SAN DIEGO STATE UNIVERSITY

Capstone Project – BDA 600 – Fall 2024

Sharmaine Lindahl (Individual Project)

Presentation Link: Community Notes on X
Website Link: Community Notes Performance on X

Table of Contents

Abstract

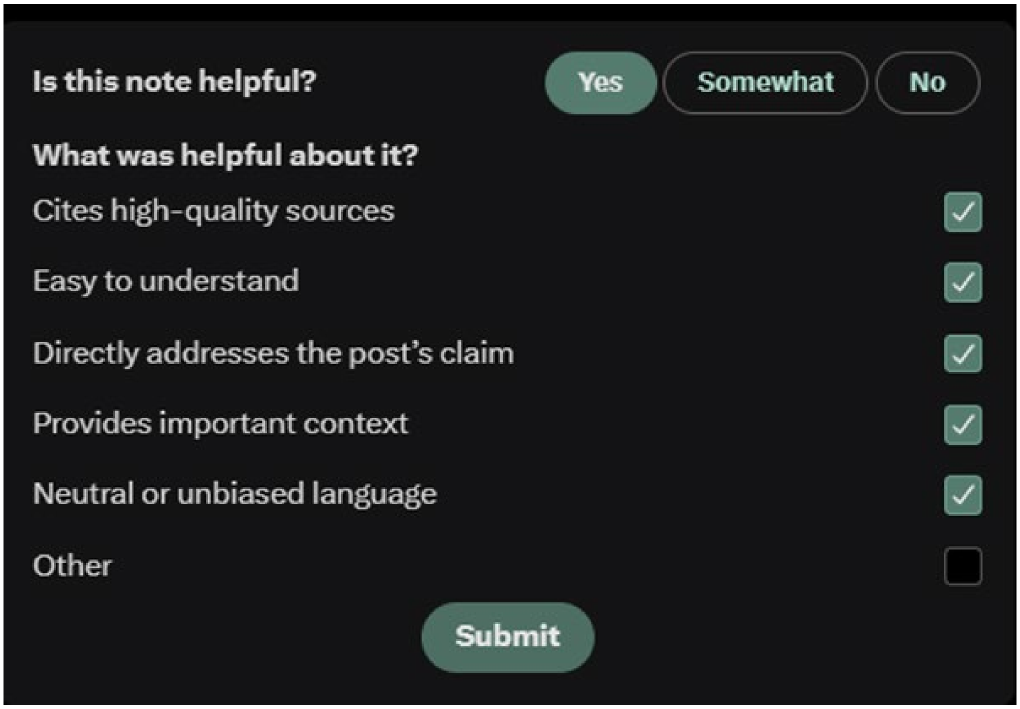
Misinformation on social media is spreading faster than ever, and people rely on these platforms more than they should for credible information (Vosoughi, Roy, & Aral, 2018). During critical events like the Ukraine-Russia war and the Israel-Hamas conflict, the need for accurate information is amplified. This research examines how well the Community Notes system on X (formerly Twitter) combats misinformation by analyzing its timeliness and effectiveness. Introduced widely after Elon Musk acquired Twitter, Community Notes relies on volunteer contributors to add context to posts, aiming to clarify misleading claims. Using public data from October 2023 to November 2024, this study investigates how quickly and effectively these notes reach visibility. Using data analysis tools, the results show that helpful notes related to the Israel-Hamas conflict took an average of 27 hours to display to the public, while notes about the Russia-Ukraine war took 17 hours. These averages, however, mask a deeper issue: while notes are displayed relatively quickly once rated, the majority remain in 'Needs More Ratings' for extended periods, creating the illusion of efficiency. While visible notes are displayed quickly, this represents only a small subset of submissions. The majority remain in 'Needs More Ratings' indefinitely, with survival analysis revealing a low probability of transitioning to visibility. This bottleneck creates the illusion of efficiency, masking systemic delays. Helpful and publicly displayed notes comprise a very small percentage of the overall notes submitted by contributors. An analysis on the note categories revealed that the majority of submitted notes for both conflicts remained in the “Needs More Ratings” category and therefore hidden from public view. These notes are awaiting ratings from contributors to help them reach the ratings score threshold calculated by the Community Notes algorithm in order to be displayed to the public. A Kaplan-Meier survival analysis found that notes in the "Needs More Ratings" status persist for extended periods, with median survival times of 384 days for Israel-Hamas posts and 292 days for Russia-Ukraine posts, indicating severe delays in attaining visibility. Furthermore, notes deemed helpful and displayed to the public tended to have less negative sentiment compared to those still awaiting ratings, suggesting that controversial topics may face more challenges in reaching the public. Aligning with Goal 16 of the UN Sustainable Development Goals, this study offers insights and future recommendations to enhance the efficacy of Community Notes in curbing misinformation on one of the world's most influential social media platforms. Future research suggestions include continuing to optimize the Community notes algorithm for faster response times, functionality to address the bottleneck of notes remaining in Needs More Ratings for extended periods of time, and incentivizing the recruitment of Community Notes participants to have a vast network of writers and raters of submitted notes. These findings contribute valuable insights on the current performance as X continues to refine the Community Notes System to mitigate the spread of misinformation on its platform.

Introduction

Social media has become the go-to source for information, often replacing traditional media for millions of users worldwide. While this accessibility has its advantages, it also accelerates the spread of misinformation. This is particularly dangerous during high-stakes events like wars, where public perception can influence international relations and policymaking. Research has shown that false information travels faster and reaches more people than truthful information on social media (Vosoughi, Roy, & Aral, 2018). For conflicts like the Ukraine-Russia war and the Israel-Hamas conflict, misinformation can distort public opinion and escalate tensions.

In response to the challenge of misinformation, platforms typically deploy in-house teams for moderation and fact-checking. However, X has taken a different approach. After acquiring Twitter in 2022, Elon Musk expanded the Birdwatch program, renaming it Community Notes and opening it to the public. This system relies entirely on volunteer contributors to fact-check and contextualize posts. While innovative, the program has faced heavy criticism. Investigations, such as those by the Center for Countering Digital Hate (2023), reveal gaps in its coverage and delays in providing corrections. These shortcomings raise questions about its reliability and impact.

This study focuses on how Community Notes addresses misinformation specifically related to the Ukraine-Russia war and the Israel-Hamas conflict. Using data from October 2023 to November 2024, we examine the timeliness and sentiment of the notes published. Our analysis is limited to English-language notes, though future research could expand to other languages for broader insights. By evaluating Community Notes' performance, this research contributes to the growing discourse on combating misinformation in the digital age and highlights areas for improvement to make fact-checking tools more effective and accessible. Figures 1 and 2 below show the contributor ratings system choices for rating a note helpful or unhelpful, and the general note submission interface.



Is this note helpful? [✓ Yes] Somewhat No

What was helpful about it?

Cites high-quality sources ☑

Easy to understand ☑

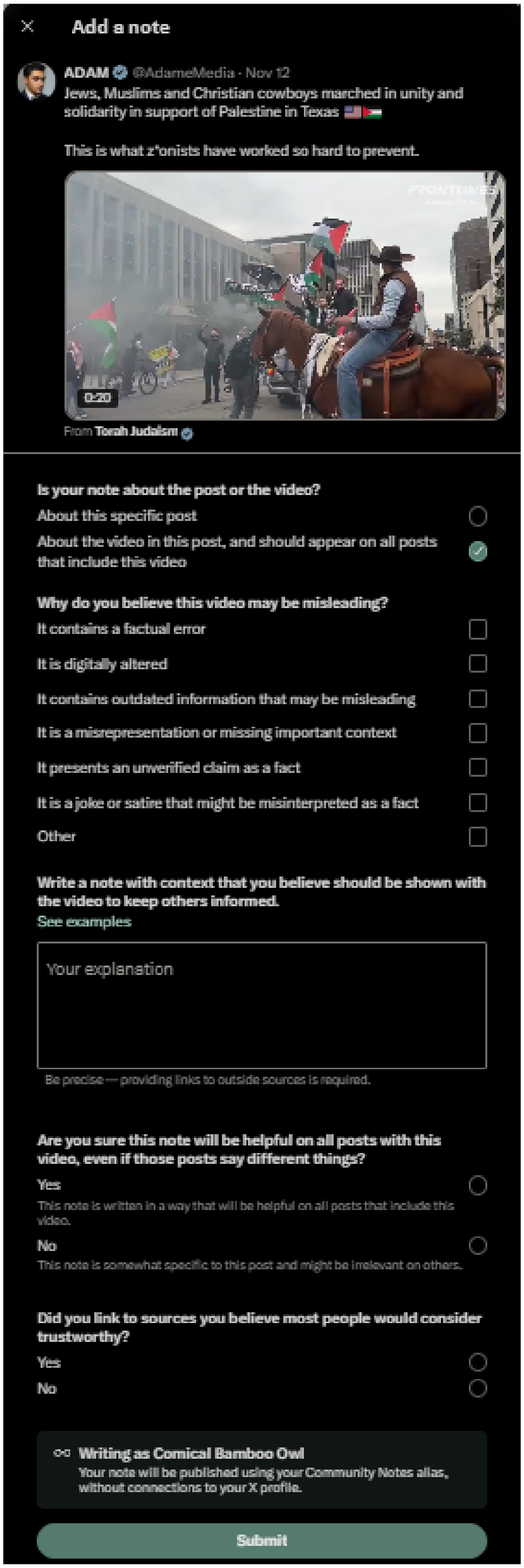
Directly addresses the post’s claim ☑

Provides important context ☑

Neutral or unbiased language ☑

Other □

Figure 1: Contributor interface for rating a note



× **Add a note**

ADAM @AdameMedia · Nov 12
Jews, Muslims and Christian cowboys marched in unity and solidarity in support of Palestine in Texas 🇺🇸🇵🇸

This is what z*onists have worked so hard to prevent.

0:20

From **Torah Judaism**

Is your note about the post or the video?

About this specific post ○

About the video in this post, and should appear on all posts that include this video ✓

Why do you believe this video may be misleading?

It contains a factual error □

It is digitally altered □

It contains outdated information that may be misleading □

It is a misrepresentation or missing important context □

It presents an unverified claim as a fact □

It is a joke or satire that might be misinterpreted as a fact □

Other □

Write a note with context that you believe should be shown with the video to keep others informed.
See examples

Your explanation

Be precise — providing links to outside sources is required.

Are you sure this note will be helpful on all posts with this video, even if those posts say different things?

Yes ○
This note is written in a way that will be helpful on all posts that include this video.

No ○
This note is somewhat specific to this post and might be irrelevant on others.

Did you link to sources you believe most people would consider trustworthy?

Yes ○

No ○

Writing as Comical Bamboo Owl
Your note will be published using your Community Notes alias, without connections to your X profile.

Submit

Figure 2: Contributor interface for creating and submitting a Community Note

Literature Review

The spread of misinformation on social media platforms has significantly impacted public perception during geopolitical conflicts. This literature review examines existing research on misinformation spread during these conflicts and evaluates the effectiveness of social media interventions, with a particular focus on X's (formerly Twitter) Community Notes feature. Research by Vosoughi, Roy, and Aral (2018) found that false information posted to social media platforms spread significantly faster and reached more people than truthful verified information. This contributes to the creation of a dangerous feedback loop during the Russia-Ukraine war and the broader Israel conflict in the middle east, where misinformation can influence global perceptions and influence decision making (Iskoujina, Gnatchenko, & Bernal, 2023).

Misinformation in the Russia-Ukraine Conflict

The Russia-Ukraine conflict has been marked by extensive information warfare, with social media serving as a critical battleground. The study by Iskoujina et al. (2023) highlights the role of social media as an information warfare tool, emphasizing the rapid spread of propaganda and disinformation. Perri et al. (2023) analyzed the spread of propaganda and misinformation on Facebook and Twitter during the earlier stages of the conflict, identifying the significant role of Russian efforts in publishing propaganda on these platforms. The RAND Corporation (2023) noted that Russia's use of social media exploded after the conflict began in 2014, aiming to exploit and influence Western responses to the Ukraine conflict to gain support for their efforts to retake Ukrainian land by annexation.

Misinformation in the Israel-Hamas Conflict

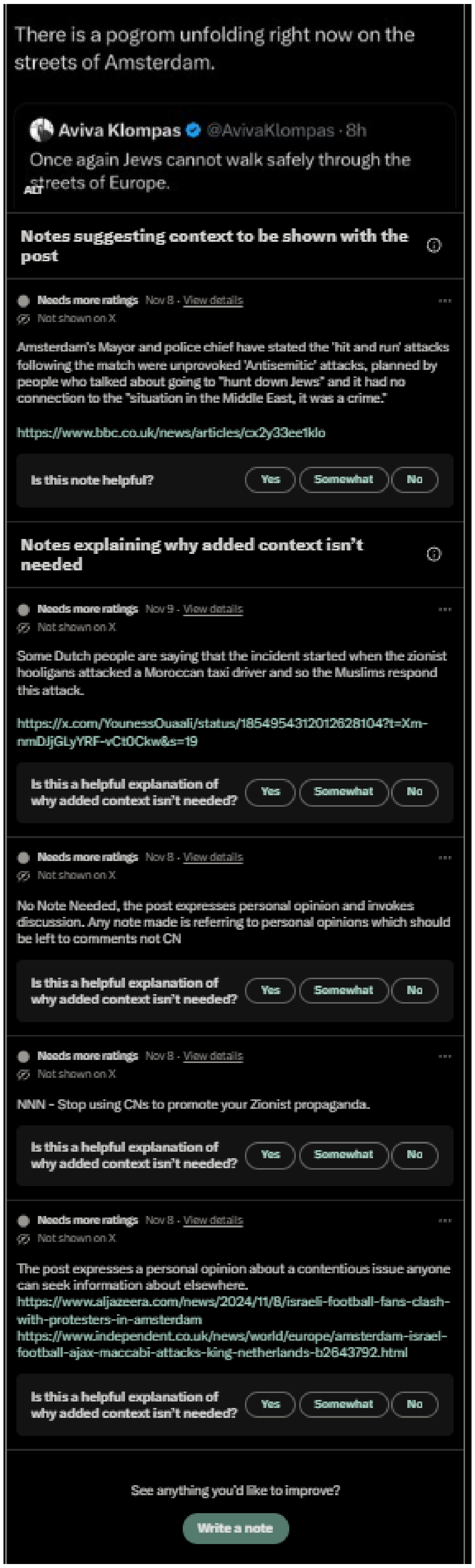
Similarly, the Israel-Hamas conflict has experienced a surge in misinformation after Hamas led rebels attacked Israel on October 7th, 2023. Analysts from the Center for Strategic and International Studies observed that social media platforms have become quite instrumental in spreading both pro-Israel and pro-Palestinian narratives, often confusing the readers by making it incredibly difficult to understand whether they are reading fact-based reporting, or propaganda (Analysts from the Center for Strategic and International Studies [CSIS], 2023). The RAND corporation (2023) emphasized the role of disinformation in amplifying polarization during the conflict, noting that both sides are guilty of using social media to try and garner support and influence public opinion.

Community Notes on X

The Community Notes feature on X utilizes an open-source algorithm to rank notes based on the ratings of contributors. This algorithm helps to identify notes that achieve a consensus across diverse perspectives. While this research project did not look at the algorithm directly, its features and functionality provide important context for understanding how note visibility is managed through this system. The background of this functionality supports the decision made to focus on temporal aspects, including category averages, survival analysis, and sentiment analysis regarding the handling of potential misinformation on X (X, 2024).

Community driven fact-checking systems like X's Community Notes have been created as innovative solutions to combat and slow the spread of misinformation by allowing volunteers who have signed up to participate in writing notes and in rating them with, the goal of helpful notes displaying to the public and adding important context, like links to verified information and reports. Pröllochs (2021) analyzed an earlier iteration of Community notes, previously named Birdwatch, and identified its potential to utilize crowdsourcing to fact-check posts made by X users. However, Allen, Martel, and Rand (2022) discovered that partisanship significantly influences how well this type of system performs since users are often evaluating content through already biased lenses.

Figures 3 and 4 below illustrate examples of conflicting perspectives within the Community Notes system, where contributors submit notes for context and make choices on whether context is needed or not. These examples underscore the significant challenges that Community Notes faces in gaining consensus amongst contributors and the delays that can be caused by these disagreements, a limitation that was also noted in previous research by Chuai, Tian, and Pröllochs (2023).



There is a pogrom unfolding right now on the streets of Amsterdam.

Aviva Klompas @AvivaKlompas · 8h

Once again Jews cannot walk safely through the streets of Europe.

Notes suggesting context to be shown with the post

Needs more ratings Nov 8 · View details
Not shown on X

Amsterdam's Mayor and police chief have stated the 'hit and run' attacks following the match were unprovoked 'Antisemitic' attacks, planned by people who talked about going to "hunt down Jews" and it had no connection to the "situation in the Middle East, it was a crime."

https://www.bbc.co.uk/news/articles/cx2y33ee1klo

Is this note helpful?

Notes explaining why added context isn't needed

Needs more ratings Nov 9 · View details
Not shown on X

Some Dutch people are saying that the incident started when the zionist hooligans attacked a Moroccan taxi driver and so the Muslims respond this attack.

https://x.com/YounessOuaali/status/1854954312012628104?t=Xm-nmDJjGLyYRF-vCt0Ckw&s=19

Is this a helpful explanation of why added context isn't needed?

Needs more ratings Nov 8 · View details
Not shown on X

No Note Needed, the post expresses personal opinion and invokes discussion. Any note made is referring to personal opinions which should be left to comments not CN

Is this a helpful explanation of why added context isn't needed?

Needs more ratings Nov 8 · View details
Not shown on X

NNN - Stop using CNs to promote your Zionist propaganda.

Is this a helpful explanation of why added context isn't needed?

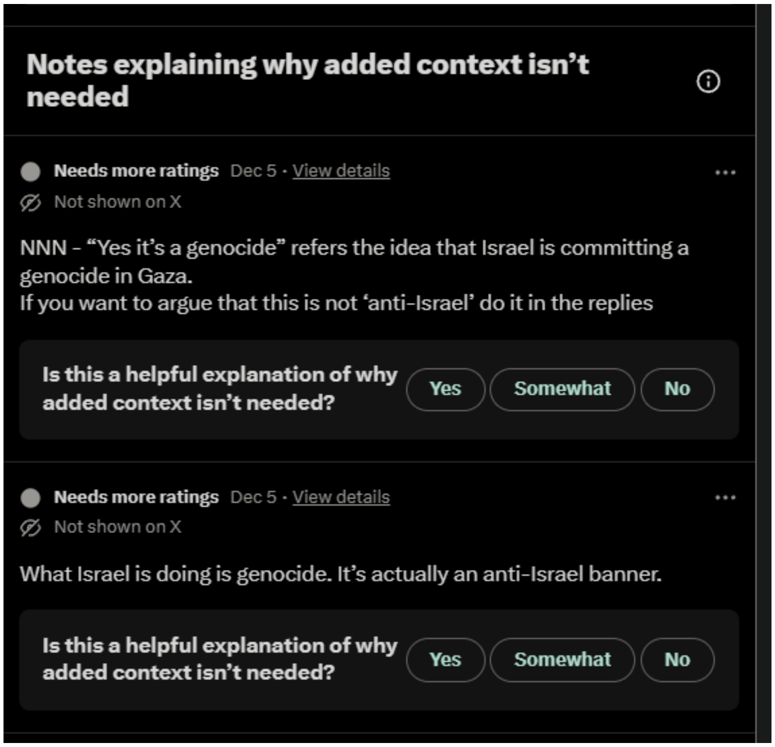
Needs more ratings Nov 8 · View details
Not shown on X

The post expresses a personal opinion about a contentious issue anyone can seek information about elsewhere.
https://www.aljazeera.com/news/2024/11/8/israeli-football-fans-clash-with-protesters-in-amsterdam
https://www.independent.co.uk/news/world/europe/amsterdam-israel-football-ajax-maccabi-attacks-king-netherlands-b2643792.html

Is this a helpful explanation of why added context isn't needed?

See anything you'd like to improve?

Figure 3: A highly contentious representation of a Community Note on X as seen from a contributor's perspective



Notes explaining why added context isn't needed

● Needs more ratings Dec 5 · View details

Not shown on X

NNN - "Yes it's a genocide" refers the idea that Israel is committing a genocide in Gaza.
If you want to argue that this is not 'anti-Israel' do it in the replies

Is this a helpful explanation of why added context isn't needed?

● Needs more ratings Dec 5 · View details

Not shown on X

What Israel is doing is genocide. It's actually an anti-Israel banner.

Is this a helpful explanation of why added context isn't needed?

Figure 4: Contributor disagreements on whether context is necessary.

Despite its promise, Community Notes has faced some significant challenges in their goal to combat misinformation on X. Chuai et al. (2023) also found that the introduction of the Community Notes system did not stop other users from engaging in with misleading tweets and suggested that the large delays they discovered in processing and publishing notes adding context to tweets can reduce its effectiveness. Additionally, the timeliness of correction is critical because people scroll through and digest social media posts incredibly fast. Any delays can contribute to a larger and more viral spread of misinformation (Allen et al., 2022).

Overall, while community Notes represents a noted step toward addressing the challenge of fighting misinformation, its current implementation, even after improvements, reveals limitations and bottlenecks in the system. While many improvements have sped the system up, and attempt to fix some of the identified issues, delays in visibility, and submitted notes remaining stuck needing more ratings for long periods of time remain problem areas. This literature review underscores the need for continued adjustments to the Community Notes algorithm and the overall program.

Methodology

Data Collection and Preprocessing

Data for this study was sourced from X's publicly available datasets, spanning posts and Community Notes from October 6, 2023, to November 6, 2024. The raw datasets can be accessed here: X - Community Notes Data. The dataset was filtered using conflict-specific keywords to focus on content relevant to the Ukraine-Russia war and the Israel-Hamas conflict. This approach of keyword filtering is well-documented in misinformation research for isolating targeted data (Vosoughi, Roy, & Aral, 2018). After filtering, the data was consolidated into a single combined dataset for further analysis. Entries flagged as "No Note Needed" were excluded to ensure the analysis concentrated solely on notes intended for public display. All preprocessing was conducted in Python using libraries such as Pandas and NumPy, which are standard tools for data cleaning and manipulation in academic research (McKinney, 2010).

Sentiment Analysis

Sentiment analysis was performed on the processed Community Notes dataset using the Valance Aware Dictionary and Sentiment Reasoner (VADER) within Python, a rule-based tool that was specifically designed for analyzing sentiment in social media text. VADER is highly effective for short summary and informal text, making this particularly suitable for this study's dataset (Hutto & Gilbert, 2014). Sentiments are categorized on a scale of -1 to +1 with 0 indicating neutral sentiment. Sentiments of the notes were categorized as positive, negative, or neutral and assigned a score to analyze whether emotional tone may contribute to the visibility, or lack of, when looking at notes regarding these two major conflicts.

Data Visualization

Data visualizations were generated to enhance understanding and interpretation of the results and to uncover patterns and trends in the data. Survival curves for notes in the "Needs More Ratings" category was generated by using the Kaplan-Meier analysis method and visualized with Python's Matplotlib library. This method has been extensively utilized in misinformation studies to help explore the temporal dynamics of content visibility (Zhou, Chen, & Zafarani, 2020). The curves were then inverted to enhance user understanding about the probability of a note becoming visible. Tableau was used to create graphics and interactive dashboards that depict key findings including the sentiment analysis scores and note processing efficiency. These data visualization techniques were specifically chosen for the ability to present complex results in an easily understood format that can be used to address a wider audience (Kirk, 2016). Additionally, a word cloud illustrating the frequency of the keywords in the dataset is included in the appendix and on the website, as word cloud illustrations are widely accepted tools for visualizing text data (Heimerl, Lohmann, Lange, & Ertl, 2014).

Results

The analysis of the Community Notes data for both the Israel-Hamas conflict and the Russia-Ukraine conflict revealed some significant disparities in the distribution of current note statuses (Fig 5, 6, 7). These results highlight patterns in how notes are categorized and processed within the Community Notes system.

Figures 5, 6, and 7 illustrate the distribution of Community Notes categorized as "Helpful," "Not Helpful," and "Needs More Ratings" for the Israel-Hamas and Russia-Ukraine conflicts. For the Israel-Hamas conflict, 6.34% of notes were categorized as "Helpful," 3.83% as "Not Helpful," and 89.83% as "Needs More Ratings." In contrast, the Russia-Ukraine conflict saw 10.12% of notes marked as "Helpful," 3.38% as "Not Helpful," and 86.50% as "Needs More Ratings."

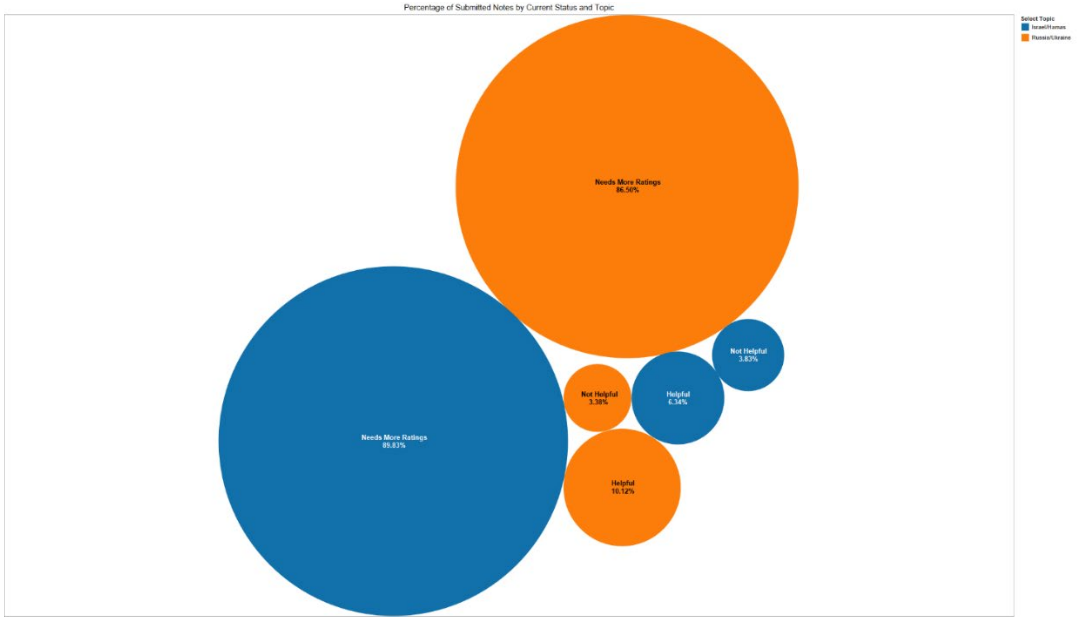


Figure 5: Bubble Chart showing the percentage of notes in each category by Topic.

Figure 6 displays the raw counts of notes submitted for each conflict. For the Israel-Hamas conflict, there were 2,521 notes classified as "Helpful," 1,520 as "Not Helpful," and 35,697 as "Needs More Ratings." In comparison, the Russia-Ukraine conflict had 401 notes classified as "Helpful," 134 as "Not Helpful," and 3,427 as "Needs More Ratings." These raw figures highlight the total volume of notes generated for each conflict across all categories.

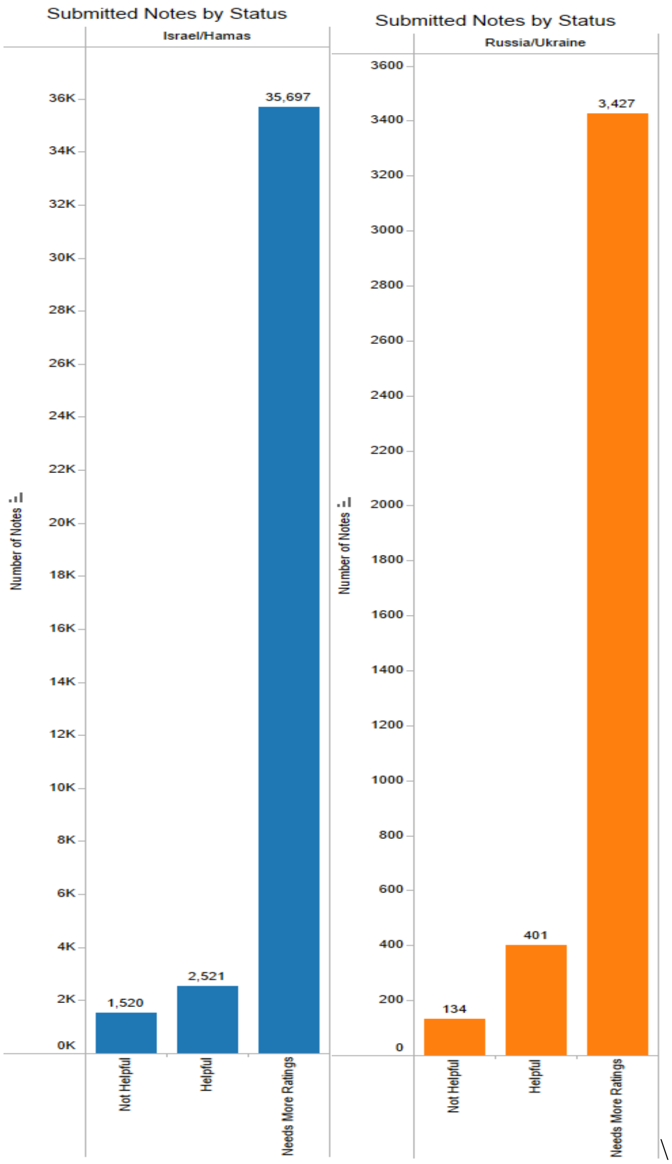


Figure 6: Number of Notes in each category by topic. Notes in the status of "Needs More Ratings" and "Not Helpful" are not displayed to the public.

The median time to visibility across all note categories shows that Israel-Hamas related notes required 195 days to transition, whereas the Russia-Ukraine related notes required 130 days, indicating a 55-day difference between the two conflicts (Fig 7).

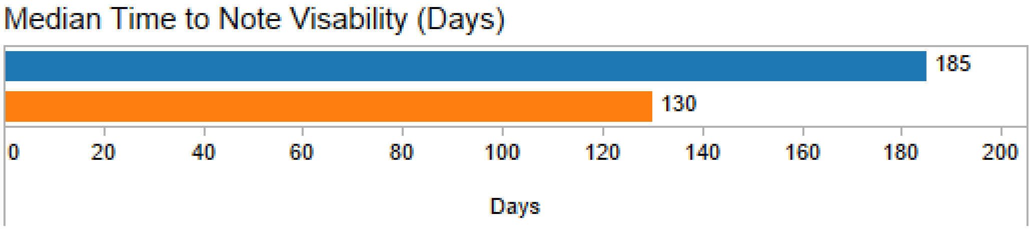


Figure 7: The median time it takes a note to become visible across all note categories

The average time for notes to transition to 'Helpful' and thus visible to the public was 27.36 hours for the Israel-Hamas conflict and 17.28 hours for the Russia-Ukraine conflict. Notably, the average time for notes to transition to 'Not Helpful' showed a significant disparity, with Israel-Hamas notes taking 227.04 hours compared to just 15.36 hours for Russia-Ukraine notes (Fig. 8).

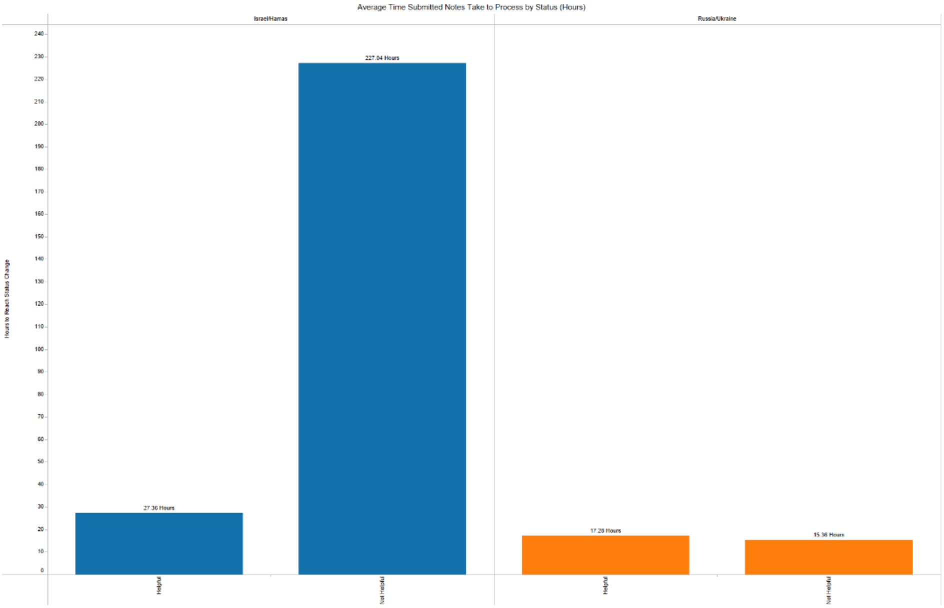
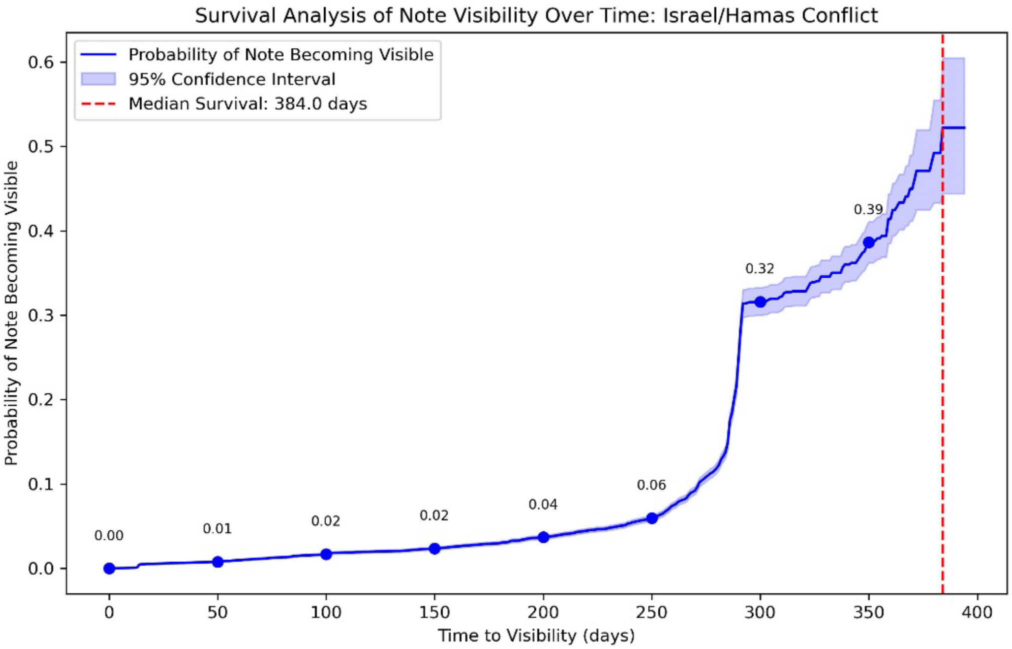


Figure 8: Average time notes take to reach a status other than "Needs More Ratings"

The Kaplan-Meier survival curves display the probability that a note stuck in “Needs More Ratings” will transition to “Helpful” over time and gain visibility to the public. For the Israel-Hamas conflict, the median survival time was 384 days, while the Russia-Ukraine conflict was 292 days. This shows a 92-day difference between the two topics. As an example, at 150 days, the probability of notes transitioning out of “Needs More Ratings” to “Helpful” is just 2% and 7% for the Russia-Ukraine conflict (Fig 9, 10).



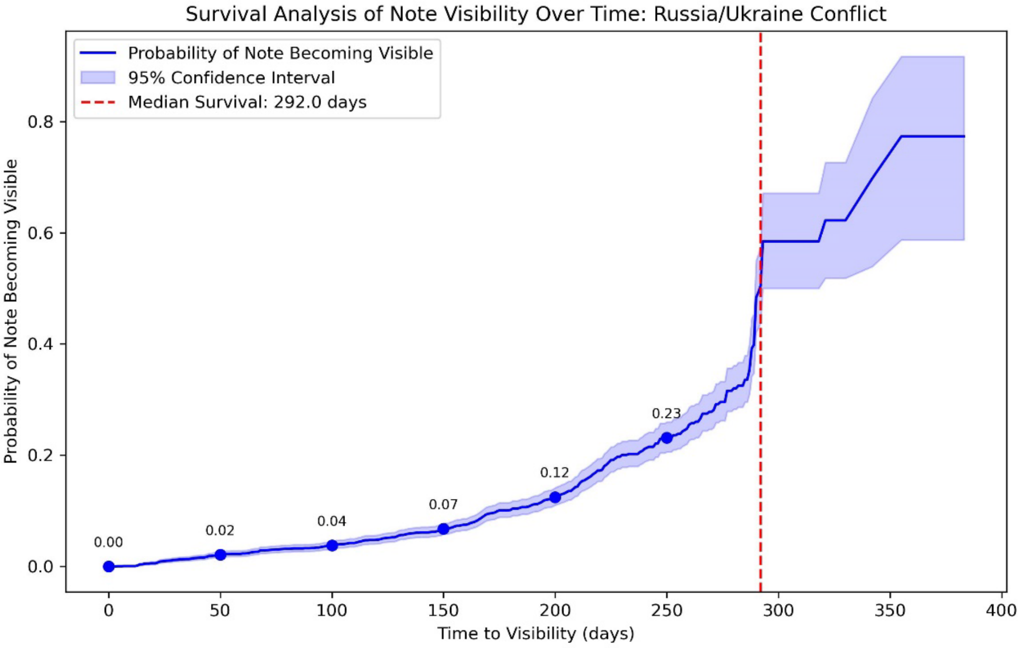


Figure 9 and 10: *Kaplan-Meier survival curve showing the probability of notes transitioning from "Needs More Ratings" to visibility (helpful) over time. The median survival times are 384 days for the Israel-Hamas conflict and 292 days for the Russia-Ukraine conflict respectively, indicating that 50% of notes remain in "Needs More Ratings" beyond this point.*

Figure 11 illustrates the average sentiment scores of Community Notes categorized as “Helpful”, “Needs More Ratings”, and “Not Helpful” for both conflicts. Sentiment scores are measured on a scale of -1 (most negative) to +1 (most positive), with 0 representing neutral sentiment.

For the Israel-Hamas conflict, notes in the “Helpful” category exhibit a sentiment score of -0.185, whereas notes in the “Needs More Ratings” and “Not Helpful” categories show more negative scores of -0.306 and -.326, respectively. Similarly for the Russia-Ukraine conflict, “Helpful” notes have a score of -0.228, while “Needs More Ratings” and “Not Helpful” notes have sentiment scores of -0.226 and -0.269, respectively (Fig 11).

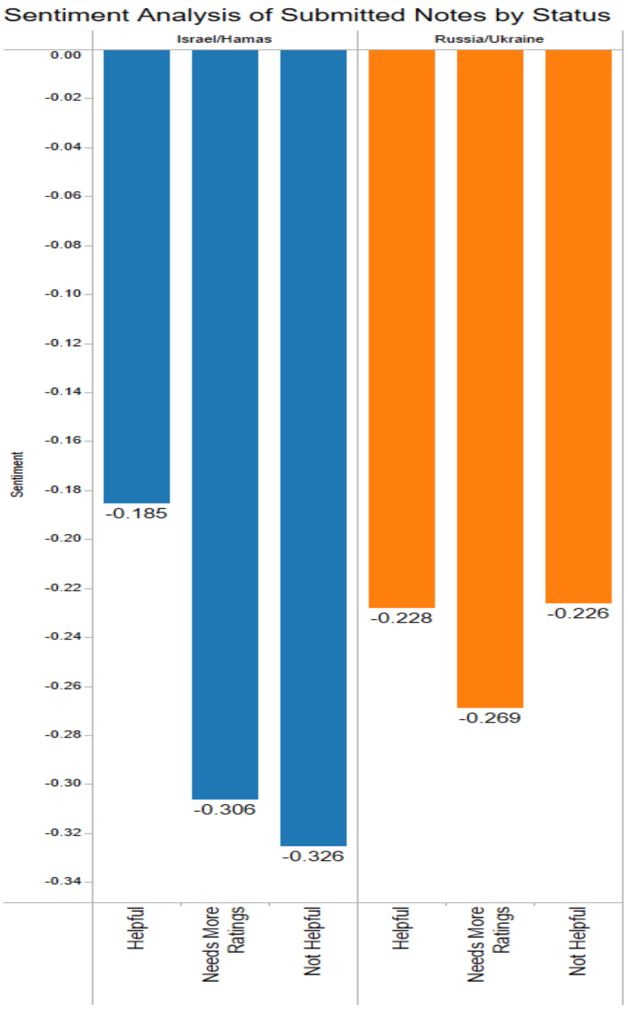


Figure 11: Average sentiment scores of submitted notes by category

Discussion, Future Research, and Limitations (Including SWOT Analysis)

The results of this study reveal significant inefficiencies and disparities in the processing and visibility of Community Notes on X, despite some improvements compared to prior research. While the system seeks to combat misinformation, the data suggests it struggles to meet this goal effectively.

Temporal Patterns and Systemic Bottlenecks

The temporal analysis highlights significant inefficiencies in the Community Notes system, which are obscured by the rapid transition of a small subset of notes to visibility. For example, the median time for visibility across all note statuses (Needs More Ratings, Not Helpful, and Helpful) was 185 days for Israel-Hamas conflict notes and 130 days for Russia-Ukraine conflict notes. Although this marks an improvement over prior studies (Chuai, Tian, and Pröllochs, 2023), which reported delays of 450 to 650 days, these median times still indicate considerable systemic delays in note visibility.

Interestingly, while the average time for notes to become "Helpful" is relatively quick (27.36 hours for Israel-Hamas and 17.28 hours for Russia-Ukraine), this metric is incredibly misleading. It suggests that the system processes notes efficiently, but it overlooks the fact that a significant portion remains stuck in the "Needs More Ratings" category for long periods. Specifically, 89.93% of Israel-Hamas conflict notes and 86.50% of Russia-Ukraine notes in my dataset remained in "Needs More Ratings," unable to reach public visibility. The small number of notes that transition to "Helpful" within a day create the illusion of efficiency and represent only a fraction of the total submissions.

The Kaplan-Meier survival analysis reveals the severity of this bottleneck: notes in the "Needs More Ratings" category have median survival times of 384 days for Israel-Hamas and 292 days for Russia-Ukraine. After 150 days, only 2% of Israel-Hamas notes and 7% of Russia-Ukraine notes transition to "Helpful," highlighting the stark disparity between the small subset of fast-tracked notes and the majority that stagnate for extended periods.

This discrepancy is significant: the apparent efficiency, driven by a few notes becoming "Helpful" quickly, hides the underlying systemic inefficiency affecting the majority. These delays are likely compounded by the algorithm's processing priorities and potential biases toward controversial topics, such as the Israel-Hamas or Russia-Ukraine conflicts Additionally, once notes enter the "Needs More Ratings" category, their chances of being visible drop significantly. This reinforces and provides evidence of this illusion of efficiency, where only a few notes are swiftly processed while most languish in the queue. The lack of visibility for these notes, especially for sensitive topics, prevents them from contributing to public discourse and may inadvertently exacerbate misinformation on the platform.

These findings warrant further investigation into how the algorithm handles these notes. Understanding how rating mechanisms and algorithmic priorities contribute to these delays could help improve the flow of notes, particularly for high-priority or controversial topics.

Sentiment and Visibility

The sentiment analysis adds another layer of complexity to these findings. Notes marked as "Helpful" consistently exhibit less negative sentiment than those categorized as "Needs More Ratings" or "Not Helpful." For example, "Helpful" notes for Israel-Hamas had an average sentiment score of -0.185, compared to -0.306 for "Needs More Ratings" and -0.326 for "Not Helpful." Similarly, "Helpful" notes for Russia-Ukraine scored -0.228, while "Needs More Ratings" and "Not Helpful" notes scored -0.226 and -0.269, respectively. These findings suggest that the system may prioritize less polarizing notes, potentially sidelining those addressing contentious issues. While this might reduce conflict on the platform, it risks perpetuating bias by prioritizing "safer" corrections that lack critical impact. Or it could be that notes with a more neutral tone reach helpful status more often based on the ratings criteria requiring agreement across diverse perspectives. Notes perceived as more negative by contributors, or more controversial, may cause delays in meeting the ratings threshold required by the algorithm.

A Need for Scale and Incentivization

The overwhelming volume of notes in the "Needs More Ratings" category—over 35,000 for Israel-Hamas and more than 3,400 for Russia-Ukraine—highlights the system's lack of scale. Without a larger, more active contributor base, the system cannot process notes efficiently. An incentivization program could address this gap by encouraging broader participation. For instance, contributors could earn "X Credits," redeemable for X merchandise, premium subscriptions, or gifting subscriptions to others. Such incentives could attract a more diverse pool of contributors, strengthening the system's capacity to address misinformation.

Relevance to the United Nations Sustainable Development Goals

This study aligns with United Nations Sustainable Development Goal (SDG) 16, which seeks to promote peace, justice, and strong institutions (United Nations, 2015). SDG 16 emphasizes the importance of providing access to reliable information and combating misinformation to foster informed decision-making and public trust. While Community Notes embodies this objective, the systemic delays and inefficiencies revealed in this analysis underscore the urgent need for X to refine its algorithms and expand its contributor base. Without these improvements, the system risks undermining its potential as a mechanism for enhancing digital accountability and public trust.

SWOT Analysis

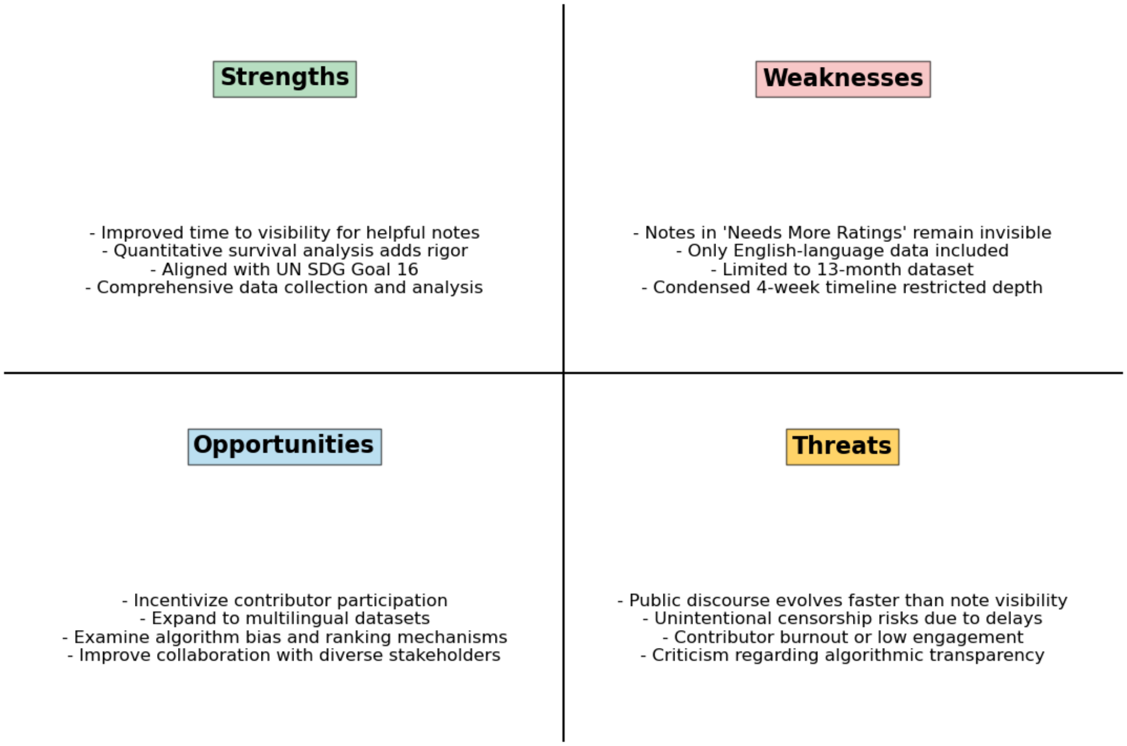
The SWOT analysis provides a stark but quite necessary critique of the Community Notes system and my research project, capturing the potential and the pitfalls (Figure 12). There are commendable strengths: the research does show improved time-to-visibility metrics for helpful notes compared to earlier research, which is a small but noteworthy step forward. Using rigorous survival analysis methods adds credibility to the findings, while aligning the system's intent with the United Nations Sustainable Development Goal (SDG) 16, promoting peace, justice, and reliable institutions, highlights its noble ambitions. Coupled with comprehensive data collection and analysis, these aspects suggest that the system is at least trying to grapple with the wildfire that is social media misinformation.

That said, the weaknesses are glaring. The fact that notes stuck in the "Needs More Ratings" category remain invisible for extended periods borders on negligence. This is not just a bug in the system; it is effectively unintentional censorship, denying critical context when it is needed most. Limiting the analysis to English-language data further narrows the lens, excluding vital non-English contributions to global discourse. And let us not forget the inherent limitations of a 13-month dataset. It provides a snapshot rather than a complete understanding. The four-week project timeline for this analysis only compounded these issues, forcing major compromises on depth and scope.

Opportunities, while promising, come with challenges. Incentivizing contributors through programs like "X Credits," offering perks like free merchandise or premium subscriptions, could broaden participation and inject much-needed energy into the system. Expanding to multilingual datasets and refining algorithms could address glaring inefficiencies, particularly delays and prioritization biases. There’s also room to foster collaboration with diverse stakeholders, not only to enhance algorithmic transparency but also to restore some semblance of trust in the system’s intentions as there has been a lot of people abandoning the X social media platform.

Regarding threats, the elephant in the room is timing: public discourse moves at breakneck speed, and a system that takes months to make notes visible is, frankly, irrelevant in the moment it is most needed. The longer notes languish in obscurity, the greater the risk of perpetuating misinformation through omission. This, coupled with contributor burnout or disengagement, threatens the scalability and sustainability of the program. And the growing scrutiny over algorithmic transparency is not going away. If anything, it is a dark cloud hanging over the system’s credibility, growing larger with every delay and questionable decision.

While Community Notes has potential, it is stuck in a cycle of good intentions undermined by continued poor performance. Without significant reform, better incentivization, more robust data inclusion, and faster turnaround times, it risks becoming just another well-meaning but ineffective tool in the uphill battle against misinformation. Figure 12 visualizes these dynamics, serving as both a summary and a warning.



Strengths

- Improved time to visibility for helpful notes
- Quantitative survival analysis adds rigor
- Aligned with UN SDG Goal 16
- Comprehensive data collection and analysis

Weaknesses

- Notes in 'Needs More Ratings' remain invisible
- Only English-language data included
- Limited to 13-month dataset
- Condensed 4-week timeline restricted depth

Opportunities

- Incentivize contributor participation
- Expand to multilingual datasets
- Examine algorithm bias and ranking mechanisms
- Improve collaboration with diverse stakeholders

Threats

- Public discourse evolves faster than note visibility
- Unintentional censorship risks due to delays
- Contributor burnout or low engagement
- Criticism regarding algorithmic transparency

Figure 12: SWOT Analysis on my Community Notes Research Project

Limitations and Future Research

This study has several limitations also mentioned briefly in the SWOT analysis. By analyzing only English-language notes, it excludes a significant portion of global discourse, potentially skewing the results. Additionally, the 13-month timeframe provides only a snapshot of the system's performance, and the highly condensed four-week project timeline limited the depth of analysis possible. There was a disparity in the size of the data for each conflict. This could be due to the relevance and controversy of the Israel-Hamas conflict during the studied time period compared to the Russia-Ukraine conflict which has been ongoing since 2014. It could also be a limitation of including only English language notes. Future research should expand to include multilingual datasets and examine longer timeframes to capture broader trends. Additionally, exploring the dynamics between algorithmic functionality, sentiment, and contributor dynamics could provide valuable insights into mitigating systemic delays.

Conclusion

Despite incremental improvements, Community Notes remains a deeply flawed system that struggles to keep pace with the rapid spread of misinformation. While the data shows progress in reducing delays compared to earlier studies, these improvements are marginal at best. Without significant reform, particularly in addressing the bottleneck in 'Needs More Ratings,' the Community Notes system risks perpetuating the illusion of efficiency while failing to address the majority of submitted notes. Only by tackling these systemic delays can X achieve its goal of combating misinformation effectively and in real time.

References

Allen, J., Martel, C., & Rand, D. G. (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–24. https://doi.org/10.1145/3491102.3502040

Center for Countering Digital Hate. (2023). *X content moderation failure: How Twitter/X continues to host posts we reported for extreme hate speech*. Center for Countering Digital Hate. Retrieved from https://counterhate.com/wp-content/uploads/2023/09/230907-X-Content-Moderation-Report_final_CCDH.pdf

Chuai, Y., Tian, H., & Pröllochs, N. (2023). *The roll-out of Community Notes did not reduce engagement with misinformation on Twitter.* arXiv. https://arxiv.org/abs/2307.07960

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. *Proceedings of the Annual Conference on Human Factors in Computing Systems*, 1837–1846. https://doi.org/10.1145/2556288.2556961

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Iskoujina, Z., Gnatchenko, Y., & Bernal, P. (2023). *Social media as an information warfare tool in the Russia-Ukraine war.* University of East Anglia. Retrieved from https://www.cmu.edu/ideas-social-cybersecurity/events/ideas2024_paper_6.pdf

Kirk, A. (2016). *Data Visualisation: A Handbook for Data Driven Design*. SAGE Publications.

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56. https://doi.org/10.25080/Majora-92bf1922-00a

Pierri, F., Luceri, L., Jindal, N., & Ferrara, E. (2023). *Propaganda and misinformation on Facebook and Twitter during the Russian invasion of Ukraine. ACM Web Science Conference.* Retrieved from https://dl.acm.org/doi/fullHtml/10.1145/3578503.3583597

Pröllochs, N. (2021). *Community-based fact-checking on Twitter's Birdwatch platform.* arXiv. https://arxiv.org/abs/2104.07175

RAND Corporation. (2023). *Lies, misinformation play key role in Israel-Hamas fight.* Retrieved from https://www.rand.org/pubs/commentary/2023/10/lies-misinformation-play-key-role-in-israel-hamas-fight.html

United Nations. (2015). *Goal 16: Promote peaceful and inclusive societies for sustainable development.* Retrieved December 6, 2024, from https://sdgs.un.org/goals/goal16

Vosoughi, S., Roy, D., & Aral, S. (2018). *The spread of true and false news online. Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

X. (2024). *Ranking notes*. Retrieved December 7, 2024, from https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes

Zhou, X., Chen, L., & Zafarani, R. (2020). *A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys* (CSUR), 53(5), 1–40. https://doi.org/10.1145/3395046

Appendix

Figure A1: Word Cloud for Topic Israel-Hamas conflict. This word cloud highlights keywords associated with the Israel-Hamas conflict in Community Notes.

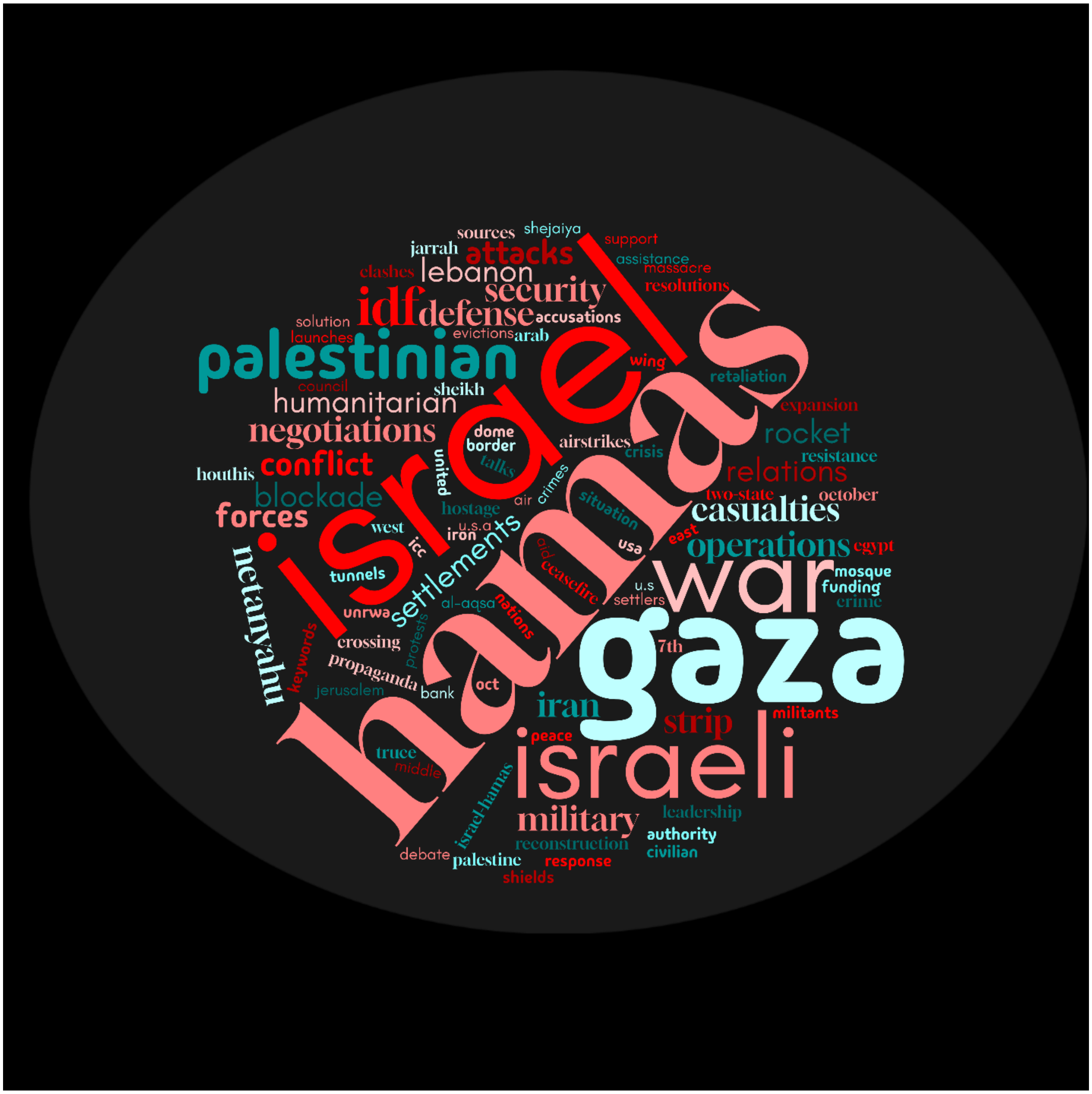


Figure A2: Word Cloud for Russia-Ukraine conflict. This word cloud highlights common keywords associated with the Russia-Ukraine conflict in Community Notes.

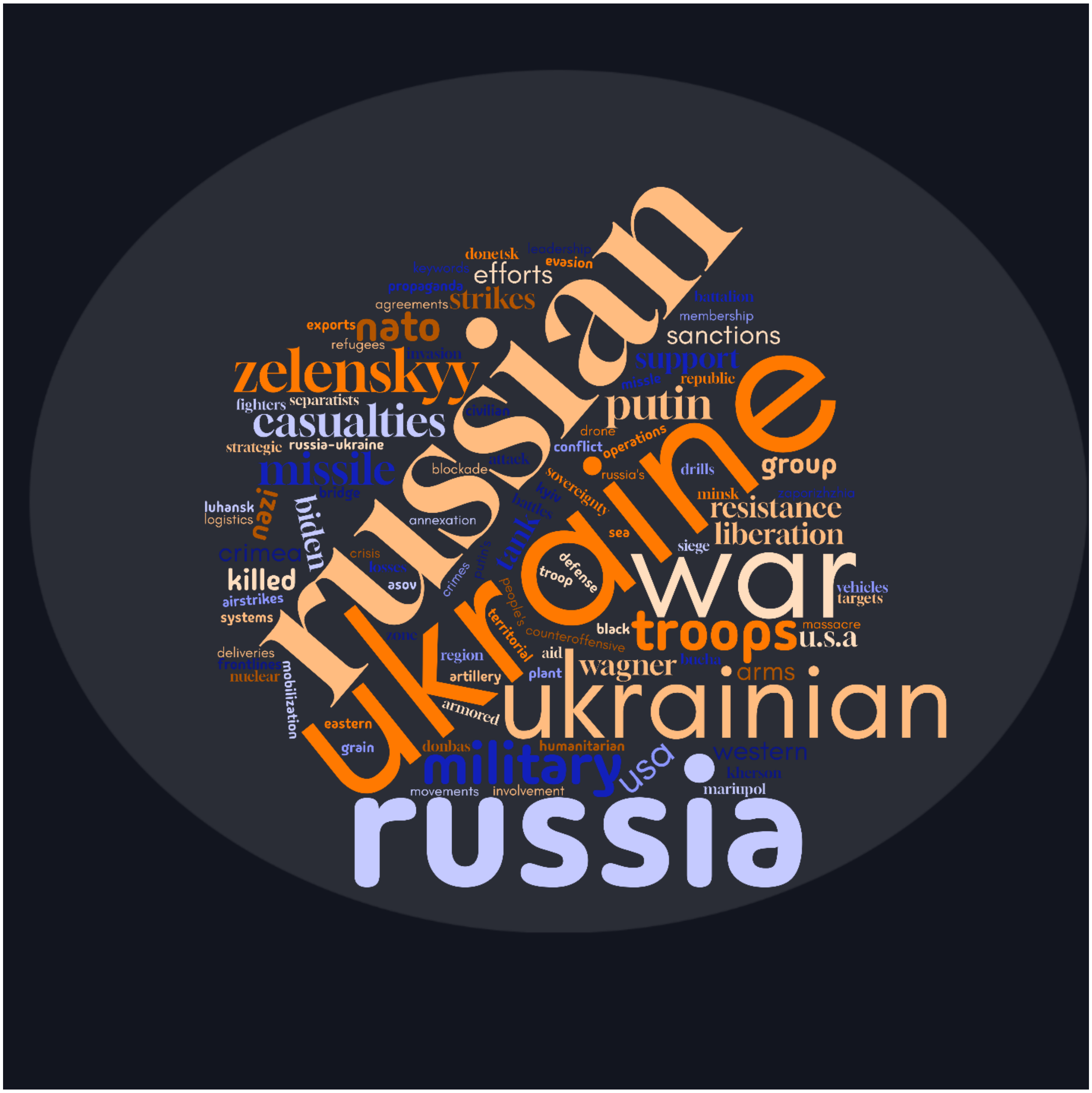
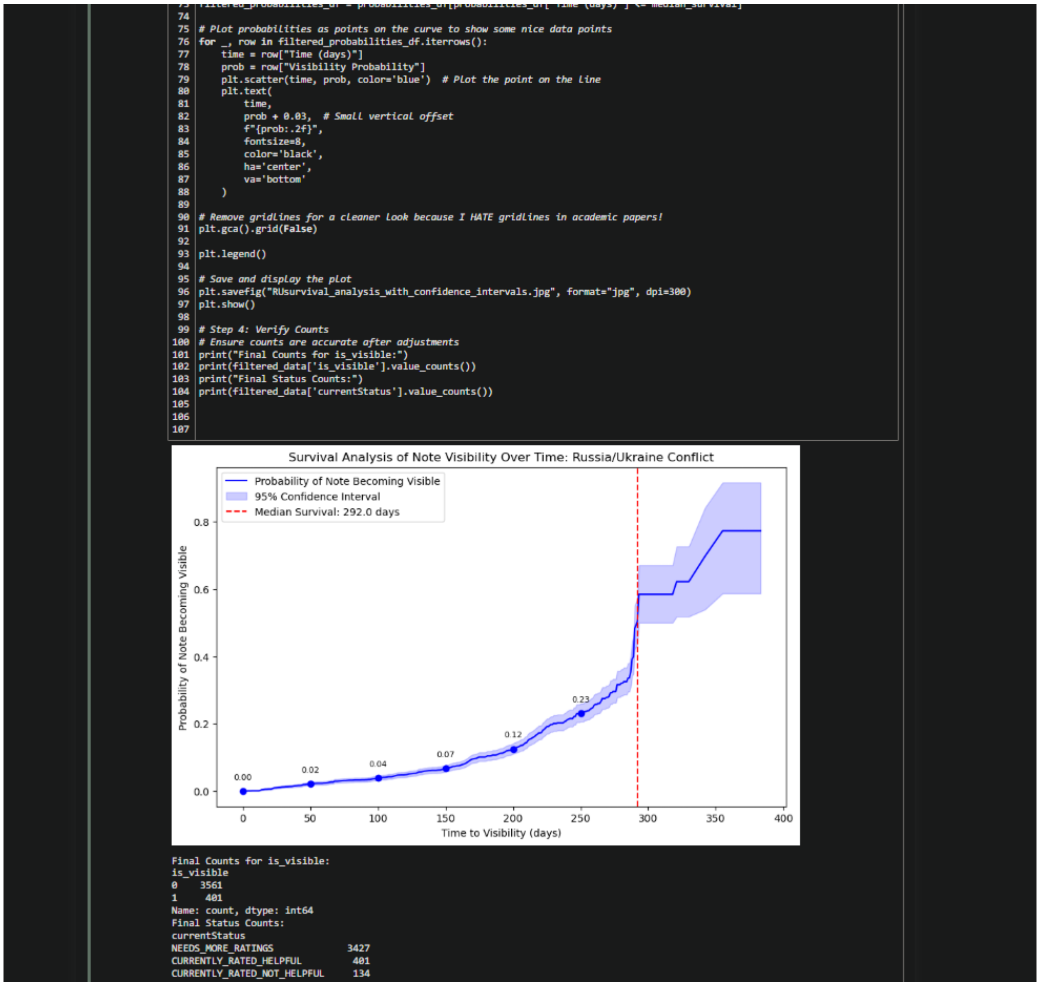


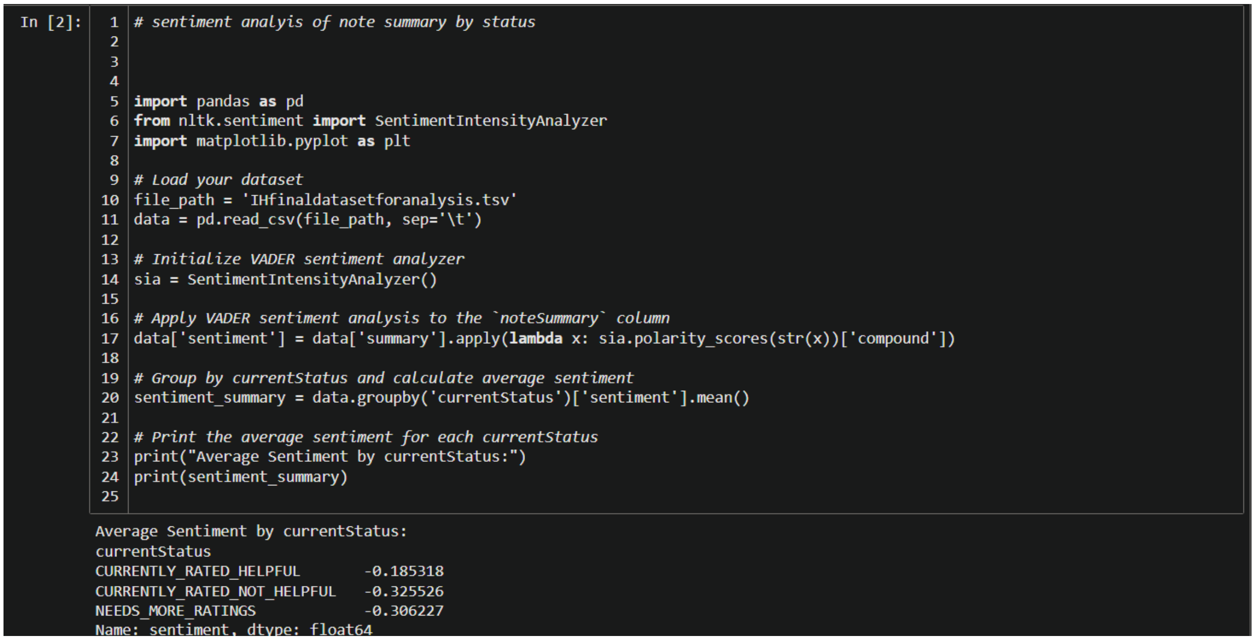
Figure A2, A3: Sample Python coding for survival analysis, sentiment analysis



```
filtered_probabilities_df = probabilities_df[probabilities_df["Time (days)"] <= median_survival]

# Plot probabilities as points on the curve to show some nice data points
for _, row in filtered_probabilities_df.iterrows():
    time = row["Time (days)"]
    prob = row["Visibility Probability"]
    plt.scatter(time, prob, color='blue')  # Plot the point on the line
    plt.text(
        time,
        prob + 0.03,  # Small vertical offset
        f"{prob:.2f}",
        fontsize=8,
        color='black',
        ha='center',
        va='bottom'
    )

# Remove gridlines for a cleaner look because I HATE gridlines in academic papers!
plt.gca().grid(False)

plt.legend()

# Save and display the plot
plt.savefig("RUsurvival_analysis_with_confidence_intervals.jpg", format="jpg", dpi=300)
plt.show()

# Step 4: Verify Counts
# Ensure counts are accurate after adjustments
print("Final Counts for is_visible:")
print(filtered_data['is_visible'].value_counts())
print("Final Status Counts:")
print(filtered_data['currentStatus'].value_counts())
```

Survival Analysis of Note Visibility Over Time: Russia/Ukraine Conflict
Probability of Note Becoming Visible
95% Confidence Interval
Median Survival: 292.0 days
Probability of Note Becoming Visible
0.0
0.2
0.4
0.6
0.8
0.00
0.02
0.04
0.07
0.12
0.23
0
50
100
150
200
250
300
350
400
Time to Visibility (days)

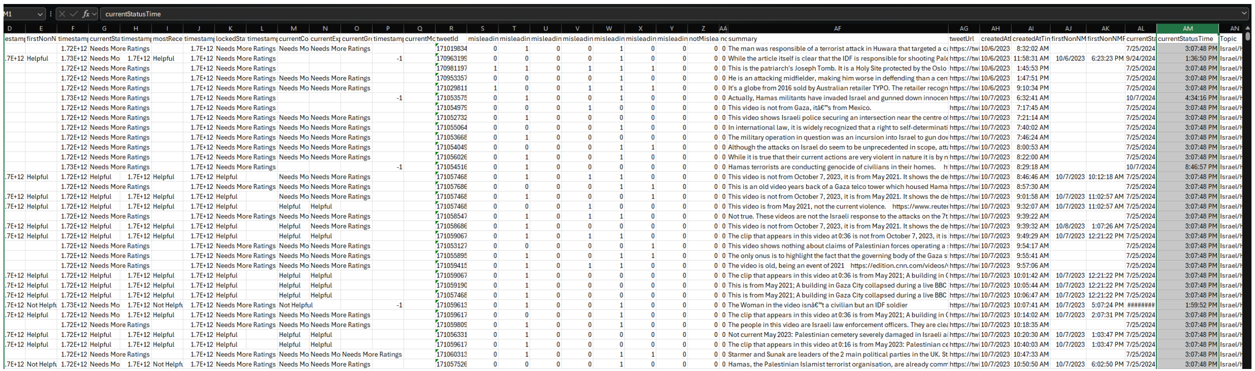
```
Final Counts for is_visible:
is_visible
0    3561
1     401
Name: count, dtype: int64
Final Status Counts:
currentStatus
NEEDS_MORE_RATINGS            3427
CURRENTLY_RATED_HELPFUL        401
CURRENTLY_RATED_NOT_HELPFUL    134
```



```
In [2]:
# sentiment analyis of note summary by status



import pandas as pd
from nltk.sentiment import SentimentIntensityAnalyzer
import matplotlib.pyplot as plt

# Load your dataset
file_path = 'IHfinaldatasetforanalysis.tsv'
data = pd.read_csv(file_path, sep='\t')

# Initialize VADER sentiment analyzer
sia = SentimentIntensityAnalyzer()

# Apply VADER sentiment analysis to the `noteSummary` column
data['sentiment'] = data['summary'].apply(lambda x: sia.polarity_scores(str(x))['compound'])

# Group by currentStatus and calculate average sentiment
sentiment_summary = data.groupby('currentStatus')['sentiment'].mean()

# Print the average sentiment for each currentStatus
print("Average Sentiment by currentStatus:")
print(sentiment_summary)

```

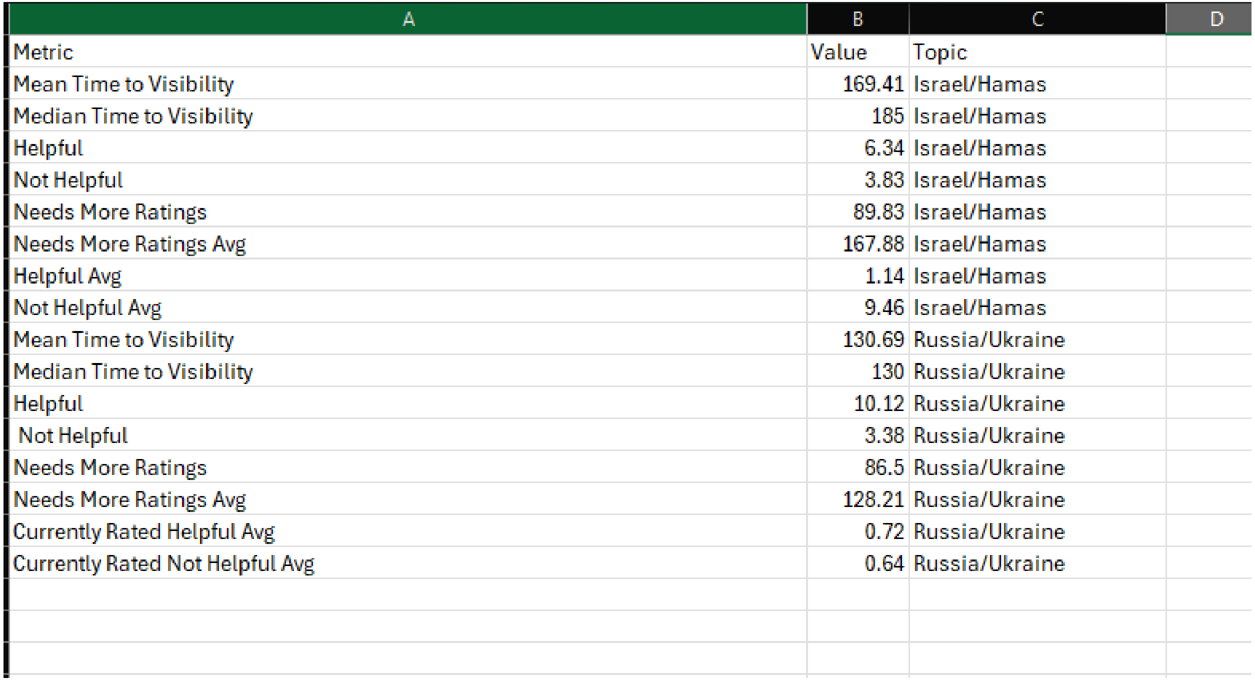
```
Average Sentiment by currentStatus:
currentStatus
CURRENTLY_RATED_HELPFUL       -0.185318
CURRENTLY_RATED_NOT_HELPFUL   -0.325526
NEEDS_MORE_RATINGS            -0.306227
Name: sentiment, dtype: float64
```

Figure A4, A5, A6: Dataset examples as it was prepared for Tableau

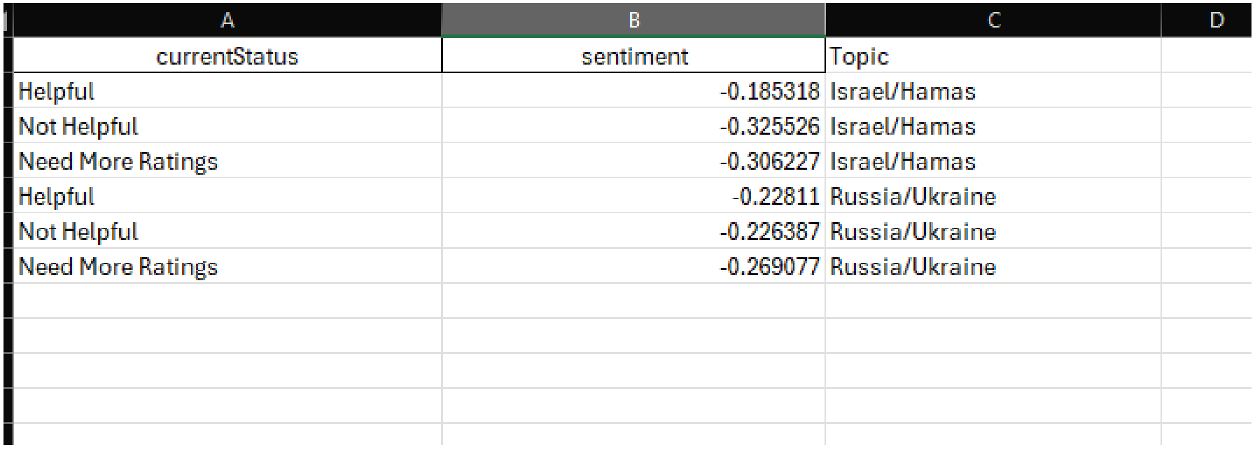


M1 | currentStatusTime

D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AF	AG	AH	AI	AJ	AK	AL	AM	AN
estamp	firstNonN	timestamp	currentSta	timestamp	mostRece	timestamp	lockedSta	timestamp	currentCo	currentExp	currentGr	timestamp	currentMc	tweetId	misleadin	misleadin	misleadin	misleadin	misleadin	misleadin	misleadin	notMislea	nc	summary	tweetUrl	createdAt	createdAtTin	firstNonNM	firstNonNMF	currentSta	currentStatusTime	Topic
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171019834	0	1	0	0	1	0	0	0	0	The man was responsible of a terrorist attack in Huwara that targeted a ca	https://twi	10/6/2023	8:32:02 AM			7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.73E+12	Needs Mo	1.7E+12	Helpful	1.7E+12	Needs More Ratings					-1		170963199	0	0	0	0	1	0	0	0	0	While the article itself is clear that the IDF is responsible for shooting Pale	https://twi	10/6/2023	11:58:31 AM	10/6/2023	6:23:23 PM	9/24/2024	1:36:50 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings							170981197	1	1	0	1	1	1	0	0	0	This is the patriarch's Joseph Tomb. It is a Holy Site protected by the Oslo	https://twi	10/6/2023	1:45:53 PM			7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				170953357	0	0	0	0	1	0	0	0	0	He is an attacking midfielder, making him worse in deffending than a cen	https://twi	10/6/2023	1:47:51 PM			7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171029811	1	0	0	1	1	1	0	0	0	It's a globe from 2016 sold by Australian retailer TYPO. The retailer recogn	https://twi	10/6/2023	9:10:34 PM			7/25/2024	3:07:48 PM	Israel/H
		1.73E+12	Needs More Ratings			1.7E+12	Needs More Ratings					-1		171053575	0	1	0	0	1	0	0	0	0	Actually, Hamas militants have invaded Israel and gunned down innocen	https://twi	10/7/2023	6:32:41 AM			10/7/2024	4:34:16 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings							171054979	0	0	0	1	0	0	0	0	0	This video is not from Gaza, itâ€™s from Mexico.	https://twi	10/7/2023	7:17:45 AM			7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171052732	0	1	0	0	0	0	0	0	0	This video shows Israeli police securing an intersection near the centre ot	https://twi	10/7/2023	7:21:14 AM			7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171055064	0	0	0	0	1	0	0	0	0	In international law, it is widely recognized that a right to self-determinati	https://twi	10/7/2023	7:40:02 AM			7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171053668	0	1	0	0	0	0	0	0	0	The military operation in question was an incursion into Israel to gun dov	https://twi	10/7/2023	7:46:24 AM			7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171054049	0	0	0	0	1	1	0	0	0	Although the attacks on Israel do seem to be unprecedented in scope, atta	https://twi	10/7/2023	8:00:53 AM			7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171056026	0	1	0	0	1	0	0	0	0	While it is true that their current actions are very violent in nature it is by n	https://twi	10/7/2023	8:22:00 AM			7/25/2024	3:07:48 PM	Israel/H
		1.73E+12	Needs More Ratings			1.7E+12	Needs More Ratings					-1		171054516	0	1	0	0	0	0	0	0	0	Hamas terrorists are conducting genocide of civilians in their homes. h	https://twi	10/7/2023	8:29:18 AM			10/7/2024	8:46:57 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Needs Mo	Needs More Ratings				171057468	0	1	0	1	1	0	0	0	0	This video is not from October 7, 2023, it is from May 2021. It shows the de	https://twi	10/7/2023	8:46:46 AM	10/7/2023	10:12:18 AM	7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171057686	0	0	0	0	1	1	0	0	0	This is an old video years back of a Gaza telco tower which housed Hama	https://twi	10/7/2023	8:57:30 AM			7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Needs Mo	Needs More Ratings				171057468	0	1	0	0	0	1	0	0	0	This video is not from October 7, 2023, it is from May 2021. It shows the de	https://twi	10/7/2023	9:01:58 AM	10/7/2023	11:02:57 AM	7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Helpful	Helpful				171057468	0	0	0	1	0	0	0	0	0	This video is from May 2021, not the current violence. https://www.reuter	https://twi	10/7/2023	9:32:07 AM	10/7/2023	11:02:57 AM	7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171058547	0	1	0	1	1	0	0	0	0	Not true. These videos are not the Israeli response to the attacks on the 7t	https://twi	10/7/2023	9:39:22 AM			7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Helpful	Needs More Ratings				171058686	0	1	0	0	0	0	0	0	0	This video is not from October 7, 2023, it is from May 2021. It shows the de	https://twi	10/7/2023	9:39:32 AM	10/8/2023	1:07:26 AM	7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Helpful	Helpful				171059067	0	1	0	1	1	0	0	0	0	The clip that appears in this video at 0:36 is not from October 7, 2023, it is	https://twi	10/7/2023	9:49:29 AM	10/7/2023	12:21:22 PM	7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171053127	0	0	0	0	0	1	0	0	0	This video shows nothing about claims of Palestinian forces operating a s	https://twi	10/7/2023	9:54:17 AM			7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171055895	0	1	0	0	1	1	0	0	0	The only onus is to highlight the fact that the governing body of the Gaza s	https://twi	10/7/2023	9:55:41 AM			7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171059415	0	1	0	1	1	0	0	0	0	The video is old, being an event of 2021 https://edition.cnn.com/videos/	https://twi	10/7/2023	9:57:06 AM			7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Helpful	Helpful				171059067	0	1	0	0	0	0	0	0	0	The clip that appears in this video at 0:36 is from May 2021; A building in G	https://twi	10/7/2023	10:01:42 AM	10/7/2023	12:21:22 PM	7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Helpful	Helpful				171059190	0	1	0	0	0	0	0	0	0	This is from May 2021; A building in Gaza City collapsed during a live BBC	https://twi	10/7/2023	10:05:44 AM	10/7/2023	12:21:22 PM	7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Helpful	Helpful				171057468	0	1	0	0	0	0	0	0	0	This is from May 2021; A building in Gaza City collapsed during a live BBC	https://twi	10/7/2023	10:06:47 AM	10/7/2023	12:21:22 PM	7/25/2024	3:07:48 PM	Israel/H
.7E+12	Not Helpfu	1.73E+12	Needs Mo	1.7E+12	Not Helpfu	1.7E+12	Needs More Ratings		Not Helpful			-1		171059613	0	1	0	0	1	0	0	0	0	The Woman in the video isnâ€™t a civilian but an IDF soldier	https://twi	10/7/2023	10:07:41 AM	10/7/2023	5:07:24 PM	########	1:59:52 PM	Israel/H
.7E+12	Helpful	1.72E+12	Needs Mo	1.7E+12	Helpful	1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171059617	0	0	0	0	1	0	0	0	0	The clip that appears in this video at 0:36 is from May 2021; A building in G	https://twi	10/7/2023	10:14:02 AM	10/7/2023	2:07:31 PM	7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171059809	0	1	0	0	1	0	0	0	0	The people in this video are Israeli law enforcement officers. They are clea	https://twi	10/7/2023	10:18:35 AM			7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Helpful	Helpful				171056331	0	1	0	0	0	0	0	0	0	Not current May 2023: Palestinian cemetery severely damaged in Israeli ai	https://twi	10/7/2023	10:20:30 AM	10/7/2023	1:03:47 PM	7/25/2024	3:07:48 PM	Israel/H
.7E+12	Helpful	1.72E+12	Helpful	1.7E+12	Helpful	1.7E+12	Helpful		Helpful	Helpful				171059617	0	1	0	0	0	0	0	0	0	The clip that appears in this video at 0:16 is from May 2023: Palestinian ce	https://twi	10/7/2023	10:40:03 AM	10/7/2023	1:03:47 PM	7/25/2024	3:07:48 PM	Israel/H
		1.72E+12	Needs More Ratings			1.7E+12	Needs More Ratings		Needs Mo	Needs Mo	Needs More Ratings			171060313	0	1	0	0	1	1	0	0	0	Starmer and Sunak are leaders of the 2 main political parties in the UK. St	https://twi	10/7/2023	10:47:33 AM			7/25/2024	3:07:48 PM	Israel/H
.7E+12	Not Helpfu	1.72E+12	Needs Mo	1.7E+12	Not Helpfu	1.7E+12	Needs More Ratings		Needs Mo	Needs More Ratings				171057526	0	0	0	0	1	0	0	0	0	Hamas, the Palestinian Islamist terrorist organisation, are already comm	https://twi	10/7/2023	10:50:50 AM	10/7/2023	6:02:50 PM	7/25/2024	3:07:48 PM	Israel/H



A	B	C	D
Metric	Value	Topic	
Mean Time to Visibility	169.41	Israel/Hamas	
Median Time to Visibility	185	Israel/Hamas	
Helpful	6.34	Israel/Hamas	
Not Helpful	3.83	Israel/Hamas	
Needs More Ratings	89.83	Israel/Hamas	
Needs More Ratings Avg	167.88	Israel/Hamas	
Helpful Avg	1.14	Israel/Hamas	
Not Helpful Avg	9.46	Israel/Hamas	
Mean Time to Visibility	130.69	Russia/Ukraine	
Median Time to Visibility	130	Russia/Ukraine	
Helpful	10.12	Russia/Ukraine	
Not Helpful	3.38	Russia/Ukraine	
Needs More Ratings	86.5	Russia/Ukraine	
Needs More Ratings Avg	128.21	Russia/Ukraine	
Currently Rated Helpful Avg	0.72	Russia/Ukraine	
Currently Rated Not Helpful Avg	0.64	Russia/Ukraine	



A	B	C	D
currentStatus	sentiment	Topic	
Helpful	-0.185318	Israel/Hamas	
Not Helpful	-0.325526	Israel/Hamas	
Need More Ratings	-0.306227	Israel/Hamas	
Helpful	-0.22811	Russia/Ukraine	
Not Helpful	-0.226387	Russia/Ukraine	
Need More Ratings	-0.269077	Russia/Ukraine	