X Community Notes Dataset - Initial Report

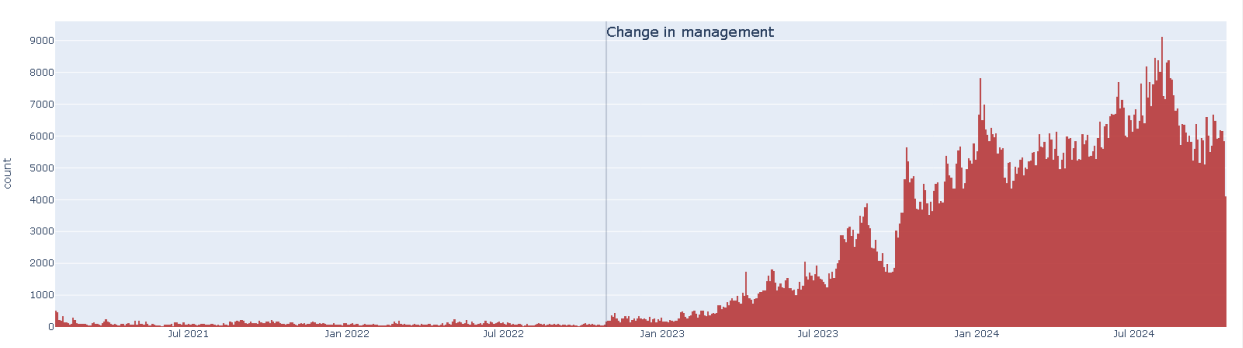
Introduction

This publicly available dataset provides information on the Community Notes (CN) crowd-sourced 'fact-checking' system implemented on X/Twitter since January 2021. Community Notes are responses to tweets contributing additional context, source information, or clarification. Notes themselves are then rated by other contributors to assess the notes' helpfulness. The scope and functionality of this dataset, as well as the statistical models underlying it, have not remained constant since it was first implemented, and we see this reflected in the data. The volume of community notes has increased dramatically since Twitter's buyout in October 2022, and initial exploration suggests that community notes relating to harmful content drive the majority of this increase.

Initial Insights

1. **Increases in Community Notes relating to harmful posts**: The total volume of notes and notes authors/contributors has dramatically increased since early 2023 (Figure 1). A multi-label classifier trained on the data before this change (in which notes were explicitly labelled as relating to potential "considerable harm," "little harm," or N/A, based on the crowd-sourced characterization of the original tweet) predicts that the majority of these notes (72% in 2024[1]) would be labelled as relating to "considerable harm." This is a significant difference vs. the pre-takeover distribution of harm labels[2], in which 63% of notes had been labelled as relating to "considerable harm." Further investigation into the source tweets, available for each note in this dataset, may confirm whether this observation aligns with any increases in harmful content (for example by topic modelling or keyword analysis).

1: Community Notes over time

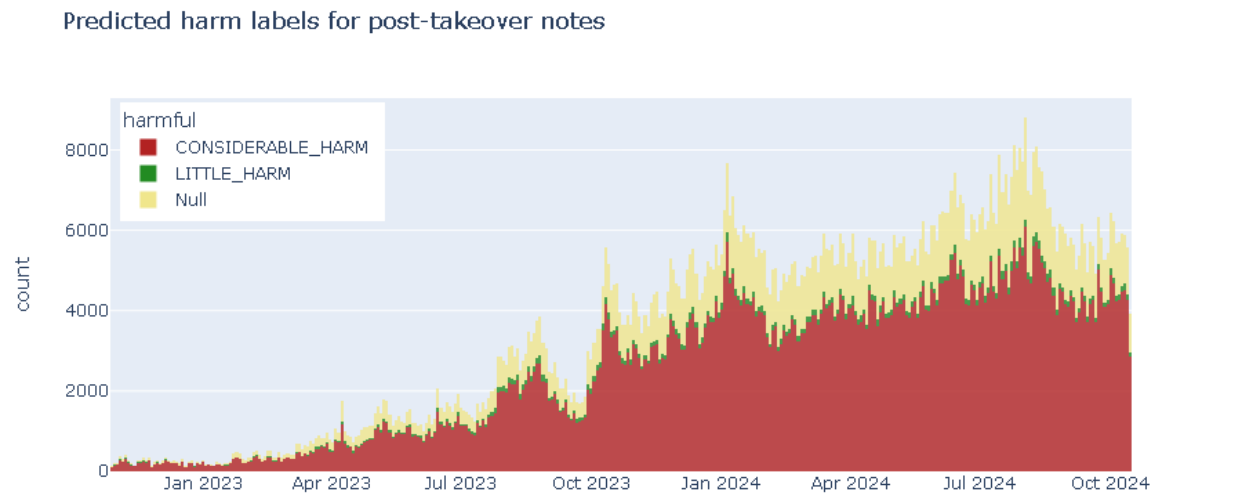


Since early 2023, the number of community notes has increased dramatically, in line with publicised expansions to the system.

2. **Modelling Populations**: Contributors are separated into 3 populations – "core" contributors (intended as a reliable baseline of longer-term users), and "expansion/expansion plus" (supposedly two increasingly wider/less-experienced groups of users, intended to test the ratings of the "core" group against a wider population). However, initial investigation does not reveal significant differences[3] in the earned user privileges (ability to rate/author notes) between these populations, or the rated helpfulness of their contributions. Further investigation should focus on assessing whether these populations are in fact meaningfully different in terms of their CN contributions.
3. **Election Misinformation Networks**: Using the 'summary' data for each note allows for the construction of networks of names associated with a given sub-selection of the dataset, for example, names commonly occurring together in predicted-harmful notes mentioning 'election' or 'voters/voting' (Figure 3). Further investigation will allow better resolution of these networks, extension to include, e.g., organizations/ideological groups, and could support the construction of an ontology/knowledge graph for a given topic, allowing future unsupervised analysis to be done with reference to known entities.

[1] Weighted average F1 score for predictions = 71%
[2] One-way Chi squared test on pre vs post takeover 'harmfulness' label distributions, Test stat = 10.32, p=0.0057
[3] Chi-squared contingency on distribution of user roles in "core" vs "expansion" vs "expansion plus": Test stat = 1.61, p=0.99

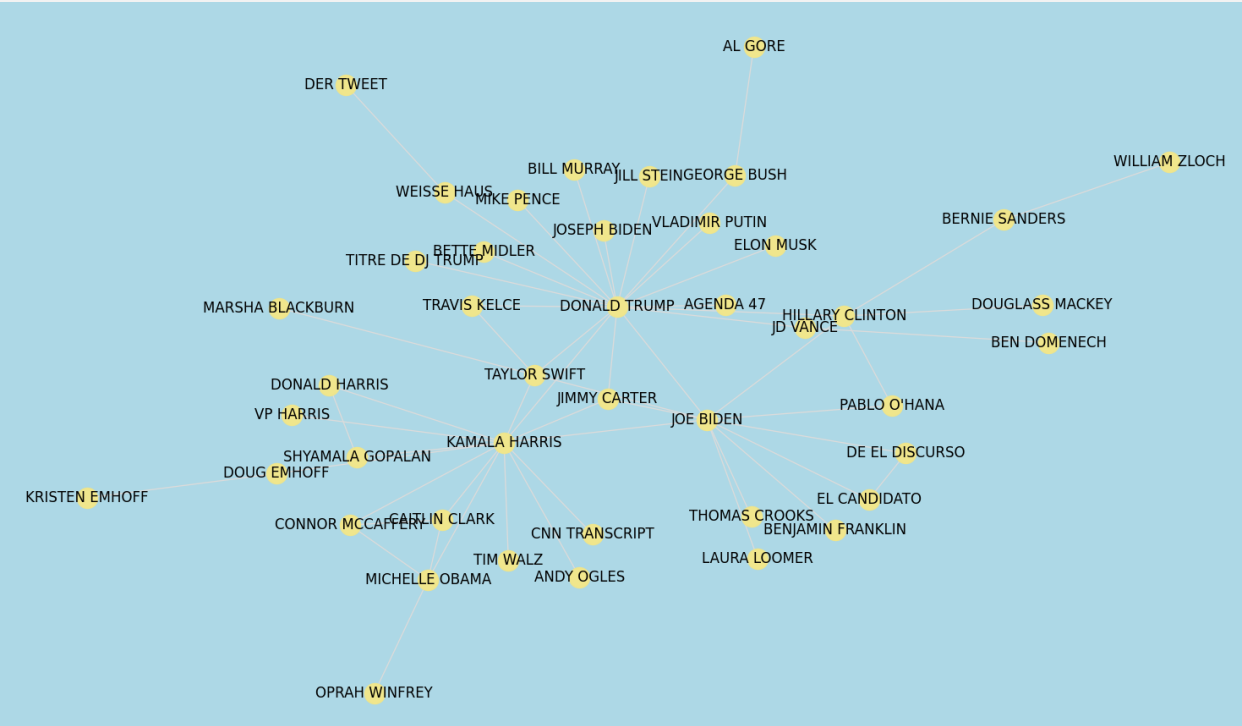
2: Predicting Community Notes relating to harmful content



Pre-takeover data was used to train a multi-label classifier to predict the potential harmfulness of the original tweet a CN was attached to. The model was trained using the crowd-sourced True/False labels attached to the CN (e.g. whether the note was misleading, missing context, based on manipulated media etc.), and a prediction of "considerable harm", "little harm" or "None" generated for each post-takeover CN.

.	Precision	Recall	F1
Considerable Harm	0.74	0.98	0.84
Little Harm	0.60	0.10	0.16
None	1.00	1.00	1.00
Weighted avg	0.75	0.77	0.71

3: Named persons in election related harmful CNs



This network was constructed by performing named entity recognition for persons on Community Notes summaries mentioning "elections", "voters" or "voting" which were predicted by a classification model to relate to considerably harmful content. Names occurring within the same summary were added to the graph with a connecting 'edge'. This approach offers immediate insight into potential clusters, and highlights key individuals prominent in potentially harmful election related discourse on X.