Hate Speech detection using graph mining

Adarsh Vermali
avermali@ucsd.edu

Anuj Goenka
angoenka@ucsd.edu

Tushar Mohan
tmohan@ucsd.edu

Abstract

Effective moderation of hate speech on social media is critical to maintaining healthy online communities. This study introduces an engagement-aware scoring system to enhance hate speech detection and prioritization. Using a dataset of 291 posts and their associated comments mined from the Reddit community "r/PoliticalDiscussion," we constructed a directed user interaction graph and scored posts using the Bertweet model. To address ambiguities in hate speech detection, we developed a composite scoring approach that integrates content-level hate scores, user-level aggregated scores, and graph-based neighborhood influences. Evaluation against a baseline Bertweet model, using Spearman correlation and human-labeled metrics, demonstrated that our system outperforms the baseline in ranking hateful content. These findings highlight the potential of leveraging user interactions and graph-based modeling to improve the prioritization of harmful content for moderation.

1 Introduction

The proliferation of user-generated content on social media platforms has transformed how individuals communicate, share opinions, and consume information. However, this transformation has also been accompanied by a rise in harmful behaviors, including the dissemination of hate speech. Hate speech, broadly defined as abusive or threatening communication that expresses prejudice against individuals or groups based on attributes such as race, religion, or gender, poses significant challenges to content moderation efforts (Castellanos et al., 2023) (Amara, 2024). Effective moderation is essential for fostering inclusive online communities and mitigating the real-world harm that can arise from online abuse (Albrecht et al., 2024) (Laub, 2019).

Despite advances in natural language processing (NLP) and machine learning, hate speech detection remains a complex task. One major challenge lies in identifying ambiguous cases where the hateful nature of the content is context-dependent or subtle. Moreover, hate speech is often intertwined with user interactions and group dynamics, which are difficult to capture using content-level analysis alone (Kovács et al., 2021). While several models, such as Bertweet and other transformer-based architectures, have demonstrated high accuracy in detecting explicit hate speech, these approaches frequently fail to consider the broader context of user interactions that can amplify harmful behavior (Yigezu et al., 2023).

Existing datasets for hate speech detection often focus solely on content-level features, ignoring the relational dynamics between users on social media platforms. For instance, user interactions, such as comments and replies, can provide valuable context for assessing the intent and impact of a particular post (Unsvåg and Gambäck, 2018). The absence of datasets that integrate user interaction graphs with post content limits the ability of existing models to account for these contextual factors. To address this gap, our study mined a dataset of posts, comments, and replies from the Reddit community "r/PoliticalDiscussion," a forum known for its heated and polarized debates, to construct a directed user interaction graph. This graph serves as the foundation for our engagement-aware scoring system, which incorporates both content-level and interaction-based features.

Our proposed engagement-aware scoring system seeks to enhance the prioritization of hateful content for moderation by leveraging three key components: (1) content-level hate scores derived from transformer-based models such as Bertweet, (2) user-level aggregated scores that reflect the average hatefulness of a user's interactions, and (3) graph-based neighborhood influences, which quantify the extent to which a user's network contributes to the propagation of hate speech. By integrating these components, we aim to capture not only the explicit hatefulness of a given post but also the implicit dynamics of its surrounding context (Fortuna et al., 2022).

The introduction of graph-based modeling into hate speech detection aligns with a growing body of research that emphasizes the importance of considering relational data in NLP tasks. Graph neural networks (GNNs) and other graph-based approaches have been successfully applied in domains such as recommendation systems, fraud detection, and sentiment analysis, highlighting their potential for capturing complex relationships between entities. In the context of hate speech detection, incorporating user interaction graphs enables a deeper understanding of how harmful content spreads and influences online communities.

Another critical aspect of our study is the evaluation framework, which emphasizes human-centered assessment. While automated metrics such as Spearman correlation provide a quantitative measure of ranking performance, we also incorporate manual labeling to validate the real-world utility of our proposed system. Human evaluators compared the ranked lists generated by our engagement-aware model with those produced by a baseline Bertweet model, assessing which lists more accurately prioritize hateful content. This dual evaluation approach ensures that our findings are both statistically robust and practically relevant (Elangovan et al., 2024) (Schuff et al., 2023).

The results of our study demonstrate that the engagement-aware scoring system outperforms the baseline in both Spearman correlation and human-labeled metrics, highlighting its effectiveness in prioritizing hateful content for moderation. By integrating content-level and interaction-based features, our approach addresses key limitations of existing models and offers a more nuanced understanding of hate speech in social media contexts. These findings underscore the importance of considering relational data in hate speech detection and open avenues for future research on graph-based moderation systems.

In the following sections, we detail the methodology employed to construct the dataset and develop the engagement-aware scoring system, describe the evaluation framework, and present the results of our comparative analysis. We also discuss the broader implications of our findings for content moderation and outline potential directions for further research.

2 Methodology

This section outlines the pipeline we developed for detecting and propagating hate speech in social media posts. It is structured into five subsections: Data Collection, Hate Speech Detection, User-User Graph Construction, User Toxicity Scoring, and Low-Confidence Hate Speech Propagation. Each subsection details the methodology employed to facilitate reproducibility and comprehensiveness. By integrating text-level classification with graph-based analysis, our approach seeks to address the nuanced challenges posed by ambiguous hate speech and relational dynamics among social media users.

2.1 Data Collection

The first hurdle in our methodology was data collection. Social media data, particularly datasets containing both post content and user metadata, is sparse and often fragmented. Publicly available datasets frequently consist of either textual content without metadata or user metadata without content, limiting their utility for building interaction-based models. To address this gap, we decided to create our own dataset by mining Reddit, a platform known for its structured discussions and research-friendly APIs (Adams, 2022).

Reddit's subreddit structure, particularly "r/PoliticalDiscussion," provided an ideal setting for this study. Given the recent elections and the high prevalence of polarized discourse, this subreddit offered a rich source of posts likely to contain varying levels of hate speech. Using Reddit's developer APIs, we retrieved all posts, comments, and replies from the past month. The dataset included both textual content and metadata such as the author's username, timestamps, and interaction details, enabling us to construct a directed user-user graph later in the pipeline.

For every comment and reply, we also collected the corresponding parent entity and its author, facilitating the mapping of interaction flows. This step was critical for constructing the user-user graph, as it allowed us to accurately capture the relationships between users based on their interactions.

This robust data collection methodology ensured that our dataset was well-suited for studying both text-based hate speech detection and interaction-driven propagation dynamics.

Entity	Description
Post	Main threads initiated by a user
Comment	Responses to posts
Reply	Responses to comments

Table 1: Entities and their description

2.2 Hate Speech Detection

Once the dataset was collected, the next step in the methodology involved classifying the collected textual entities—posts, comments, and replies—into categories based on their level of hate speech. This process is crucial for identifying and labeling the content appropriately before further analysis. For this, we relied on pre-trained transformer-based language models, which have been shown to perform effectively in text classification tasks, particularly in hate speech detection.

Initially, we experimented with two off-the-shelf models: `pysentimiento/bertweet-hate-speech` (Pérez et al., 2021) (Nguyen et al., 2020) and `unitary/toxic-bert` (Hanu and Unitary team, 2020). After a thorough comparison of their outputs on a subset of the dataset, we chose the `pysentimiento/bertweet-hate-speech` model, as it demonstrated superior performance on our data. The `unitary/toxic-bert` model, while robust in some scenarios, was less effective in capturing the nuances and ambiguities present in Reddit discussions, particularly in the `r/PoliticalDiscussion` subreddit, where contextual subtleties and sarcasm are common.

The selected model assigned a hate speech confidence score to each text entity. To categorize the entities effectively, we established the following thresholds:

- **Hateful (score > 0.7):** Texts in this category were labeled as hateful with high confidence. These texts often contained explicit slurs, threats, or other abusive content that clearly violated community guidelines.
- **Non-Hateful/Neutral (score < 0.3):** Texts with low scores were labeled as neutral or non-hateful, indicating no significant signs of toxicity.
- **Ambiguous/Low-Confidence (score 0.3–0.7):** Texts falling within this range were flagged as ambiguous. These entities were difficult to classify confidently due to subtleties in tone, context, or phrasing, necessitating further analysis.

By categorizing entities in this manner, we ensured that the downstream processes focused on resolving ambiguities in hate speech detection, which are often overlooked in traditional classification approaches. Ambiguities in hate speech detection pose a significant challenge, as they can arise from factors such as sarcasm, cultural nuances, or implicit biases in language models (Lee et al., 2024).

The ambiguous entities (score 0.3–0.7) represented a particularly interesting subset, as their classification could significantly impact the effectiveness of hate speech detection systems. To address these entities, we proposed a novel approach incorporating graph-based propagation of hate scores, as detailed in subsequent sections.

2.3 User-User Graph Construction

The next step involved constructing a user-user interaction graph to capture the relationships and interactions between Reddit users in our dataset. A graph-based approach was necessary to address the contextual and relational aspects of hate speech propagation, which are often missed by traditional text-only models (Elinas et al., 2019).

2.3.1 Graph Structure and Components

We defined the user-user graph $G = (V, E)$, where:

- **Nodes (V):** Represent individual Reddit users.
- **Edges (E):** Represent directed interactions between users. These interactions included:
  - *Comments:* Edges directed from the commenter to the original post author.
  - *Replies:* Edges directed from the replier to the comment author.
- **Edge Weights ($w(u, v)$):** Quantify the frequency of interactions between users u and v. For example, multiple replies or comments between the same pair of users resulted in higher edge weights.

2.3.2 Construction Process

To construct the graph, we leveraged the parent entity and author metadata collected during the data collection phase. For each comment or reply, we established a directed edge between the authors of the interacting entities. This step ensured that the graph accurately reflected the dynamics of user interactions in the subreddit.

The weight of each edge was calculated as the total number of interactions between the two users. For instance, if user A commented on three posts authored by user B and replied to two comments made by user B, the edge weight $w(A, B)$ was set to 5. This weighting mechanism allowed the graph to capture not only the existence of relationships but also their intensity.

2.3.3 Graph Analysis

The resulting user-user graph served as a foundation for identifying patterns of toxicity propagation. With a total of $11,155$ nodes (users) and $49,183$ edges (interactions), the graph reflected the sparsity typical of social media networks, with a density of 0.000395. The average edge weight of 1.58 indicated relatively low interaction frequency between user pairs, suggesting that most interactions occurred in small, isolated clusters.

2.3.4 Significance of the Graph

The graph construction was critical for propagating hate scores through user interactions. By capturing the relational structure of the subreddit, the graph enabled us to identify how toxicity spreads from one user to another and how interactions with toxic users influence the behavior and content of their neighbors. This graph-based perspective provided a deeper understanding of the dynamics of hate speech on social media and laid the groundwork for calculating user toxicity scores and propagating low-confidence hate scores.

By integrating user-user graph construction into the pipeline, our methodology addressed the limitations of text-only approaches and highlighted the importance of relational data in understanding hate speech dynamics.

2.4 User Toxicity Scoring

Once the hate scores for posts, comments, and replies were assigned, the next step involved calculating a toxicity score for each user. This step was critical for understanding the behavior and toxicity levels of individual users, as well as for propagating toxicity through the graph-based framework.

To measure user-level toxicity, we defined the toxicity score of a user as the percentage of their posts labeled as hateful. Formally, the toxicity score for a user u is calculated as follows:

$$Toxicity\ Score(u) = \frac{\#\ hateful\ posts\ of\ u}{\#\ total\ posts\ of\ u} \times 100$$

where $\#\ hateful\ posts\ of\ u$ is the number of posts authored by user u that were classified as hateful, and $\#\ total\ posts\ of\ u$ represents the total number of posts authored by the same user.

This toxicity score provides a quantitative measure of a user's overall behavior in the subreddit. Users with higher toxicity scores were flagged as having a greater likelihood of contributing to the propagation of hate speech. It is important to note that this metric is relative and depends on the volume of content a user has contributed to the dataset.

Furthermore, the toxicity scores were leveraged in downstream tasks, such as evaluating the influence of toxic users on their neighbors in the interaction graph. This allowed us to capture both direct and indirect contributions to hate speech dynamics, adding an additional layer of granularity to our analysis. This approach aligns with prior research emphasizing the importance of user-level metrics in content moderation (Schaffner et al., 2024).

2.5 Low-Confidence Hate Speech Propagation

After assigning toxicity scores to individual users, we focused on resolving the ambiguity of low-confidence hate speech (i.e., entities with hate scores between 0.3 and 0.7). For these cases, we proposed a novel *Engagement Aware Score*, which combines three key components: the content's hate score, the user's toxicity score, and the influence of the user's immediate neighbors in the interaction graph.

The *Engagement Aware Score* for a post is calculated as:

Engagement Aware Score = 0.6 × *content hate score* + 0.3 × *user hate probability* + 0.1 × *neighborhood hate*

Here:

- **Content Hate Score:** The hate score assigned to the post, comment, or reply by the hate speech detection model.
- **User Toxicity:** The toxicity score of the post's author, as calculated in the previous subsection.
- **Neighborhood Hate:** A weighted average of the toxicity scores of the user's neighbors in the interaction graph, multiplied by the

weights of the edges connecting them. Formally:

$$H(u) = \frac{\sum_{v \in N(u)} w(u,v) \times T(v)}{\sum_{v \in N(u)} w(u,v)} \quad (1)$$

where $H(u)$ represents the neighborhood hate score of user u, $T(v)$ is the toxicity score of user v, $N(u)$ represents the set of neighbors of user u, and $w(u,v)$ is the weight of the edge between users u and v.

By combining these components, the Engagement Aware Score integrates content-level and relational data to assign a more informed score to low-confidence posts. This approach captures both the intrinsic properties of the content and the broader social context in which it was created.

The Engagement Aware Scores were used to rank ambiguous entities, providing a prioritized list for manual evaluation by moderators. This allowed us to efficiently address the most problematic cases while minimizing false positives and negatives. By leveraging graph-based propagation, this method also highlights the importance of social dynamics in understanding and mitigating hate speech (He et al., 2019).

3 Results

3.1 Dataset Statistics

We analyzed a dataset comprising **291** Reddit posts, **46,093** comments, and **31,651** replies from a single subreddit, `r/PoliticalDiscussion`. The dataset represents contributions from **11,155** distinct users. These statistics underscore the richness and diversity of the collected data, which provides a strong basis for hate speech detection and propagation analysis.

3.2 User-User Graph

The user-user interaction graph constructed from the dataset contains **11,155** nodes (representing users) and **49,183** directed edges (representing interactions such as comments and replies). The average edge weight is **1.58**, indicating that, on average, users interacted approximately 1.58 times with others.

The graph's density is calculated as **0.000395**, which reflects the sparsity of user interactions typical of large social media networks. The graph captures meaningful relational dynamics, enabling the analysis of how toxicity propagates across the user network.

3.3 Hate Speech Classification

Using the hate speech detection model, the collected entities were classified as follows:

- **1.3%** of posts were labeled as hateful (score > 0.7).
- **95.5%** of posts were labeled as non-hateful/neutral (score < 0.3).
- **3.09%** of posts were classified as ambiguous (low-confidence hate scores, $0.3 \leq$ score ≤ 0.7).

This classification highlights the small proportion of posts explicitly identified as hateful, which aligns with the observed distribution of hate speech in social media platforms. However, the presence of 3.09% ambiguous posts indicates the need for additional analysis, as these posts may have a significant influence on the propagation of toxicity.

3.4 Low-Confidence Posts and Toxic Connections

For posts with low-confidence hate scores (0.3–0.7), we applied graph-based propagation rules to refine their classification. Using the *Engagement Aware Score*, we calculated a combined score for each post by integrating content-level, user-level, and neighborhood-level features. These scores were used to generate a ranked list, prioritizing ambiguous posts for moderation.

Our analysis revealed that posts connected to highly toxic users or highly interactive regions in the graph exhibited increased hate scores, even if their initial scores were ambiguous. This observation underscores the value of incorporating user interactions and relational dynamics into hate speech detection pipelines.

These engagement-aware scores were subsequently evaluated against the baseline content-only scores to assess their effectiveness in improving the classification and prioritization of ambiguous posts.

4 Evaluation

4.1 Spearman Rank Correlation

To evaluate the effectiveness of our engagement-aware scoring system, we used Spearman Rank Correlation, a non-parametric measure of rank similarity between two ordered lists. In this context, we compared the ranked lists generated by:

Model	Spearman Correlation	Our Metric (Unweighted Manual)	Our Metric (Weighted Manual)
Baseline	0.128 (Top 35)	16	16
Ours	0.293 (Top 35)	24	101

Table 2: Evaluation Results Comparing the Baseline and Engagement-Aware Models

- The baseline model, which ranked entities based solely on their content hate scores.
- Our engagement-aware model, which incorporated content-level, user-level, and neighborhood-level features.

We also generated a ranked list using GPT-4-o's outputs as the ground truth. GPT-4-o's rankings served as a benchmark for comparison, as they reflect a human-like understanding of hate speech intensity.

Spearman Rank Correlation is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference in ranks assigned to an entity by two models, and n is the number of entities being ranked.

Using this metric, we evaluated the top 35 entries from both models against the ground truth. Our engagement-aware model achieved a higher Spearman Rank Correlation compared to the baseline, as shown in Table 2.

The significant improvement in rank correlation demonstrates the effectiveness of our engagement-aware scoring system in better aligning with human judgments of hate speech intensity.

4.2 Manual Evaluation

We define two types of manual evaluation methods for comparing the performance of our baseline model against our proposed model: **unweighted** and **weighted manual metrics**.

4.2.1 Unweighted Manual Metric

This method involves comparing two ranked lists generated by the models. The evaluation is performed by a human judge to determine which model's content is more hateful. The process works as follows:

1. **Initialization**: Two pointers are initialized, one for each list. Let `pointer_A` refer to the baseline model's list and `pointer_B` refer to our model's list. Both pointers start at the beginning of their respective lists.
2. **Comparison**: A human evaluator compares the content at `pointer_A` and `pointer_B`:
   - If the content from list A is judged to be more hateful, a point is awarded to list A, and `pointer_A` is advanced to the next entry.
   - If the content from list B is judged to be more hateful, a point is awarded to list B, and `pointer_B` is advanced to the next entry.
3. **Addressing False Positives**: To prevent one list from dominating the evaluation due to repeated false positives, we impose a *max age limit of 3* for each pointer. This means if a pointer has not moved for 3 consecutive comparisons, it must be advanced to the next entry, regardless of the evaluation result.
4. **Termination**: The process continues until we have evaluated all 35 elements from at least one of the lists.

This method provides a straightforward comparison between the models' outputs without considering positional weighting.

4.2.2 Weighted Manual Metric

The weighted metric extends the unweighted approach by assigning scores based on the relative positions of the pointers. The weight scores are calculated as follows:

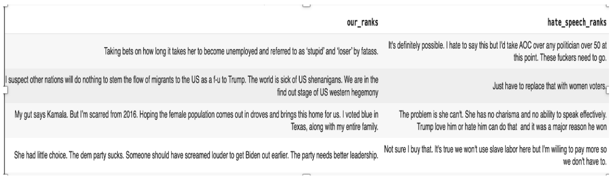
- If list A is judged more hateful, the score for list A is increased by $\max(\texttt{pointer\_A} - \texttt{pointer\_B}, 1)$. This gives higher scores when list A's more hateful content appears significantly later than list B's current entry.
- If list B is judged more hateful, the score for list B is increased by $\max(\texttt{pointer\_B} - \texttt{pointer\_A}, 1)$.

This weighting scheme rewards models more heavily when their later-ranked entries are judged more hateful than earlier entries in the competing list, emphasizing the importance of ranking quality.

4.2.3 Evaluation Process

Both metrics are applied to compare the lists until one of the lists has fewer than 35 elements left to evaluate. The results of these evaluations are summarized in Table 2, while a sample list used for manual evaluation is illustrated in Figure 1.

This dual approach provides both an unbiased baseline comparison (unweighted metric) and an emphasis on ranking quality (weighted metric), offering a comprehensive evaluation of the models' ability to identify and rank hateful content effectively.



our_ranks	hate_speech_ranks
Taking bets on how long it takes her to become unemployed and referred to as 'stupid' and 'loser' by fatass.	It's definitely possible. I hate to say this but I'd take AOC over any politician over 50 at this point. These fuckers need to go.
I suspect other nations will do nothing to stem the flow of migrants to the US as a f-u to Trump. The world is sick of US shenanigans. We are in the find out stage of US western hegemony	Just have to replace that with women voters.
My gut says Kamala. But I'm scarred from 2016. Hoping the female population comes out in droves and brings this home for us. I voted blue in Texas, along with my entire family.	The problem is she can't. She has no charisma and no ability to speak effectively. Trump love him or hate him can do that and it was a major reason he won
She had little choice. The dem party sucks. Someone should have screamed louder to get Biden out earlier. The party needs better leadership.	Not sure I buy that. It's true we won't use slave labor here but I'm willing to pay more so we don't have to.

Figure 1: Sample results of our ranked list

5 Discussion

Our results demonstrate the advantages of incorporating user interactions and relational dynamics into hate speech detection pipelines. The user-user graph enabled us to model the propagation of toxicity through a social network, capturing patterns that are not evident in text-only approaches. Specifically, the graph's structure highlighted the influence of highly toxic users on their immediate neighbors, revealing the role of network connectivity in the spread of hateful content.

The *Engagement Aware Score* proved effective in refining the classification of ambiguous posts, leveraging content hate scores, user toxicity, and neighborhood interactions. This integration of features addressed the limitations of traditional models that rely solely on content-level analysis. By prioritizing low-confidence posts for moderation, our approach aligns with the practical needs of social media platforms, where resource constraints demand a focus on the most problematic cases.

The improvement in Spearman Rank Correlation and the higher scores in our custom metrics reflect the robustness of our approach. The engagement-aware model outperformed the baseline, producing rankings that more closely align with human judgments of hate speech intensity. These results highlight the importance of combining machine learning with social network analysis to develop nuanced hate speech detection systems.

However, certain limitations of our study merit discussion. First, the reliance on a single subreddit (`r/PoliticalDiscussion`) may limit the generalizability of our findings. While this subreddit provided a rich dataset for analysis, future studies should evaluate the pipeline across diverse communities and platforms to validate its broader applicability. Second, the thresholds used for hate speech classification (e.g., 0.7 for hateful, 0.3 for neutral) were determined empirically and may require adjustment for other contexts. Finally, while the graph-based approach captured local interactions effectively, incorporating higher-order graph features such as community detection or centrality measures could further enhance the model.

6 Conclusion and Future Work

In this paper, we presented a hybrid pipeline for hate speech detection that integrates text-level classification with graph-based toxicity propagation. By leveraging a user-user graph, we modeled the spread of hate speech through social interactions, enabling the identification and prioritization of low-confidence posts. The proposed *Engagement Aware Score* effectively combined content, user, and neighborhood-level features, resulting in a significant improvement over baseline models in terms of alignment with human evaluations.

Our findings underscore the importance of incorporating relational data into hate speech detection systems. The ability to model the influence of toxic users and their interactions provides a deeper understanding of the dynamics of hateful content on social media.

Looking ahead, there are several avenues for future work:

1. **Expanding to Diverse Platforms:** Evaluating the pipeline across other social media platforms (e.g., Twitter, Facebook) to assess its generalizability and adaptability to different data structures and user behaviors.

2. **Incorporating Advanced Graph Features:** Exploring higher-order graph features such as community detection, centrality measures, and graph neural networks to capture complex interaction patterns.

3. **Developing Interventions:** Designing intervention strategies to mitigate the spread of hate speech, focusing on identifying and ad-

dressing toxic users and high-risk communities.

4. **Improving Model Explainability:** Enhancing the interpretability of the scoring system to support moderators in understanding and addressing hate speech more effectively.

5. **Evaluating Impact:** Conducting longitudinal studies to measure the real-world impact of deploying graph-based hate speech detection systems on online communities.

By addressing these directions, future research can build on the foundation established in this work to develop more comprehensive and effective approaches to combating hate speech online.

References

Norman Adams. 2022. 'scraping' reddit posts for academic research? addressing some blurred lines of consent in growing internet-based research trend during the time of covid-19. *International Journal of Social Research Methodology*, 27(1):47–62.

Joëlle Albrecht, Jérôme Endrass, Michal Dreifuss, Nina Schnyder, and Astrid Rossegger. 2024. Prevalence and impact of hate speech among politicians in switzerland. *Societies*, 14(7):98.

Naceur Amara. 2024. Hate speech, challenges and ways of confrontation. *Istanbul Journal of Advanced Studies*, 4(1):33.

Melisa Castellanos, Alexander Wettstein, Sebastian Wachs, Julia Kansok-Dusche, Cindy Ballaschk, Norman Krause, and Ludwig Bilz. 2023. Hate speech in adolescents: A binational study on prevalence and demographic differences. *Frontiers in Education*, 8:1076249.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. *arXiv preprint*.

Pantelis Elinas, Anna Leontjeva, and Yuriy Tyshetskiy. 2019. Can graph machine learning identify hate speech in online social networks? Accessed: December 7, 2024.

Paula Fortuna, Mónica Domínguez, Leo Wanner, and Zeerak Talat. 2022. Directions for nlp practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11794–11805, December 7–11, 2022. Association for Computational Linguistics.

Laura Hanu and Unitary team. 2020. Detoxify. GitHub repository, https://github.com/unitaryai/detoxify.

Yipeng He, Di Bai, and Wantong Jiang. 2019. Cs224w: Detection and analysis of hateful users on twitter. https://snap.stanford.edu/class/cs224w-2019/project/26424341.pdf. Project report for Stanford University course CS224W.

Gábor Kovács, Pablo Alonso, and Rishabh Saini. 2021. Challenges of hate speech detection in social media. *SN Computer Science*, 2(2):95.

Zachary Laub. 2019. Hate speech on social media: Global comparisons. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons. Updated June 7, 2019. Accessed: 2024-12-03.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, To be confirmed. Association for Computational Linguistics. Accessed: December 7, 2024.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. "community guidelines make this the best party on the internet": An in-depth study of online platforms' content moderation policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–16. ACM.

Hannah Schuff, Laura Vanderlyn, Hady Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, 29(5):1199–1222.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The effects of user features on twitter hate speech detection. In *Proceedings of the Second Workshop on Abusive Language Online (ALW2)*, pages 75–85, Brussels, Belgium. Association for Computational Linguistics. Presented October 31, 2018.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Transformer-based hate speech detection for multi-class and multi-label classification. In *CEUR Workshop Proceedings*, volume 3496, Jaén, Spain. CEUR-WS.org. Presented at IberLEF 2023.