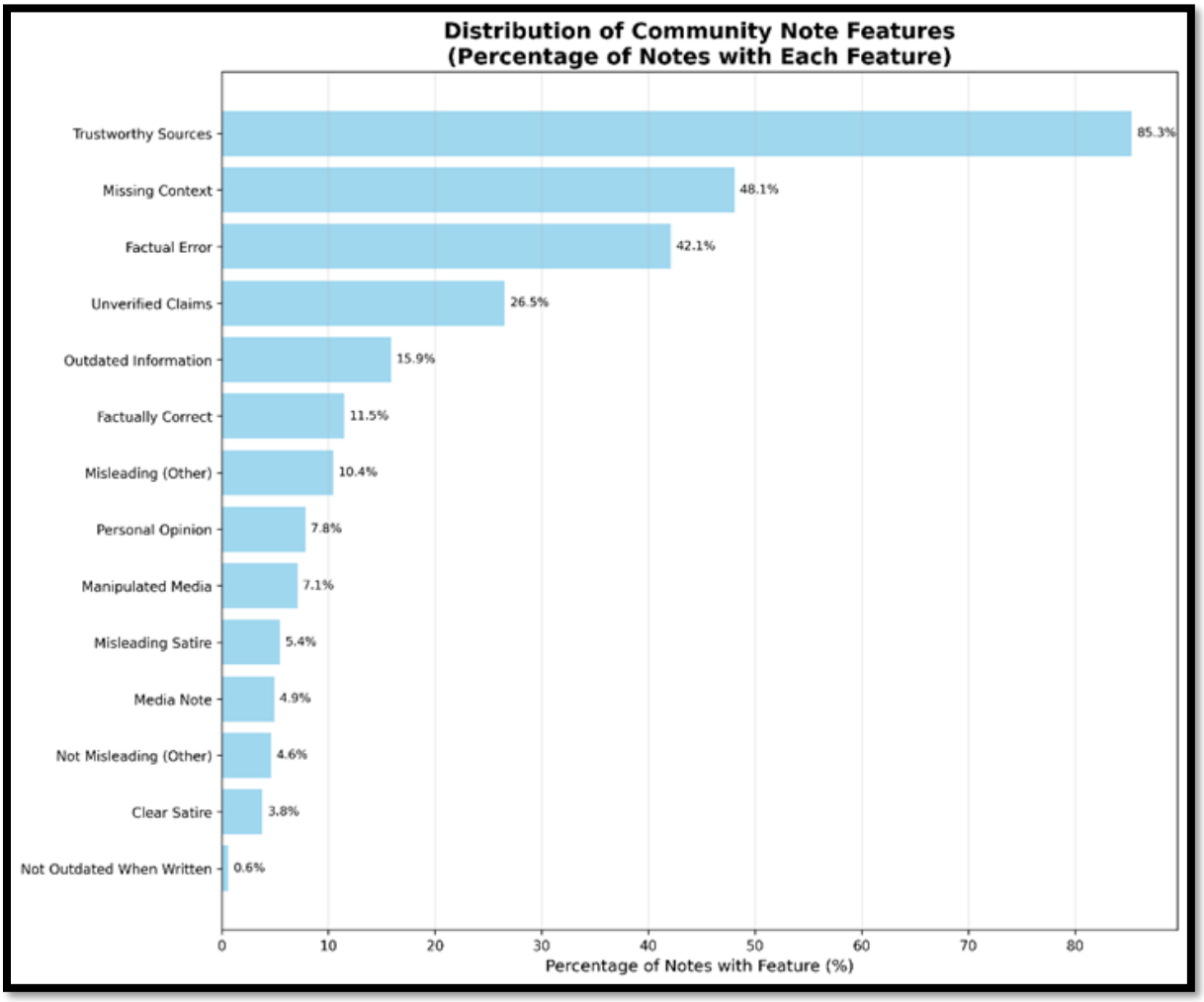

Figure 3 – Most notes are with sources/references

The real story emerges when comparing Helpful to Not Helpful notes. Here, the “trustworthy sources” label becomes the deal-breaker (Figure 4): almost 95% of Helpful notes include them, compared to only ~70% of Not Helpful ones. That 25-point gap tells us something powerful - credibility isn’t just nice to have, it’s the ticket to being heard. Features like adding context or pointing out factual errors also push notes toward acceptance, but sources stand out as the sharpest dividing line.

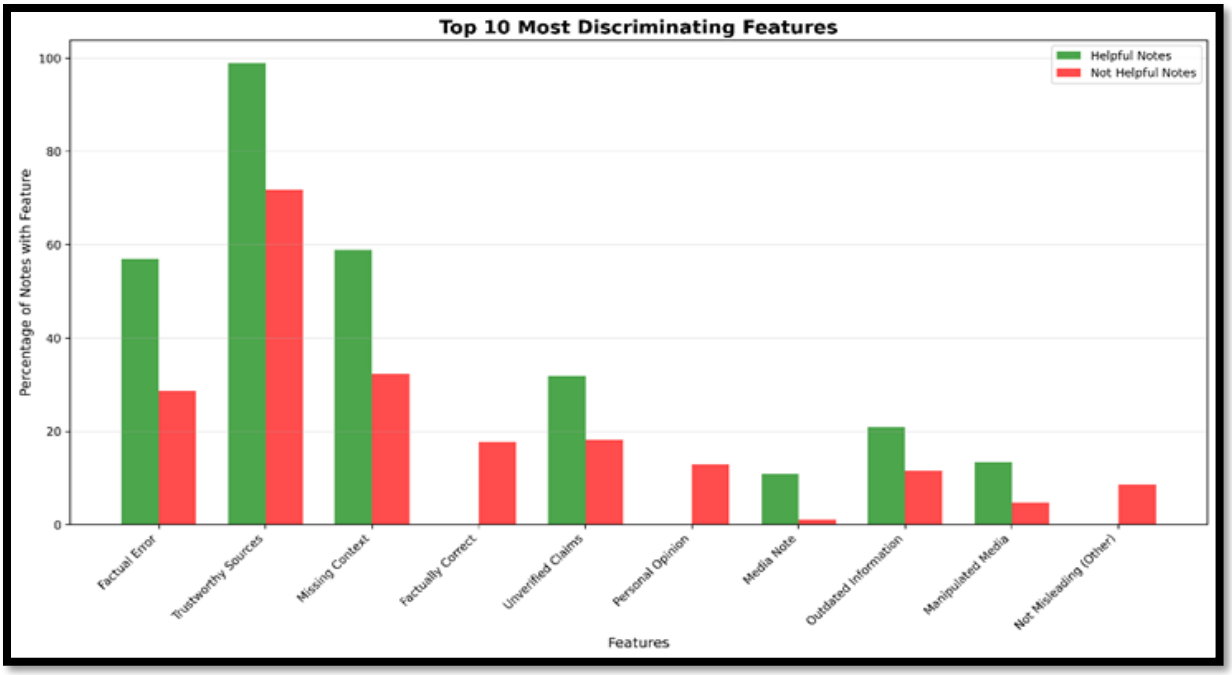

Figure 4 – Features of Helpful vs Not Helpful Notes

This early exploration gave us more than just descriptive stats. It hinted at the DNA of a successful note and set the stage for the rest of our analysis. The message was clear: in the noisy marketplace of online fact-checking, trust is currency. And that insight made us eager to see how topics, time, and contributor communities shape the bigger story.

Topic Classification of Community Notes

Motivation:

We aimed to classify Community Notes into topical categories to uncover what kinds of posts attract fact-checking and whether topical differences influence note helpfulness. Our categories balanced two goals: capturing the high-salience conflicts that dominate global discourse (Ukraine, Gaza, Syria, Iran, China-Taiwan, etc.), and including broad societal issues like Politics, Health/Medical, Scams, and Climate. This mix reflects both the "headline" geopolitical crises and the everyday issues that shape misinformation online.

Data Preparation:

We worked with 1.2M notes and first applied language filtering to focus on English-only content, removing empty or very short entries. Next, we created dictionaries of *seed keywords* for 15 categories. These dictionaries were deliberately rich: conflict categories contained place names, political figures, and event terms, while domains like Health included pandemic terms, drug names, and institutions (e.g., CDC, WHO). This ensured coverage of both explicit and subtle references. Using these seeds, we were able to label ~73% of the notes as a starting point.

Method:

We then trained a TF-IDF + Logistic Regression pipeline. The TF-IDF vectorizer was configured with:

- 50,000 unigram and bigram features,
- English stopwords removed
- terms appearing in fewer than 2 notes discarded
- terms appearing in more than 95% of notes ignored

The Logistic Regression classifier used:

- lbfgs solver
- max iteration is 1000 to ensure convergence
- we trained the model to look at one topic at a time ("one-vs-rest")
- since some topics have way more notes than others, we told the model to "pay extra attention" to the smaller classes so they wouldn't be ignored

This setup gave us an interpretable yet scalable model that could process ~8,400 notes per second.

Evaluation Criteria:

With no gold-standard labels available, we measured success in three ways:

1. Accuracy against seed labels: an 80/20 stratified split yielded 82.8% accuracy, comfortably beating random and naive baselines. Always predicting the majority class

("Politics") gives only ~16% accuracy, while random guessing yields ~7–12% depending on weighting. We also tested a keyword-matching baseline that used our seed dictionaries directly to assign topics; this reached ~65% accuracy but failed on ambiguous or nuanced notes. Our TF-IDF + Logistic Regression model clearly outperformed it by capturing richer contextual patterns beyond simple keyword hits.

2. External event validation (Figure 5): topic spikes aligned strongly with real-world events (e.g., Gaza notes after October 7, politics notes during U.S. debates).
3. Manual review & confidence (Figure 6): mean confidence was 45.4%, with geopolitical conflicts classified more reliably than broad categories like Politics.

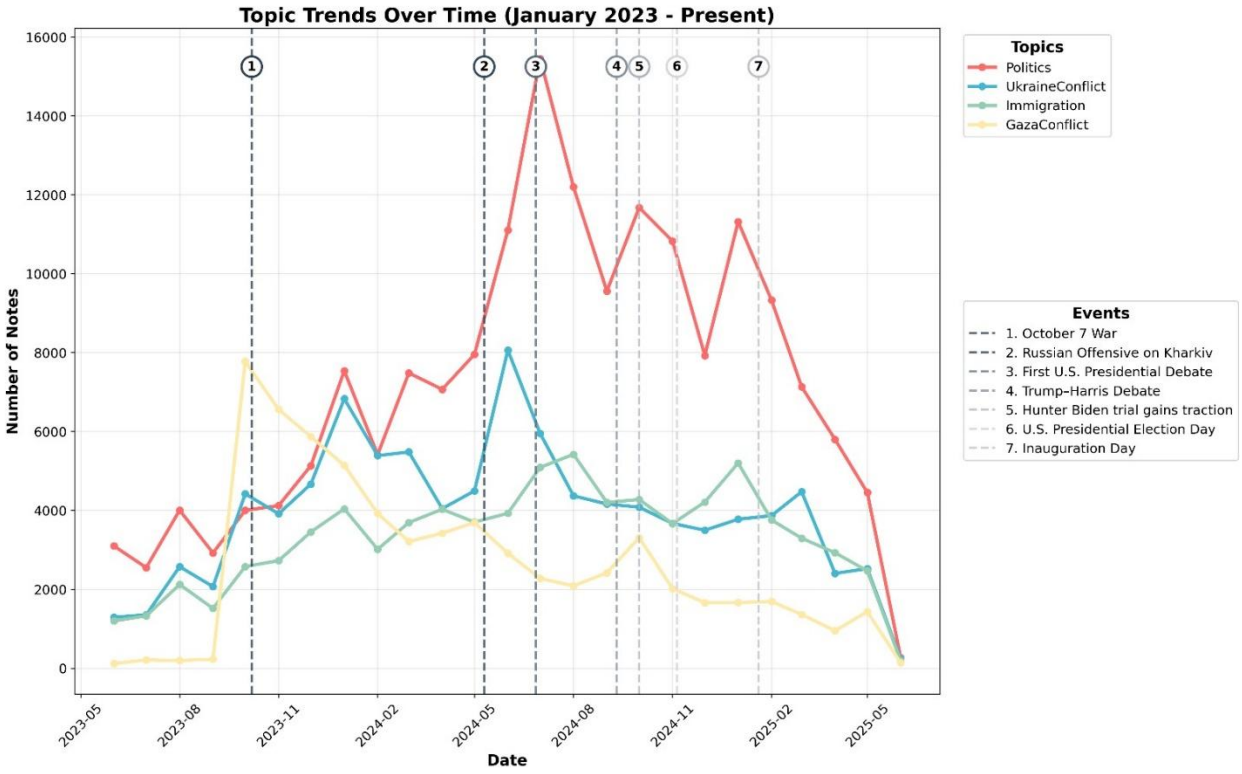

Figure 5 – Sanity check of the topic classifier: Spikes aligned well with real events

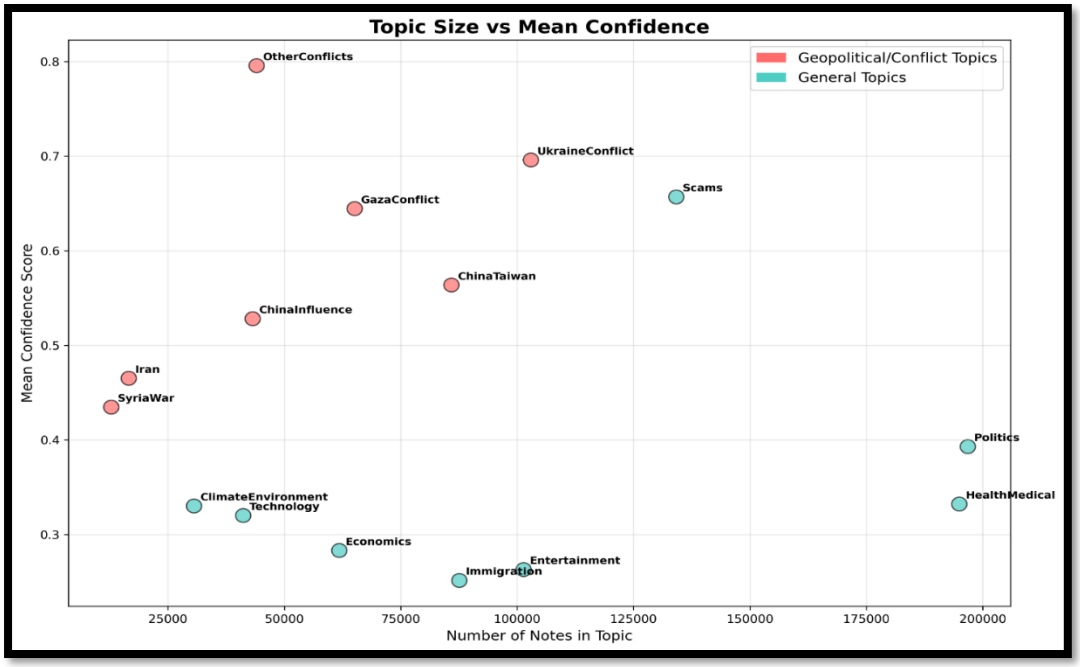

Figure 6 – More 'subtle' topics get higher confidence score

Results:

The classifier surfaced clear distributions (Figure 7): Politics (16.1%) and Health/Medical (16.0%) dominated, followed by Scams (11.0%), Ukraine Conflict (8.4%), and Entertainment (8.3%). Geopolitical issues accounted for ~30% of notes. Importantly, our analysis revealed a paradox (Figure 8): *conflict-related notes were less often rated Helpful*. This likely reflects X's design, which requires agreement across users with "different perspectives" - a condition that is hardest to achieve on divisive topics like Gaza or Ukraine.

Visualization:

We used discriminating-topic bar charts to highlight contrasts between Helpful and Not Helpful notes (see figure). While ~40% of Health or Immigration notes were rated Helpful, conflict-heavy topics skewed strongly toward Not Helpful, making polarization visible at a glance.

Impediments:

The biggest limitation was the lack of human-labeled ground truth, forcing reliance on seed-based supervision. This left ~27% of notes unclassified. Still, the hybrid approach of keyword seeding, interpretable models, and validation against external events gave us both credibility and insight.

Another limitation was our computation resources, which enforced us to train a light-weight model rather than fine-tuning an LLM.

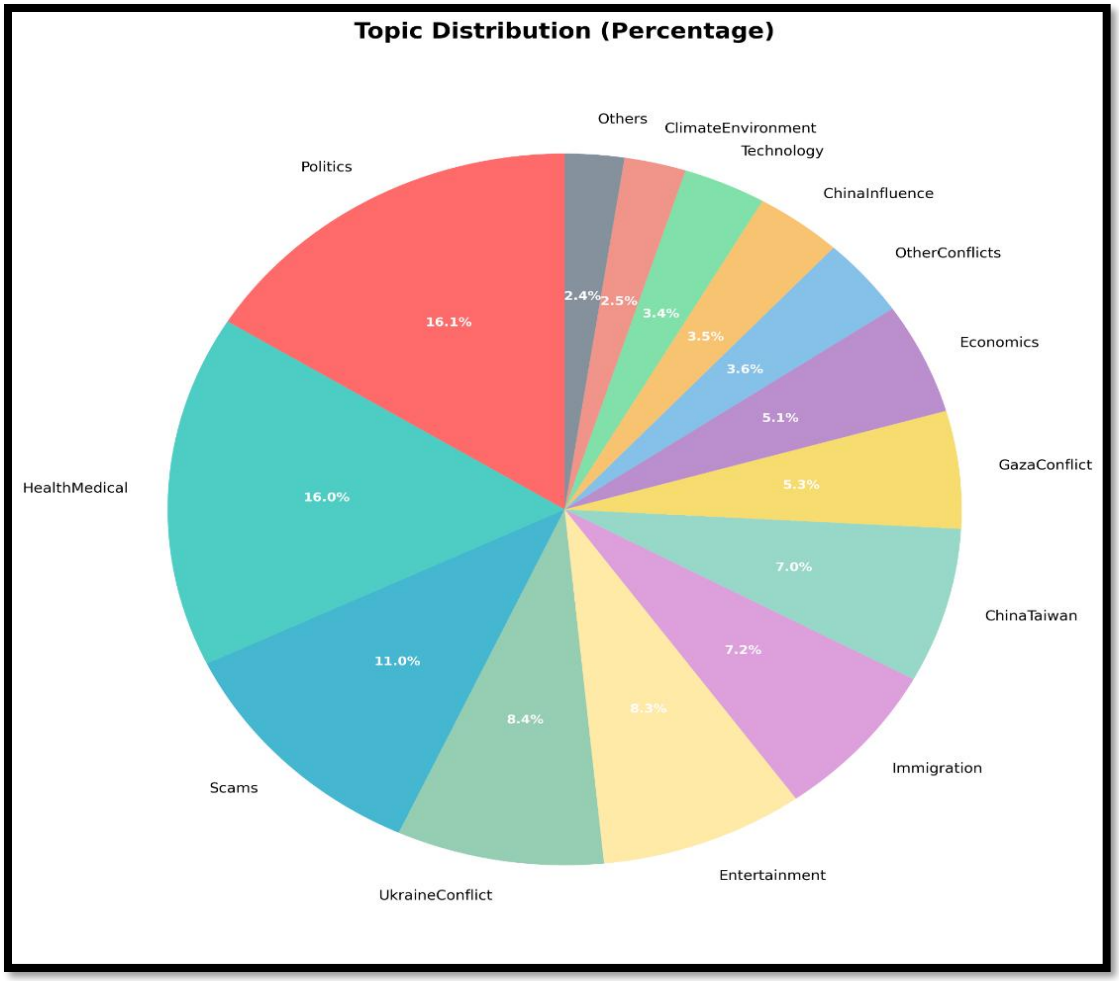

Figure 7 – Distribution of topics assigned by the classifier model

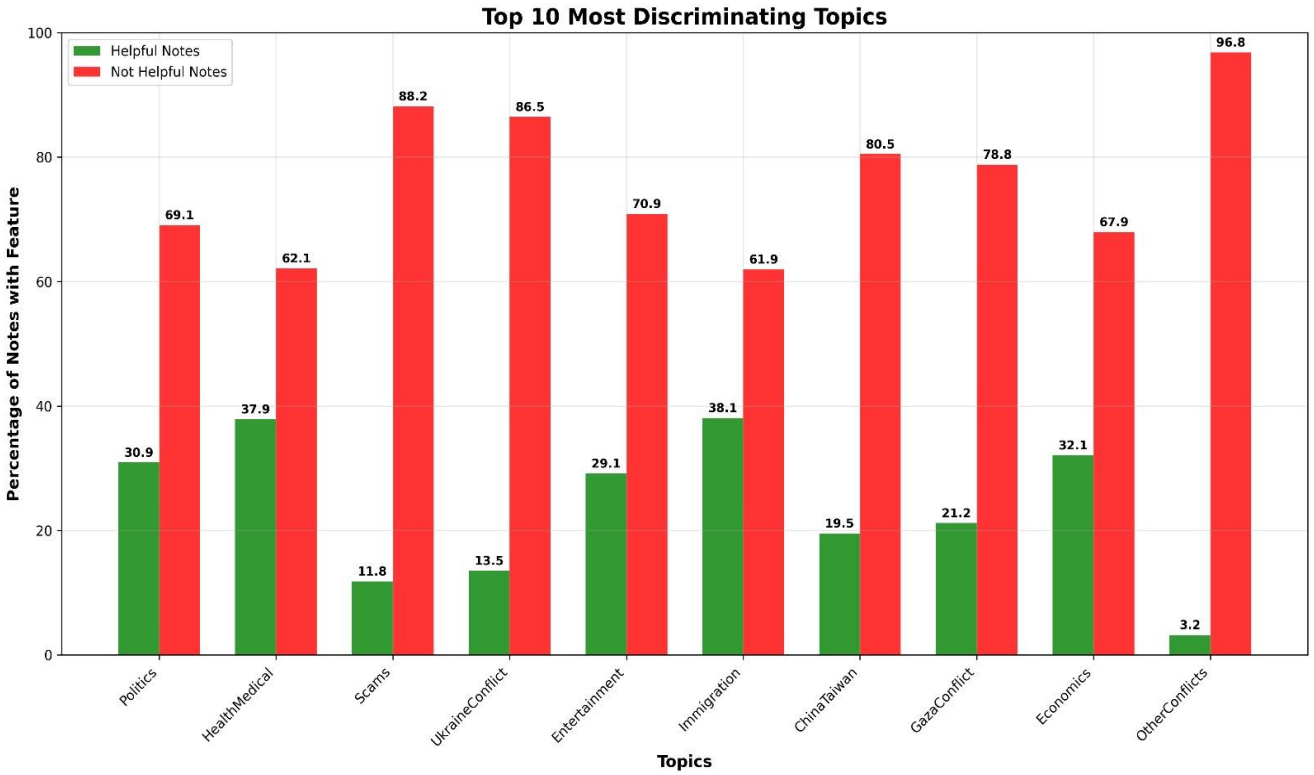

Figure 8 – Notes Helpfulness by Topic

Clustering of Contributors

Motivation:

Community Notes are not just about the notes themselves - they are also about the *people* behind them. We wanted to know: do contributors naturally split into distinct communities? And if so, are these communities driven by topical specialization (e.g., politics junkies vs. health experts) or do they reflect the same echo chambers that dominate the broader platform?

Method:

We built a contributor–contributor graph where each node is a contributor, and edges connect contributors who rated the same note as Helpful. This design captures *agreement in action* rather than mere co-occurrence. To detect structure, we applied the **Louvain community detection algorithm**, which partitions the graph by maximizing modularity. Louvain was chosen for its scalability - critical for a graph with hundreds of thousands of users - and because it produces interpretable communities. Parameters were kept at defaults, but we pre-filtered extremely low-activity users to reduce noise and improve stability.

Evaluation Criteria and Setup:

Since there is no "ground truth" for contributor communities, we evaluated success by:

1. **Modularity score:** Louvain achieved modularity >0.6, indicating strong community structure.
2. **Topical coherence:** We mapped each community's notes back to our topic classifier. Coherence was measured by entropy/diversity scores, where low entropy indicated specialization.

Results:

The Louvain algorithm uncovered multiple communities of varying size, from small niche clusters to mega-communities. Most communities showed broad topical engagement (diversity ~0.9), but a few stood out as specialists (Figure 11):

- **Politics specialists** (Communities 3 & 7), responsible for massive volumes of notes centered on U.S. debates and elections.
- **Health/Medical specialists** (Communities 0, 1, 9, 10), consistently focused on vaccines, pandemics, and health misinformation.
- **Conflict-oriented clusters,** especially around Ukraine and Gaza, which were smaller but disproportionately active during global crises.

Visualization:

The community scatter plot (Figure 9) shows clear clusters, each dot a contributor, each color a community. Radar plots of the top 6 communities (Figure 10) highlight topical focus: Politics-heavy groups peak sharply on that axis, while Health specialists display the opposite profile. The topic leadership heatmap (Figure 11) makes dominance even clearer - a few communities consistently lead in shaping discourse on divisive topics. Finally, activity distributions (Figures 12-13) reveal a classic *long-tail*: most contributors rate a handful of notes, but a tiny hyperactive minority drive the bulk of the activity.

Interpretation:

This paints a sobering picture. While most contributors are generalists, the ecosystem is tilted by a few dominant communities and hyperactive users. Fact-checking is therefore not evenly distributed but shaped by concentrated groups with specific interests. On divisive topics, this structure risks reinforcing the very echo chambers Community Notes was meant to break - a resonance box where agreement is hardest to achieve and polarization is amplified.

Impediments:

A key challenge was the sheer scale of the dataset - millions of notes and contributor interactions. To handle this efficiently, we leveraged GPU acceleration, which allowed us to process data and run models at scale without bottlenecks. Another limitation was interpretability: while Louvain identifies clusters, labeling them required careful cross-referencing with topic classification outputs. Despite these hurdles, the clustering provided a powerful new lens into how contributor dynamics shape the platform.

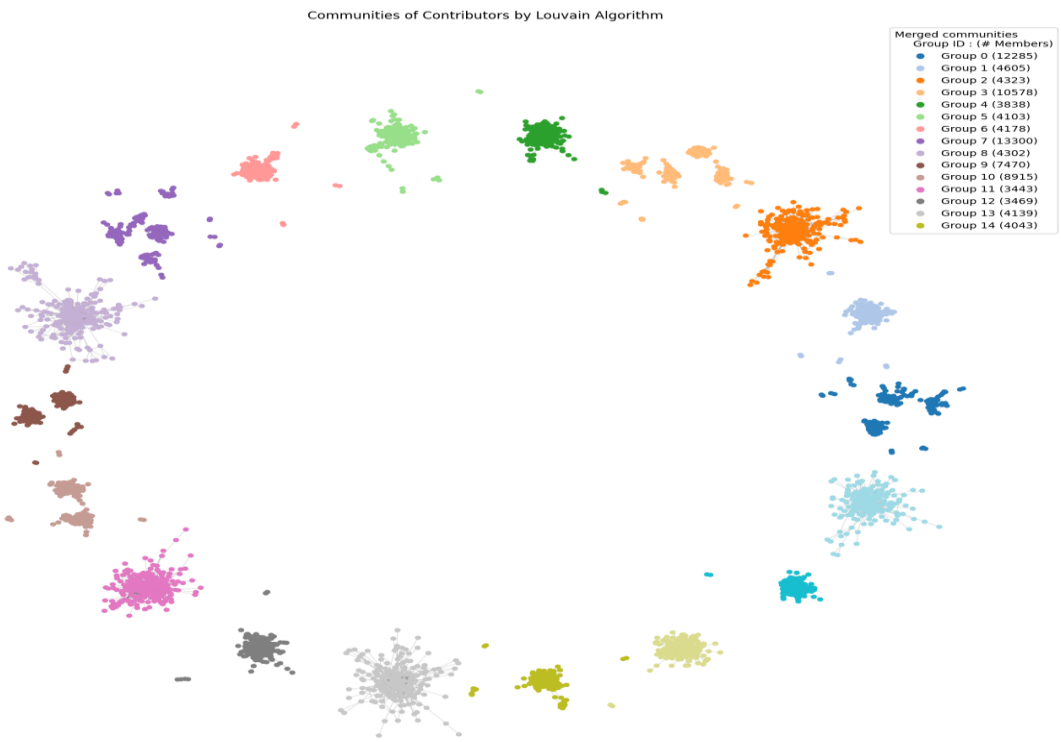

Figure 9 – Initial result of Louvain algorithm, 15 communities varying in size

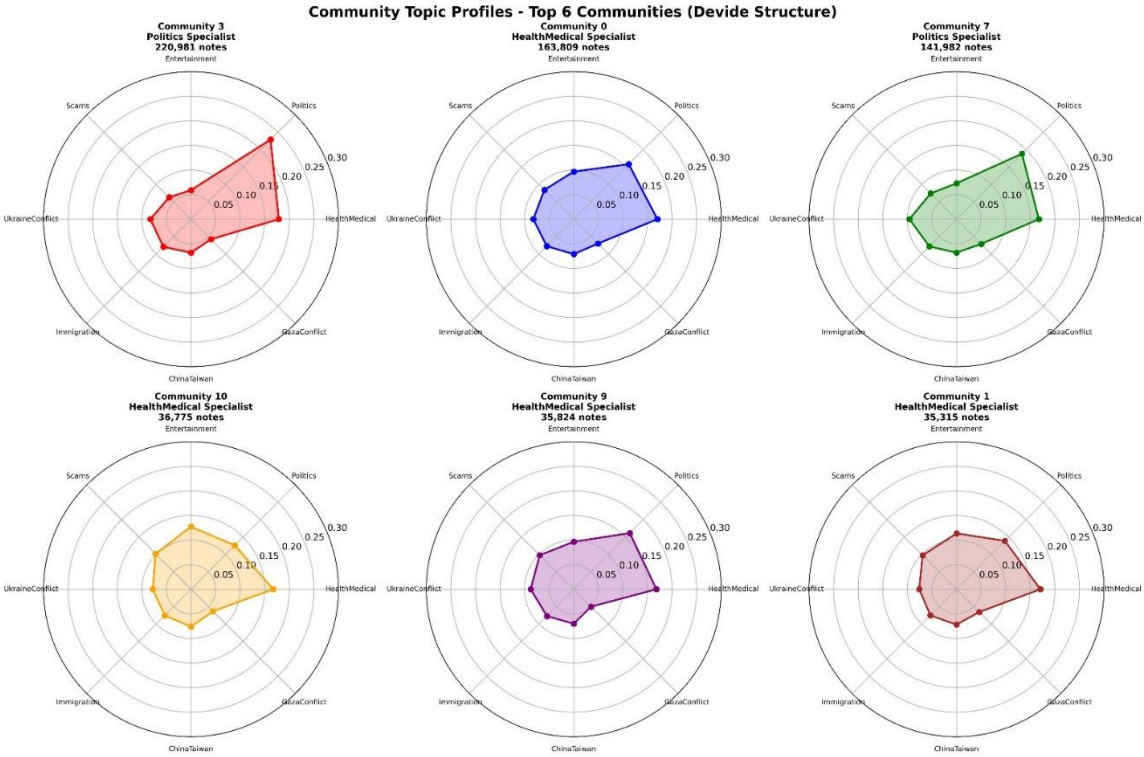

Figure 10 – Top 6 Communities topic orientation

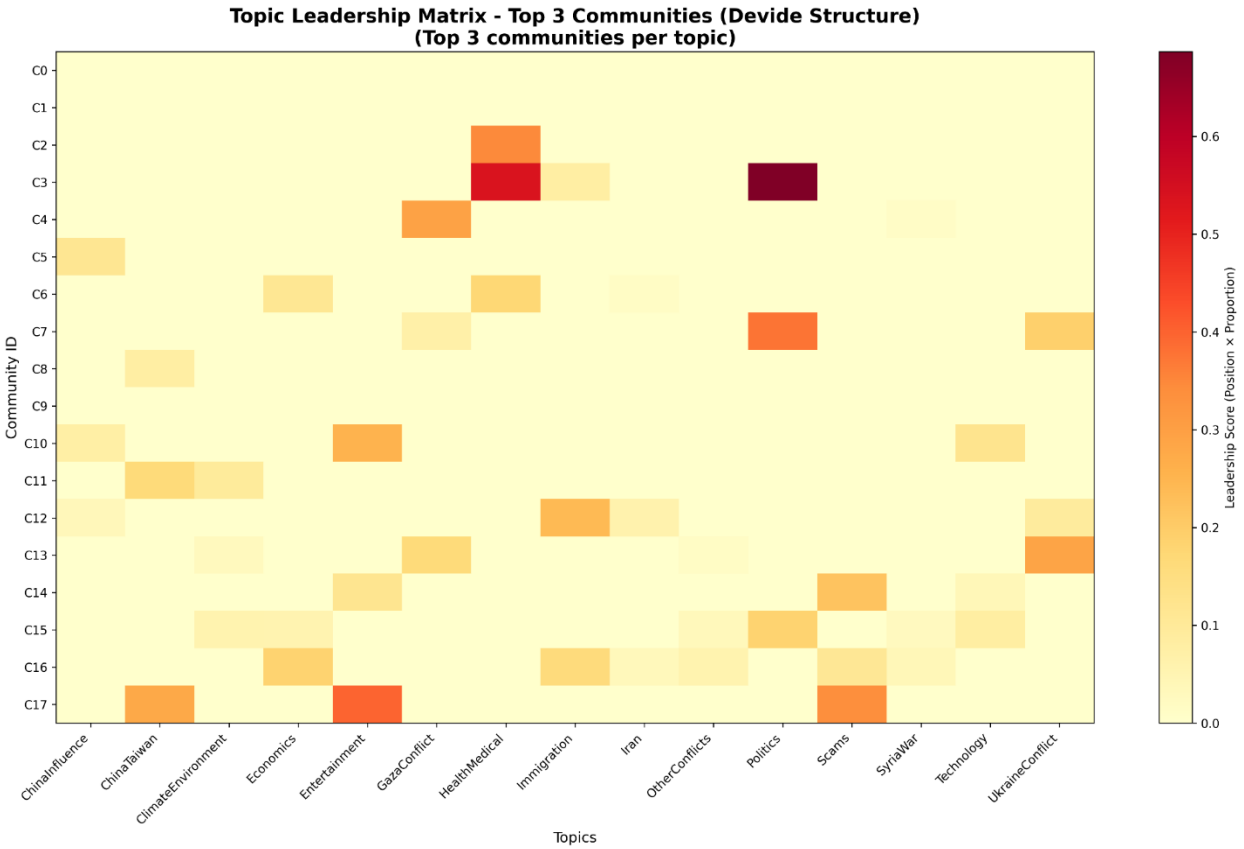

Figure 11 – Top 3 communities per topic

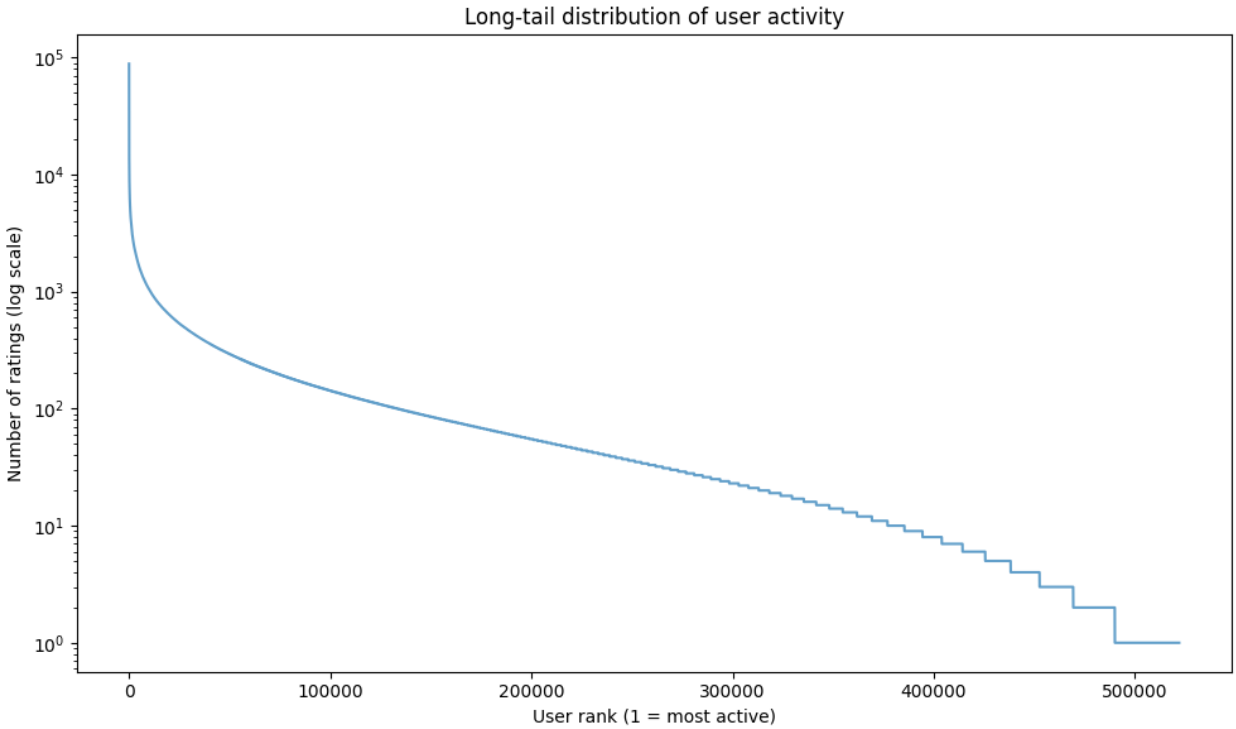

Figure 12 – Small portion of contributors make most ratings

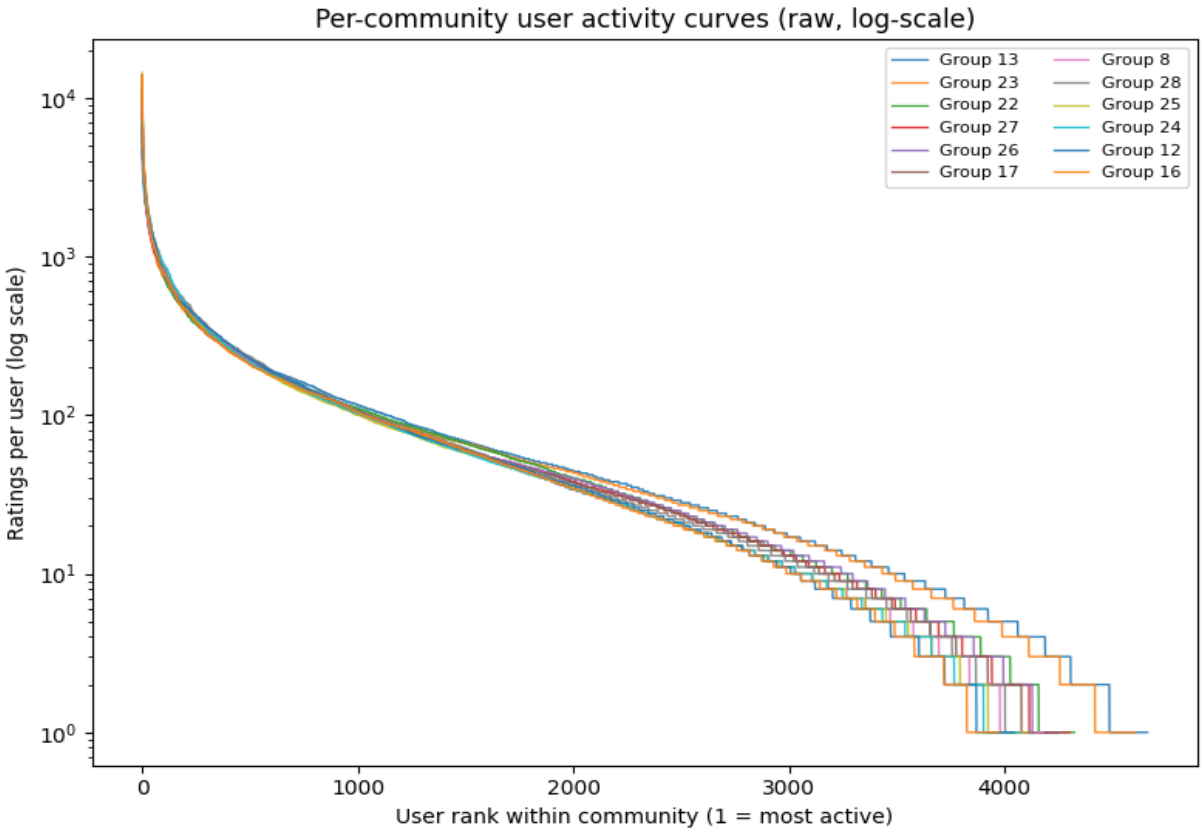

Figure 13 – Different communities show similar long-tail user-ratings

Future Work

Our project focused on understanding topical patterns, contributor clustering, and predictors of note helpfulness, but many exciting directions remain. One major open question is the *real-world impact* of Community Notes on users and tweets themselves. Do tweets that receive notes lose traction in terms of likes, shares, or follower growth? Do notes actually reduce misinformation spread, or do they sometimes deepen polarization by attracting more heated engagement? Linking note assignments to downstream engagement metrics would allow us to measure influence at scale - moving from helpfulness as a proxy toward direct evidence of behavioral change on the platform.

Another frontier is the rise of AI-assisted note writing. X has already begun experimenting with *Grok*, its in-house AI, to generate or suggest notes. This raises critical questions about quality, bias, and trust. How do AI-generated notes compare to human-authored ones in accuracy, clarity, and legitimacy? Will contributors accept or reject them differently? Studying this intersection of AI and crowdsourced fact-checking could shape the future of Community Notes, for better or worse.

Conclusion

Community Notes is a bold experiment in crowdsourced fact-checking. Our analysis showed that trustworthy sources are the clearest predictor of success: notes backed by credible evidence were far more likely to be rated Helpful, underscoring the value of verifiability in online discourse.

Topic classification revealed that divisive domains dominate - politics, health, and geopolitical conflicts surge during major events, yet these very areas struggle most to achieve consensus. This tension illustrates how polarization can limit the system’s ability to function where it is most urgently needed.

Clustering of contributors added another dimension: while many participants engage broadly, we found specialist communities centered on politics or health, and a long-tail pattern where a tiny minority of hyperactive users drive disproportionate influence. This raises important questions about diversity, expertise, and the potential role of bots or coordinated actors.

Together, these findings highlight both the promise and fragility of Community Notes. The platform can elevate timely, credible information, but its consensus-driven design makes it most vulnerable in exactly the spaces where truth is contested. As X experiments with new tools, including AI-assisted notes, the challenge will be whether Community Notes can bridge divides - or whether it risks amplifying them.