Analyzing Hate Speech Amplification on Fringe Platforms

Anika Ghosh Basu
The Harker School
California
anika.ghosh.basu@gmail.com

Humberto Carlon
A.M.M. Bloomfield High School
California
humbertocarlon05@gmail.com

Junyi Liu
St. Paul's School
Massachusetts
liujunyi2006@gmail.com

Abstract

Fringe platforms like Gab harbor high volumes of hate speech due to minimal moderation and insular communities. Our study examines the factors that determine how hate speech amplifies on these platforms. We prepared a novel dataset of 5K+ threads and 50K+ responses from four fringe platforms (Gab, 4chan, Stormfront, and Vanguard), including both structural features (e.g., timestamps, metadata) and content features (e.g., original text, hate intensity of posts), where hate speech amplification was measured using platform-specific engagement metrics. We trained both Generalized Linear Models and Gradient Boosted Tree models to estimate how several features influence the amplification of hate speech on fringe platforms, and used Shapley value estimates to identify the relative importance of the features. Our analysis shows that research insights from social network analysis (SNA) of mainstream sites like X do not directly generalize to fringe platforms. For instance, our experiments show that using features like thread structure and disagreements in early response windows can give up to 74% lift in Root Mean Squared Error (RMSE) of predicting reply counts for hateful posts on fringe platforms, compared to a baseline model that has features like hate intensity and thread age (which would be considered predictive by regular SNA methods).

1 Introduction

In 2018, Robert Bowers carried out one of the deadliest antisemitic attacks in recent U.S. history, killing 11 people at the Tree of Life synagogue. Moments before the shooting, he posted explicit antisemitic threats on Gab (Byman, 2021; Maarouf et al., 2024). While mainstream platforms such as Twitter/X have strengthened moderation policies, fringe platforms like Gab continue to operate with minimal oversight, offering anonymity and ideological insulation that facilitate extremist discourse. This study examines the factors that drive hate-speech amplification on fringe platforms, which can guide the design of moderation tools.

Prior work shows that online forums can amplify both the intensity and spread of hate speech. The *Online Disinhibition Effect* (Hollenbaugh and Everett, 2013; Watanabe et al., 2018) describes how reduced social accountability and interface-mediated distance enable more extreme and dehumanizing expressions of hate. The absence of face-to-face cues further lowers empathy, while delayed or uncertain consequences encourage riskier behavior.

On fringe platforms, these dynamics are compounded by distinct linguistic and structural characteristics. In the largest analysis of Gab to date, (Tawkat et al., 2023) show that classifiers trained on Twitter/X perform poorly on Gab due to pronounced lexical and structural drift, including the prevalence of coded language, unconventional slur variants, and meme-based hate expressions, as well as longer post lengths and looser syntax. Platform-specific hate vocabularies have also been studied, with (Rieger et al., 2021) reporting that 24% of comments in alt-right communities on 8chan, 4chan, and Reddit contain hateful slurs rarely observed on mainstream platforms.

We study the amplification of hate speech on four fringe platforms: Gab, 4chan, Vanguard, Stormfront. We filter the data to only include posts related to a group of ethnicities that has historically been the target of hate speech, namely Jews, African Americans, Latinos, East Asians, and South Asians. Our contributions in this paper are as follows:

1. ***Novel Hate Speech Amplification Dataset:*** We created an anonymized dataset of 5K+ threads and overall 50K+ responses (with personally identifiable information removed), having both structural features (e.g., timestamps and counts of quotes, links, images in responses) and content features (e.g., text

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 509–514

and hate intensity of the original message and responses), by scraping data from the four fringe platforms we studied. This is the first dataset of its kind that captures thread-level hate speech dynamics across multiple fringe platforms. We plan to augment this dataset with hate speech posts from mainstream platforms (e.g., X/Twitter), and release the data soon as a benchmark for the study of hate speech virality.

2. ***Novel Features Predicting Hate Speech Engagement:*** We trained Generalized Linear Models and Gradient Boosted Trees models to predict post engagement in the hate speech datasets, and analyzed the importance of features using Shapley value estimates. Our analysis identified that apart from hate intensity and thread age, which have been identified as predictors for hate speech virality in earlier work (Mathew et al., 2019), aspects of early responses in the post like targeted replies, disagreement, and challenges play a strong role in driving hate speech amplification (e.g., driving a lift of up to 74% in predicting hate speech engagement).

2 Methodology

We study hate speech dynamics across four major fringe platforms. Gab is a far-right social media platform structurally similar to X; 4chan is a thread-based pseudonymous forum board with extreme unmoderated content; Stormfront and Vanguard are fringe forums promoting white supremacy.

2.1 Data

We first outline the methodology we used to collect the data and relevant features from each platform.

Data scraping: Gab dynamically resurfaces older posts, so we scraped the platform homepages over several days and merged the resulting data into a unified dataset after filtering duplicate posts. For 4chan, we collected our data from the /pol/ board in the 4plebs archive (`https://archive.4plebs.org/`) that has no explicit license – it has white supremacist, xenophobic, antisemitic hate speech. For Stormfront and Vanguard, we scraped posts from the "Newslinks & Articles" and "This Just In" forums. The data was downloaded using the Selenium (SeleniumHQ, 2026) scraping API.

Feature extraction: We extracted three categories of data for each post in a platform: metadata (the post ID, URL, author username, and timestamp), content (the full text of the thread and any associated image URLs, including both the original post and all replies), and engagement metrics (the number of reposts, quotes, views, and reactions). We further used the Tesseract OCR model (Smith, 2007) to extract the text present in images from the image URLs. Table 1 gives an overview of the engagement metrics we extracted from each platform.

Stability of engagement metrics: While collecting engagement metrics, we controlled for temporal variability in forums like Gab that resurface content by using a 2-phase approach. We stored the postIDs upon initial scraping, and then a week later we re-scraped the same posts to capture stabilized engagement metrics. A study of Gab posts suggests that engagement levels plateau around the one-week mark, making this approach more reliable for longitudinal and comparative analysis.

Platform	# Posts	Engagement metrics
Gab	1000	Reposts, quotes, likes views, reactions
4chan	1000	Reply count
Stormfront	1825	View count, reply count
Vanguard	1825	View count, reply count

Table 1: Overview of engagement data collected.

2.2 Models

Our main goal is to study which features in our fringe datasets predict the community engagement on hate speech posts. To accomplish this, we first trained a model to identify the hate speech intensity of the content of the original post and subsequent responses. We subsequently used the hate speech intensity estimates and other extracted features to predict hate speech amplification, as measured by the thread engagement metrics.

2.2.1 Hate Speech Intensity Modeling:

We used Microsoft's DeBERTa-v3 (He et al., 2021) model to predict hate speech intensity of posts, since DeBERTa shows strong performance on language understanding benchmarks, particularly in capturing nuanced, context-dependent semantic differences that are crucial for detecting subtle hate.

The Measuring Hate Speech dataset (Kennedy et al., 2020; Sachdeva et al., 2022) has *hate_speech_score*, a continuous score that takes

human annotation data for hate speech and adjusts for annotator bias across 135K+ comments, made by 8K individuals, using Item Response Theory (De Ayala, 2022). DeBERTa was fine-tuned on this dataset to predict *hate_speech_score*. The fine-tuned DeBERTa prediction had a 0.97 Pearson's correlation coefficient compared to the true scores on the holdout test set of 1K points, showing the effectiveness of the fine-tuned DeBERTa model – we used this model to label the hate speech intensity of all posts in the fringe dataset.

2.2.2 Model Features:

Features were grouped conceptually and evaluated both jointly and via ablation. Only features that exist for a given platform were included in modeling:

1) **Baseline.** Controls capturing thread-level priors, consisting of the hate intensity score of the original post and thread age in days.

2) **Structural.** Platform-specific structural image features, e.g., number of images in the full thread, the original post, and the first K replies.

3) **Early-Interactions.** Features extracted from first K replies, e.g., number of quoted replies, nesting of quoted content, replies with external links.

4) **Early-HateVariance.** Measures of heterogeneity in early hateful responses, e.g., hate intensity max value and variance in early replies.

2.2.3 Hate Speech Amplification Modeling:

We trained two hate speech amplification prediction models on each fringe dataset:

(1) **Generalized Linear Models (GLM):** We modeled thread-level engagement using a GLM (Nelder and Wedderburn, 1972) with a Negative Binomial response distribution, appropriate for overdispersed count data such as reply counts. Let y_i denote the total number of replies for thread i, and $\mathbf{x}_i$ be the feature vector. Our model was:

$$y_i \sim \text{NegBin}(\mu_i, \theta), \quad \log(\mu_i) = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta},$$

where μ_i is the expected reply count, θ is the dispersion parameter estimated from data, and $\boldsymbol{\beta}$ is the vector of regression coefficients. A log link function was used to ensure non-negative predictions.

Training and evaluation. For each platform, data was split into a single fold of training set (80%) and test (20%) set using a fixed random seed, and the same split was reused for all model variants to ensure comparability. We estimated three GLM variants: (i) a Baseline model, (ii) a Full model that included all feature groups, and (iii) Ablation models where each feature group was removed in turn to estimate their predictive efficacy. Performance was evaluated by Root Mean Square Error (RMSE) on held-out data, and computed on $\log(1 + y)$ to stabilize variance and reduce sensitivity to extreme counts.

Design choices. View counts were excluded as features from the models since replies increase views via visibility, violating causal ordering assumptions. Threads with massive counts of replies, images, or external links (top 1%, capped) were filtered, to filter out spam threads and mega threads. We define early interactions as those occurring within the first $K = 10$ replies of a thread.

(2) **Gradient Boosted Decision Trees (Tree Boost):** As a nonlinear complement to GLMs, we fit TreeBoost regressors using XGBoost (Chen and Guestrin, 2016). Instead of modeling raw reply counts y, we use $\log(1 + y)$ with a squared error objective, which gives stability for heavy-tailed count data while retaining nonlinearities.

Model specification. The XGBoost regressor is configured with the following hyperparameters: *num_estimators* = 800, *max_depth* = 5, *learning_rate* = 0.03. The platform-specific feature sets and train-test split used for GLM experiments were also used here, to ensure direct comparability.

Interpretability. To interpret learned nonlinear effects, we compute feature importance using SHAP (SHapley Additive exPlanations) values via a TreeExplainer on a random subsample of up to 2,000 training instances. Features are ranked by mean absolute SHAP value.

Table 2 shows that using structure and early response features can give up to 74% RMSE lift in predicting reply counts for hateful posts in these fringe platforms, on top of baseline features like hate intensity and thread age.

3 Results and Discussion

We analyzed the GLM and TreeBoost results for all four datasets in detail. However, we only have space to report detailed plots for Gab. Figures 1 shows how the full GLM model significantly outperforms the baseline model on the Gab dataset. The feature importance analysis in the figure also shows that feature groups like early hate variance and structure can play a more dominant role compared to the hate intensity of the post in the baseline,

Table 2: RMSE lifts across four platforms for TreeBoost, comparing results using all features to baseline model.

Platform	Baseline	Overall	Lift (%)	Top Features (SHAP)
4chan	0.9551	0.2520	73.62%	*early_quote_count* (0.568), *num_images_op* (0.186), *early_reply_images* (0.093)
Gab	0.5987	0.4272	28.64%	*max_reply_hi_early* (0.216), *early_challenge_count* (0.165), *num_images_total* (0.162)
Stormfront	0.8481	0.6779	20.07%	*early_quote_count* (0.296), *thread_age_days* (0.220), *early_link_count* (0.095)
Vanguard	0.8919	0.5707	36.01%	*early_link_count* (0.240), *early_quote_count* (0.216), *thread_age_days* (0.203)

demonstrating that debate or challenges early on in the replies to a hate post can catalyze more replies and drive engagement. Figure 2 shows the result of Shapley value estimation run on TreeBoost, confirming that features like maximum (or variance) of reply hate intensity or number of challenges in early interactions contribute to hate speech amplification, along with hate intensity of the post and thread age (features in the base model).

The Shapley value estimates from the TreeBoost model in Table 2 also confirm this result on the other three platforms, showing that features like number of early challenges, early quotes (indicating targeted responses), max or variance of hate intensity in early replies, etc. are important in predicting community engagement with hate speech posts. The RMSE lift on the other three platforms is also significant.

Platform-level differences reflect forum affordances. On 4chan, image-related features (e.g., number of images in the OP and early replies) are highly salient, reflecting the platform's imageboard model where memes and reaction images drive discussions. On Gab, early hate variance and challenge counts carry more weight, indicating that heterogeneity in early responses better predicts engagement in a platform that has a more ideologically mixed audience. On Stormfront and Vanguard, the prominence of early quoting, linking, and temporal features aligns with slower, discussion-oriented structures where sustained textual exchange, rather than multimedia amplification, governs thread growth.

4 Related Work

Models for virality generally identify structural features (e.g., network density), temporal features (e.g., the speed of the first ten reshares), and content-based features (e.g., sentiment) as most

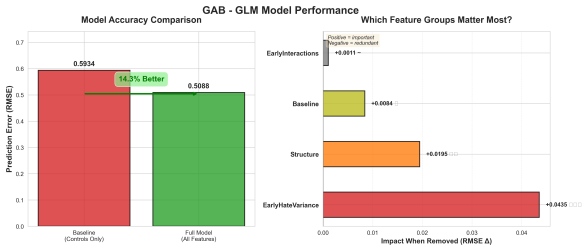

Figure 1: GLM Summmary on Gab Platform.

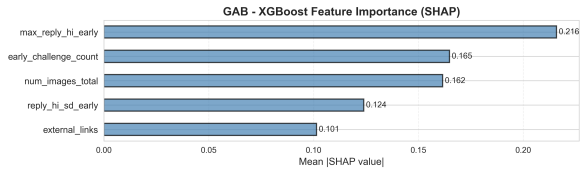

Figure 2: TreeBoost Shapley Values on Gab Platform.

predictive in online platforms. Research also shows that user-level influence and community structure are strong predictors of long-term reach (Cheng et al., 2014). Effective fringe platform models often incorporate "community embeddings" to capture the shifting semantics of hate in places like 4chan or Gab, since "in-group" signaling can be as important as the hate itself (Papasavva et al., 2020). Domain-adaptive Transformers (e.g., HateBERT) have been used capture the specific linguistic toxicity found in fringe platforms (Caselli et al., 2021). The few models specifically targeting the virality of hate speech on fringe platforms combine content toxicity with "user reputation" metrics. (Mathew et al., 2019) demonstrates that "hateful" users on Gab have more followers and their content spreads more widely than non-hateful counterparts Models like "Hate-Prophet" use dynamical "self-exciting" processes to show how one toxic reply can trigger a burst of activity (Gao et al., 2017).

Our work builds on top of prior research, but shows a novel interesting result. We show how

contrary to previous work on mainstream platforms, a higher hate intensity in the main post alone is not a strong predictor of amplification of the post on fringe platforms; instead, features related to thread structure and targeted replies, disagreement, or challenges in the early responses can often play a significant role in driving community engagement with the hate speech post.

5 Limitations

This study has some limitations. First, our dataset spans four fringe platforms, so the findings may not generalize to all extremist or lightly moderated online communities. Second, platform differences may make direct comparison difficult: Gab includes reposts, quotes, likes, views, and reactions, while 4chan, Stormfront, and Vanguard rely mainly on reply or view counts as engagement measures. Third, our hate intensity labels depend on a DeBERTa model fine-tuned on the Measuring Hate Speech dataset, which may not fully capture coded language, platform-specific slang, memes, or image-based hate common on fringe sites. Fourth, while early reply features are predictive of amplification, our models show association rather than causation; early challenges or quotes may reflect an already-engaging thread rather than directly causing later engagement. Finally, our analysis could be deepened with comparison to mainstream platforms, which is an active area of our future work.

6 Future Work

In future work, we want to run a more large-scale analysis on other fringe platforms, collect more ground truth data, and use LLMs to explain why a post or reply is considered hateful.

We also noticed in our study that hate speech intensity in fringe platforms can vary across race and ethnicities — we would like to further study this aspect of hate speech on fringe platforms. We are also collecting data from mainstream platforms (e.g., X/Twitter) — in the future, we would like to compare and contrast hate speech propagation trajectories across fringe and mainstream platforms.

References

Daniel Byman. 2021. How hateful rhetoric connects to real-world violence. *Brookings*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (ALW)*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Justin Cheng, Lada Adamic, P Alex Dow, Jon Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted? In *Proceedings of the WWW Conference*, pages 925–936.

Rafael J. De Ayala. 2022. *The Theory and Practice of Item Response Theory*, 2nd edition. Guilford Press.

Rui Gao and 1 others. 2017. Detecting emotional contagion in social media. In *Proceedings of the IJCAI*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

Elizabeth E. Hollenbaugh and Marc K. Everett. 2013. The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication*, 18(3):283–302.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Ahmed Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel. 2024. The virality of hate speech on social media. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–22.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the ACM Web Science Conference*, pages 173–182.

John Ashworth Nelder and Robert WM Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Jeremy Blackburn, and Gianluca Stringhini. 2020. Raiders of the lost kek: 3.5B comments are analyzed from 4chan's politically incorrect board (/pol/). In *Proceedings of the WWW Conference*.

Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Tim Kiening, and Georg Groh. 2021. Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and reddit. *Social Media + Society*, 7(4).

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The Measuring Hate Speech Corpus: Leveraging Rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

SeleniumHQ. 2026. Selenium browser automation.

Ray Smith. 2007. An overview of the tesseract ocr engine. *International Journal on Document Analysis and Recognition (IJDAR)*, 10(1):1–8.

M. Tawkat, I. Khondaker, Laks Muhammad, and V. Lakshmanan. 2023. Cross-platform and cross-domain abusive language detection with supervised contrastive learning. In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835.