AggregHate: An Efficient Aggregative Approach for the Detection of Hatemongers on Social Platforms

Anonymous ACL submission

Abstract

The automatic detection of online hate speech serves as a crucial step in the detoxification of the online discourse. Moreover, accurate classification can promote a better understanding of the proliferation of hate as a social phenomenon. While most prior work focus on the detection of hateful *utterances*, we argue that focusing on the *user* level is as important, albeit challenging. In this paper we consider a multimodal aggregative approach for the detection of hate-mongers, taking into account the potentially hateful texts, user activity, and the user network. We evaluate our methods on three unique datasets X (Twitter), Gab, and Parler showing that a processing a user's texts in her social context significantly improves the detection of hate mongers, compared to previously used text and graph-based methods. Our method can be then used to improve the classification of coded messages, dog-whistling, and racial gas-lighting, as well as inform intervention measures. Moreover, our approach is highly efficient even for very large datasets and networks.

Disclaimer: The illustrative examples in Table 1 may be offensive to some readers.

1 Introduction

The rising popularity of social platforms enhanced the hateful content targeting minorities and led to the proliferation of online hate speech (Waseem and Hovy, 2016; Laub, 2019). Accordingly, there is a growing body of research on the appearance and magnitude of hate speech on social media, in general, (Knuttila, 2011; Chandrasekharan et al., 2017; Zannettou et al., 2018), and on hate speech detection, in particular, (Saleem et al., 2017; Waseem and Hovy, 2016; Davidson et al., 2017). Hate speech is not merely an online inconvenience as shooting, bombing, stabbing, beating, and vandalism are often linked to online activity (Munn, 2019; Malevich and Robertso, 2019; Thomas, 2019; McIlroy-Young and Anderson, 2019; Mathew et al., 2019; ADL, 2023).

Hate is not promoted by isolated individuals, but rather by communities that often exist within larger communities. Shifting the focus from the post level to the user level can be beneficial on various levels (pun intended): it provides a better understanding of the dynamics of hate groups and the ways hate infects individuals and communities; it could be reapplied in the post level, improving the detection of coded language, dog whistling and gas-lighting; finally, it could provide the social structure in which the impact of different intervention techniques[1] could be tested (Thomas and Wahedi, 2023).

Many of the challenges inherent to hate speech classification on the utterance level are discussed by Talat et al. (2018); Fortuna et al. (2022); Velankar et al. (2022); Kim et al. (2024), among others. A recent survey of the applicability of state-of-the-art LLMs highlights the shortcomings of the use of LLMs for nuanced tasks such as hate, toxicity and empathy (Ziems et al., 2024).

We provide some illustrative examples in Table 1, complemented with a short analysis, demonstrating the specific challenges and outlining the possible benefits of shifting our attention from the utterance (textual) level to the speaker (user) level.

The explicit racism of utterance #1 is detected by all models, however, deHateBERT

[1] Intervention alternatives depend on legal regulation and platform policy and range from the removal of a specific post, the flagging of a specific post, to the temporal or indefinite suspension of an account.

	Text	DHB	JS	FTDB	GPT4	Gemini 1.5Pro
1	"Muslim terrorists did this and obama brought them here to run our government. Gotta get rid of Obammy, Omar and all these other fucking Muslims"	.57	.9	.98	✓	✓
2	"is there a more debased creature in our modern pop-cultural cesspool than (((Amy Schumer)))? Probably not."	.22	.35	.92	✓?	✓
3	"We must secure the existence of our people and a future for white children!"	.03	.36	.02	✓	✓?
4	"If white people are going to continue existing - we must support our race"	.06	.4	.69	✓?	✓?
5	"When the youngest white generation becomes adult they will be 5% of the world population, down from 1/3 in 1900"	.06	.25	.01	✗	✓?
6	"Did you know that most dog breeds were bred by whites? We're awesome and they're awesome!"	.2	.25	.65	✗	✗
7	"White lives matter!"	.02	.24	.02	✗	✗?
8	"Blue lives matter!"	.02	.01	.01	✗	✗?

Table 1: Examples of hate-promoting texts. DHB: the hate-likelihood of the text assigned by deHateBERT (Aluru et al., 2020); JS: hate prediction by Google's Jigsaw; FTDB: DistillBERT (Sanh et al., 2019a) fine-tuned on our datasets; GPT-4 and Gemini 1.5 Pro, using the prompt "Which of the following texts should be classified as hate speech:". ("Which of the following utterances may be perceived as hate speech?"). A **'?'** postscript indicates that a nuanced prediction was generated by the model (see full generated predictions in Appendix A.

(Aluru et al., 2020) (DHB) – a BERT model fine-tuned for the detection of hate speech – assigned a likelihood of only 0.57 for it to be classified as hate speech. Three of the models, ChatGPT-4 included, failed to detect the antisemitism conveyed in utterance #2 as it requires "external" knowledge, namely that Schumer is of Jewish heritage and that the (((echo))) symbol is a (newly) recognized hate symbol[2]. The text in utterance #3, known as the '14 words' – "the most popular white supremacist slogan in the world"[3]. However, it may appear innocent to the uninformed (human) eye, as well as to models that did not encounter it (in context) in training. As a result, it was identified only by ChatGPT-4 and Gemini 1.5 Pro. Posts #4-7 are thinly veiled allusions to the "14 words". The fine-tuned distilBERT (FTDB) captures #4 and #6 but misses the "14 words". Both #7 and #8 twist the slogan 'Black Lives Batter' – originally a protest against police brutality. While #7 carries an explicit white-supremacist tone[4], #8[5] is not recognized as hate-speech, although it is often used alongside racial slurs. Even state-of-the-art models such as ChatGPT-4 and Gemini 1.5 Pro either completely fail to detect the hate conveyed in utterances 4-8, or flag it with a low confidence.

While all of the texts in Table 1 pose a challenge to computational models and humans alike, contextual information and careful aggregation can be used to achieve classification on the user level: multiple implicit posts (coded, ambiguous, dog-whistling, or gas-lighting) posted by a single user can reinforce the weak signal obtained from a single post. Similarly, association with other users or explicit posts can reinforce a weak signal of a context-less post.

Contextual aggregation, however, is not straightforward as it depends on a number of philosophical and practical considerations stemming from the likelihood or the confidence of a model (or a human) to assign the class for a specific post. For example, one can argue that a single, though explicit, hateful post may not be enough to label the user as a racist or a hate-monger. However, even if a single explicit post is sufficient – what would be the user label in case of two *implicit* posts, or a few dozen of posts, each is predicted to be hateful with a low confidence? Looking at the illustrative posts in Table 1, we propose a principled way to effectively combine predictions and modalities in order to achieve an accurate classification.

To this end we explore three fundamentally

[2] www.adl.org/resources/hate-symbol/echo

[3] www.adl.org/resources/hate-symbol/14-words

[4] www.adl.org/resources/hate-symbol/white-lives-matter

[5] A response to the BLM movement, suggesting that attacking policemen should be considered a hate crime.

different approaches for contextual aggregation: (i) using binary weights with a fixed threshold, (ii) using a relational aggregation conditioned on the social context, and (iii) using sidtributional aggregation conditioned on aggregated confidence levels. Finally, we combine these methods to forn a multimodal classification model.

Contribution Our contribution of this work is threefold:

1. We propose a robust and efficient multimodal aggregative approach for the detection of hate-mongers.
2. We demonstrate the benefits of contextual aggregation over three unique datasets (Twitter, Gab, and Parler).
3. We share a novel annotated dataset of Parler hate.

2 Related Work

A comprehensive overview of methods and benchmarks for hate-speech detection is provided by Alkomah and Ma (2022), while many of the challenges in current approaches are surveyed by ElSherief et al. (2021); Velankar et al. (2022); Fortuna et al. (2022), among others. Subjectivity and the incomplete definition of hate-speech are addressed by Khurana et al. (2022) while the limitations of transfer learning for the task are demonstrated by Israeli and Tsur (2022), and a set of functional tests to evaluate the performance of different models trained on different benchmarks, types, targets, and languages were proposed by Röttger et al. (2021, 2022).

A taxonomy of implicit hate was developed and shared by ElSherief et al. (2021). Unfortunately, most of the annotated tweets are no longer available, thus user level aggregation cannot be achieved.

A growing number of works has shifted the attention from the utterance level to the user level. Waseem and Hovy (2016) analyze the relation between demographic features and hate speech, while Ribeiro et al. (2017) explores the differences between account meta-features of hateful and non-hateful Twitter users. Both works are of an exploratory nature.

Arviv et al. (2021) detects hate mongers using a multi-modal architecture that combines three streams of post-level predictions: the tweets of the target user, her followers, and her followees.

A two-step approach considering both the textual and the network modalities was proposed by Ribeiro et al. (2018) and extended by Israeli and Tsur (2022): In the first step seed nodes (users) are detected based on the textual signal (keyword matching in Ribeiro et al. and a fine-tuned BERT in Israeli and Tsur). In the second step a diffusion model is applied in order to propagate the initial hate assignments across the social network.

A number of works use Graph Neural Networks (GNN) in order to detect hate speech or hate mongers. Li et al. (2021) presented HateGNN, using textual similarity and the appearance of predefined hate terms as part of the objective that produces the node embeddings to be used for classification. The learned embeddings depend heavily on the training data, the type of hate, and the supervision. Miao et al. (2022) proposed an end-to-end framework, enriching a BERT classifier with Graph Attention Networks. However, in spite of the use of graph networks, this approach is applied for the detection of hate only on the post level rather than on the user level. Several supervised, unsupervised and semi supervised models were explored by (Das et al., 2021), including the state-of-the-art GNN models such as AGNN (Thekumparampil et al., 2018) in order to classify users as hateful and non-hateful.

Nirmal et al. (2024) proposed *SHIELD* - a framework that leverages LLM-extracted rationales to augment a base hate speech detection model to facilitate faithful interpretability. However, it was demonstrated that fine-tuned classification models outperform state-of-the-art LLMs over tasks that involve social nuances, e.g., detection of humor, empathy, toxicity and hate (Ziems et al., 2024).

3 Multimodal Aggregative Approaches

3.1 Aggregative Approaches

Utterance-level Classification (C^T) The basic building block of all aggregative approaches is the classification of a single utterance. Any classification model can be used for the utterance level as long as its output can be interpreted as a probability (e.g., by applying the sigmoid function). The probability is needed in order to assign a "confidence" threshold, before making the binary decision. Formally:

$$C^T(t) = \begin{cases} 1 & \theta(t) \geq \tau^T \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Where t is a text snippet (post, tweet, utterance), θ is a classification model, and the hyper parameter τ^T is the post-level threshold reflecting the sensitivity of the model (or the community) to implicit forms of hate speech. In this work we follow Israeli and Tsur (2022) and use a DistilBERT classifier (Sanh et al., 2019b) fine-tuned for hate-speech detection on our three datasets (see Section 4).

User-level Classification (C^U) The detection of hateful *users* is inherently related to the user's posts. Given a user u and a the user level threshold τ^U, the generic user classification function is given by:

$$C^U(u) = \begin{cases} 1 & \Theta(u) \geq \tau^U \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Naive Aggregation with Fixed Threshold Given T^u – the set of posts published by u, a naive aggregative approach could simply count the number of hateful posts (decided by the C^T) published by a user u:

$$\Theta(u) = \sum_{t \in T^u} C^T(t) \quad (3)$$

While naive aggregation does not require any training, the threshold τ^U can be used to control the sensitivity of the model: setting $\tau^U = 1$ implies zero-tolerance (often too harsh[6]) approach, while setting $\tau^U = a$, $a \in \{10, 20, 30, ...\}$ implies a more conservative approach, requiring a user to be a "repeated offender" in order to be labeled as a hate-monger. For convenience, we denote to the naive aggregation (fixed threshold) Θ_F, and the naive classification function C_F.

In order to achieve flexibility and robustness one should consider other aggregation approaches that account for the nuanced and implicit utterances as well as for the social context in which a user is embedded. The different aggregative approaches could be combined to a user feature vector and Θ could be trained to optimize the weight of each feature. Specifically, instead of a naive aggregation we consider two types of aggregative functions: Relational Aggregation (Θ_R) and Distributional Aggregation (Θ_D). Given a an annotated dataset, we train Θ to optimize C^U, where Θ is a logistic regression classifier.

In the remainder of this section we motivate the use of each aggregative approach and provide the formal definition of the model.

Relational Aggregation The Aristotelian concept of man as a social and political animal informed decades of sociological research concerning the importance of community to individual identity (McMillan and Chavis, 1986; Wellman and Gulia, 1999). Unfortunately, identity is often forged by association with hate groups, off and on-line (Gordon, 2017; Govers et al., 2023). It is therefore reasonable to assume that hateful content circulating in u's ego network should be taken into account and inform the label assigned to u. That is, even if a conservative τ^U is used, the associates of u may push her over the threshold.

Formally, given $G(V, E)$ – a directed social network where V is the set of users and $(u, v) \in E$ indicates a directed edge $u \to v$, we define $\overleftarrow{u}$ and $\overrightarrow{u}$ as the sets of followers and followees of u, respectively. In the relational case, $\Theta_R(u)$ is a linear combination of three terms – the naive aggregation over u's texts and the percentages of hateful users among his followers and followees:

$$\Theta_R(u) = \alpha \cdot C_F(u) + \beta \cdot \frac{1}{|\overleftarrow{u}|} \sum_{v \in \overleftarrow{u}} C_F(v) + \gamma \cdot \frac{1}{|\overrightarrow{u}|} \sum_{v \in \overrightarrow{u}} C_F(v)$$

The values of α, β, and γ are optimized through training, essentially reflecting the importance of the user's posts and the posts of his followers and followees.

Distributional Aggregation While relational aggregation takes the social context into account, it does not address variations the intensity of the promoted hate in terms

[6]Remember that in the standard case $\tau^T = 0.5$ thus a user will be labeled a hate-monger even if $\theta(t) = 0.51$ for one of her posts and $\theta(t') < 0.5 \quad \forall t' \in T^u / \{t\}$.

of “comitement” (what percentage of a user's stream is hate) or implicitness (e.g., the user tries to tread the fine line without violating the platform rules). For example, consider two users u and v posting utterances #3–#8 in Table 1. However, imagine that these are the only posts published by u, while v published hundreds of other posts, non of which is hateful. Intuitively, one may argue that v's questionable posts are diluted, compared to u's, thus we should require a higher level of confidence in order to assign him the 'hate monger' label (and compare to a user v' posting the more explicit #1 & #2 along with hundreds of non-hateful posts).

In order to address these variations we use distributional aggregation: instead of counting the user's hateful posts (Θ_F), we look at the distribution of the user's posts as a k-dimensional vector where k determines the number of bins used to approximate the probability density function.

Specifically, we consider two distributional spaces – bins and quantiles. In the bin-based representation the [0,1] range is divided to k equal bins, each bin (entry in the vector) holds the number of posts with the corresponding hate score assigned by $\theta(t)$. In the quantile-based representation the k bins are unique for each user u as they are defined over the range $[min(\theta(t)), max(\theta(t)]$ for $t \in T^u$. We therefore want to optimize

$$\Theta_D(u) = \sum_{i=1}^{k} w_i \cdot \sigma(B_i(u)) \qquad (4)$$

where $B_i(u)$ denotes the number of utterances for which $\theta(t)$ falls in the ith bin and σ is the softmax function.

Multimodal Aggregation Finally, the different aggragation methods can be used together in a multimodal manner – combining the hate-score distribution of the user utterances and the hate levels in the user's ego network. Using Θ_D^b and Θ_D^q to denote the bin-based and quantile-based representations we can formulate the combined model as

$$\Theta(u) = \Theta_D^b(u) + \Theta_D^q + \Theta^R$$

3.2 Social-aware Baselines

In this section we briefly describe five strong baseline algorithms we use for comparison. These five algorithms, all leverage the structure of the social network and have proved useful in an array of node classification tasks, including the detection of hate speech and hateful users.

DeGroot's Diffusion The DeGroot's model is a simple yet strong framework to classify nodes in a network through belief propagation. The model achieved good results in the detection of hateful users on Gab and Parler (Ribeiro et al., 2018; Israeli and Tsur, 2022).

Graph Neural Networks (GNNs) Das et al. (2021) explored several GNN methods for detecting hateful users on Gab and Twitter. The network embeddings incorporate nodal features, e.g., textual representations of the user's utterances in learning the node and graph embeddings. In this work we use the following four algorithms:

1. **GCN**: Graph Convolutional Networks (Kipf and Welling, 2016) uses localized approximation in learning node embeddings through convolutional layers.
2. **GAT**: Graph Attention Network learn nodal representations by combining features of the nodes in the ego network of the focal node, setting their importance via attention layers.
3. **GraphSAGE**: The Graph Sample and Aggregate (Hamilton et al., 2018) learns nodal embedings by sampling a predefined number of nodes from the ego network of a focal node u, then applying attention layers in order to assign the importance of each node in the aggregative representation.
4. **AGNN**: The Attention-based Graph Neural Network (Thekumparampil et al., 2018) uses attention layers to learn a dynamic and adaptive local summary of the neighborhood of each focal node.

4 Datasets and Annotation

We evaluate our aggregative approaches over three very different datasets: Twitter-Echo, Gab, and Parler. The number of posts, users, and label breakdown for each dataset are provided in Table 2. The remainder of this section provides further details regarding the datasets and our annotation process.

Dataset	Source	Raw Data #Posts	Raw Data #Users	Annotated Data #Posts	Annotated Data % Hate	Annotated Data #Users	Annotated Data % Hate
Echo (Twitter)	Arviv et al. (2021)	18M	7.07K	4630	8.2%	1000	15.4%
Gab	Arviv and Tsur (2021)	22M	336.7K	5000	5.1%	1000	24.8%
Parler	This work	$183M^{\dagger}$	$4.08M^{\dagger}$	$8262^{\ddagger}$	$32.9\%^{\ddagger}$	890	25.4%

Table 2: Datasets statistics. The raw Parler data (marked †) were shared by (Aliapoulios et al., 2021) and the post-level annotations (‡) were shared by (Israeli and Tsur, 2022). User-level annotations are shared as part of this paper.

Echo (Twitter) The triple parentheses, or triple brackets, also known as the (((echo))), is an antisemitic symbol that is used to highlight the names of individuals of Jewish background (e.g., actress and comedian Amy Schumer, see utterance #2 in Table 1), organizations owned by Jewish people (e.g., Ben & Jerry's), or organizations accused of promoting "Jewish globalist values"' (e.g., the International Monetary Fund). The Echo dataset curated by Arviv et al. (2021) contains over 18M English tweets posted by ∼7K echo users between May and June 2016. Annotations are provided at the tweet and the user level. An important feature of this dataset is that all users have utterances containing the echo symbol, although some users use it in a non-hateful manner, e.g., to symbolize a hug. This ambiguous nature of the symbol makes hate detection challenging.

Gab Gab, launched on August 2016, was created as an alternative to Twitter, positioning itself as putting "people and free speech first", welcoming users suspended from other social networks. Gab posts (called *gabs*) are limited to 300 characters, and users can repost, quote or reply to previously created gabs. Gab permits pornographic and obscene content, as long as it is labeled *NSFW* ('not safe for work').

The raw Gab dataset was introduced by Zannettou et al. (2018). It was collected using Gab's API with the snowball methodology. More specifically, the researchers obtained data for the most popular users as returned by Gab's API and iteratively collected data from all of their followers and their followees. They collected three types of information: basic details about Gab accounts (including username, score, and date of account creation); all the posts for each Gab user in the dataset; and all the followers and followees of each user, which allow the reconstruction of a social network. Overall, this dataset contains 22.1M posts from 336.8K users, posted between August 2016 and January 2018. Arviv and Tsur (2021) shared an annotated dataset based on the aforementioned above, containing 60K labeled posts and 1K labeled users.

Parler Alluding to the French verb 'to speak', Parler was launched on August 2018.[7] The platform branded itself as "The World's Town Square" a place to *"Speak freely and express yourself openly, without fear of being "deplatformed" for your views"*[8].

Parler users post texts (called *parlays*) of up to 1000 characters. Users can reply to parlays and to previous replies. Parler supports a reposting mechanism similar to Twitter's retweets (called 'echos', not to confuse with the (((echo))) hate symbol, see above). Parler's official guidelines[9] explicitly allowed "trolling" and "not-safe-for-work" (NSFW) content, include only three "principles" prohibiting "unlawful acts", citing "Obvious examples include: child sexual abuse material, content posted by or on behalf of terrorist organizations, intellectual property theft".

Aliapoulios et al. (2021) presented a dataset of tens of millions of Parler messages. Israeli and Tsur (2022) used this dataset to introduce an annotated dataset for hate speech (post level). Their 10K dataset consists of 3224 posts (32.8%) labeled as hateful and 6053 (59.8%)

[7] On April 2023 the platform was acquired by Starboard and was taken offline to "undergo a strategic assessment" (Starboard announcement on Parler's landing page `https://parler.com/`, accessed: 5/8/2023). The platform was relaunched in February 2024 announcing it is "breaking free from the constraints of conventional platforms" (accessed: 6/5/2024).

[8] Parler branding on its landing page (accessed: 3/10/2022)

[9] https://parler.com/documents/guidelines.pdf (accessed: 4/17/2022)

as non-hateful. However, a Parler *user-based* annotated dataset has yet to be introduced. Hence, as part of this research, we create the first annotated dataset of Parler users.

Annotation of Parler Users The Parler dataset presented by Aliapoulios et al. (2021) consists of ∼4M users. We focus on a subset of users matching the following criteria: (i) The account exists for at least six months; (ii) The user showed some activity (posted at least 30 posts); (iii) The primary language of the user is English. These three rules left us with a subset of users, denoted U^*, from which we sampled users for annotation. Following the protocol used by Ribeiro et al. (2018) and Israeli and Tsur (2022), we used stratified subsampling mitigate bias (most users and vast majority of posts are not hateful).

Annotation was done by 94 senior year Data Science B.Sc students for bonus course credit. Annotators were introduced to Parler and were given explicit instructions about the annotation task. The annotation process involved rating each account on a 1–5 scale (non-hateful – extremely hateful). We ensured that each user is annotated by three annotators. The full annotation guidelined and further details regarding the annotation protocol are available in Appendix C.

5 Results and Analysis

5.1 Experimental Settings

For the utterance-level classification, we finetune DistilBERT (Sanh et al., 2019a) on each datasets. We used a batch size of 32, a maximum number of epochs of 20 and a validation split of 0.2. We also used an early stopping callback with a patience of 5 epochs on the validation loss.

For the user-level classification, we use 5-Fold cross-validation for all of the methods. For each dataset, we considered the largest (weakly) connected component. Specifically, for the GNN methods we used the same experimental settings as proposed by Das et al. (2021). Appendix B provides further details regarding the full networks, the number of connected components and the statistics regarding the largest connected component in each dataset.

For the DeGroot's method, we followed the protocol used by Israeli and Tsur (2022).

5.2 Results

Utterance Level Prediction All aggregative models require an initial step of classification of individual utterances. While the focus of this paper is the classification on the *user* level, we report the results on the utterance level in order to highlight the challenge and the benefit of the aggregative methods.

Results for each dataset are presented in Table 4. Note the very low F-score achieved ob the Gab dataset and compare to the significant improvement achieved by shifting to the user level (below). The density of the mean hate score per user is presented in Figure 1, highlighting the differences between the datasets.

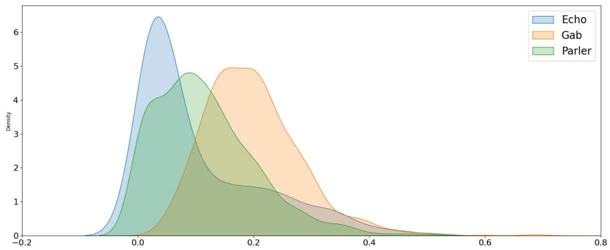

Figure 1: The Density of the mean utterance scores per user for each Dataset

User level results Detailed results of each method over each of the three datasets are presented in Table 3.

Looking at the F-score achieved by the different methods we observe that our aggregative approach consistently outperform the baseline algorithms. While the multimodal aggregation yields the best results in the Echo and Gab datasets, it is ranked third on the Parler dataset, with the relational aggregation ranked first. We speculate that this is may be attributed to the size and the unique characteristics of the Parler network (see Appendix B). A thorough analysis of the social networks is out of the scope of this paper.

Breaking the multimodal aggregative model to separate aggregative models (Relational, Distributional-bins, Distributional-quantiles and bins and quantiles combined) shows that in most cases all these approaches are competitive with the top performing models and outperform the strong baselines.

	Method	Precision	Recall	F1	ROC AUC
Echo	DeGroot's Diffusion	0.472 ± 0.389	0.255 ± 0.261	0.320 ± 0.310	0.610 ± 0.122
	GCN	0.443 ± 0.118	0.914 ± 0.096	0.585 ± 0.104	0.797 ± 0.102
	GraphSAGE	0.629 ± 0.124	$\mathbf{0.950 \pm 0.041}$	0.752 ± 0.088	0.944 ± 0.024
	GAT	0.548 ± 0.131	0.653 ± 0.241	0.574 ± 0.123	0.772 ± 0.091
	AGNN	0.759 ± 0.072	0.914 ± 0.054	0.826 ± 0.035	0.963 ± 0.032
	Fixed-Threshold	0.654 ± 0.063	0.627 ± 0.095	0.633 ± 0.040	0.836 ± 0.040
	Relational Aggregation	$\mathbf{0.820 \pm 0.060}$	0.834 ± 0.055	0.825 ± 0.042	0.956 ± 0.013
	Distributional (bins)	0.772 ± 0.045	0.871 ± 0.058	0.817 ± 0.042	0.944 ± 0.019
	Distributional (quantiles)	0.747 ± 0.064	0.899 ± 0.047	0.815 ± 0.053	0.942 ± 0.018
	Distributional (bins+quantiles)	0.757 ± 0.058	0.885 ± 0.052	0.815 ± 0.049	0.946 ± 0.021
	Multimodal (relational+bins+quantiles)	0.781 ± 0.023	0.899 ± 0.048	$\mathbf{0.836 \pm 0.028}$	$\mathbf{0.961 \pm 0.011}$
Gab	DeGroot's Diffusion	0.314 ± 0.001	$\mathbf{0.777 \pm 0.000}$	0.447 ± 0.001	0.604 ± 0.000
	GCN	0.241 ± 0.109	0.678 ± 0.428	0.334 ± 0.169	0.594 ± 0.059
	GraphSAGE	0.317 ± 0.066	0.559 ± 0.206	0.388 ± 0.077	0.582 ± 0.077
	GAT	0.194 ± 0.115	0.457 ± 0.387	0.264 ± 0.174	0.501 ± 0.060
	AGNN	0.340 ± 0.031	0.600 ± 0.190	0.423 ± 0.030	0.679 ± 0.024
	Fixed-Threshold	$\mathbf{0.497 \pm 0.078}$	0.351 ± 0.066	0.411 ± 0.070	0.722 ± 0.039
	Relational Aggregation	0.408 ± 0.061	0.437 ± 0.086	0.419 ± 0.063	0.675 ± 0.043
	Distributional (bins)	0.461 ± 0.034	0.649 ± 0.044	0.538 ± 0.024	0.763 ± 0.016
	Distributional (quantiles)	0.429 ± 0.027	0.702 ± 0.056	0.532 ± 0.033	0.770 ± 0.018
	Distributional (bins+quantiles)	0.435 ± 0.026	0.714 ± 0.043	0.540 ± 0.029	0.769 ± 0.016
	Multimodal (relational+bins+quantiles)	0.452 ± 0.027	0.702 ± 0.042	$\mathbf{0.550 \pm 0.032}$	$\mathbf{0.777 \pm 0.020}$
Parler	DeGroot's Diffusion	0.395 ± 0.221	0.441 ± 0.247	0.417 ± 0.233	0.644 ± 0.081
	GCN	0.284 ± 0.054	$\mathbf{0.760 \pm 0.404}$	0.348 ± 0.130	0.644 ± 0.145
	GraphSAGE	0.309 ± 0.092	0.649 ± 0.189	0.394 ± 0.028	0.497 ± 0.061
	GAT	0.379 ± 0.051	0.731 ± 0.164	0.488 ± 0.013	$\mathbf{0.746 \pm 0.044}$
	AGNN	0.369 ± 0.081	0.552 ± 0.255	0.416 ± 0.082	0.667 ± 0.069
	Fixed-Threshold	0.470 ± 0.050	0.369 ± 0.040	0.412 ± 0.035	0.693 ± 0.026
	Relational Aggregation	$\mathbf{0.519 \pm 0.073}$	0.509 ± 0.082	$\mathbf{0.513 \pm 0.074}$	0.730 ± 0.051
	Distributional (bins)	0.284 ± 0.037	0.500 ± 0.075	0.362 ± 0.049	0.575 ± 0.024
	Distributional (quantiles)	0.324 ± 0.018	0.734 ± 0.057	0.449 ± 0.025	0.611 ± 0.040
	Distributional (bins+quantiles)	0.324 ± 0.021	0.738 ± 0.051	0.450 ± 0.027	0.618 ± 0.034
	Multimodal (relational+bins+quantiles)	0.370 ± 0.023	0.680 ± 0.082	0.478 ± 0.035	0.699 ± 0.044

Table 3: 5-Fold CV results on the test sets of Echo, Gab, and Parler datasets using the best (F1-score-wise) configuration.

Dataset	Precision	Recall	F1 Score
Echo	0.4122	0.8026	0.5446
Gab	0.2057	0.5472	0.2990
Parler	0.6316	0.8177	0.7127

Table 4: Performance metrics of the utterance-level model for Echo, Gab, and Parler Datasets

Interestingly, looking at the results of the relational method, we observe that the importance of the different components differ across datasets: best F1-score was achieved using $\alpha = 0.608$, $\beta = 0.776$, $\gamma = 1.467$ for the Echo dataset; $\alpha = 0.776$, $\beta = 0.085$, $\gamma = 0.108$ for the Gab dataset and $\alpha = 0.239$, $\beta = 0.254$, $\gamma = 0.24$ for Parler. This result highlights the importance of the network structure and the dynamics and norms of each platform. This is also evident from the distribution of the mean utterance score in each platform.

6 Conclusion

We proposed a robust and efficient multimodal aggregative method, combining text and social context through relational and distributional aggregations. We demonstrated the benefits of this approach for the task of hate speech and hatemonger detection over three unique and very different datasets from three social platforms: X (Twitter), Gab and Parler.

Future work takes three trajectories: (i) Improving the integration of the different modalities, and (ii) Better understanding the unique characteristics of each of the platforms and the datasets and the ways they impact the results, and (iii) Reapplication of the aggregative methods to improve classification of evasive utterances and emerging forms of coded hate speech.

7 Limitations and Ethical Considerations

7.1 Limitations

This work has a number of limitations, two of which are listed below (and other are referred to in the directions for future work):

(i) All three aggregation procedures depend on the basic classifier $\theta(t)$. An evasive user that is careful with his words, using only coded language or consistent gas-lighting may not be identified at all if $\theta(t) < \tau^T$ for all $t \in T^u$ for a user u.

(ii) We use the fixed threshold τ^U as a strong baseline, however, a more comprehensive comparison to other existing methods (e.g., diffusion-based and GNNs) should be considered in future work.

7.2 Ethical Considerations

The detection of hate mongers on social platforms and the detoxification of the discourse are of great importance. However, wrongly labeling a user, may have severe emotional and legal implications. In this work, we explore computational approaches that can be used to study the phenomena at scale. As a practical policy, user labeling should be done carefully, ideally involving a 'man-in-the-loop' procedure.

References

Center on Extremism ADL. 2023. Shooter in allen, texas, embraced antisemitism, misogyny and white supremacy.

Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. 2021. An early look at the parler online social network. *arXiv preprint arXiv:2101.03820*.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13:273.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Eyal Arviv, Simo Hanouna, and Oren Tsur. 2021. It'sa thin line between love and hate: Using the echo in modeling dynamics of racist online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 61–70.

Eyal Arviv and Oren Tsur. 2021. From individuals to communities : community-aware language modeling for the detection of hate speech.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31.

Mithun Das, Punyajoy Saha, Ritam Dutt, Pawan Goyal, Animesh Mukherjee, and Binny Mathew. 2021. You too brutus! trapping hateful users in social media: Challenges, solutions & insights. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 79–89.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Paula Fortuna, Monica Dominguez, 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*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Linda Gordon. 2017. *The second coming of the KKK: The Ku Klux Klan of the 1920s and the American political tradition*. Liveright Publishing.

Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Down the rabbit hole: Detecting online extremism, radicalisation, and politicised hate speech. *ACM Computing Surveys*, 55(14s):1–35.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive representation learning on large graphs. *Preprint*, arXiv:1706.02216.

Abraham Israeli and Oren Tsur. 2022. Free speech or free hate speech? analyzing the proliferation of hate speech in parler. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 109–121.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jaehoon Kim, Seungwan Jin, Sohyun Park, Someen Park, and Kyungsik Han. 2024. Label-aware hard negative sampling strategies with momentum contrastive learning for implicit hate speech detection. *Preprint*, arXiv:2406.07886.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lee Knuttila. 2011. User unknown: 4chan, anonymity and contingency. *First Monday*, 16(10).

Zachary Laub. 2019. Hate speech on social media: Global comparisons.

Shu Li, Nayyar A Zaidi, Qingyun Liu, and Gang Li. 2021. Neighbours and kinsmen: hateful users detection with graph neural network. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11–14, 2021, Proceedings, Part I*, pages 434–446. Springer.

Simon Malevich and Tom Robertso. 2019. Violence begetting violence: An examination of extremist content on deep web social networks.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Reid McIlroy-Young and Ashton Anderson. 2019. From "welcome new gabbers" to the pittsburgh synagogue shooting: The evolution of gab. In *Proceedings of the international aaai conference on web and social media*, volume 13, pages 651–654.

David W McMillan and David M Chavis. 1986. Sense of community: A definition and theory. *Journal of community psychology*, 14(1):6–23.

Zhenxiong Miao, Xingshu Chen, Haizhou Wang, Rui Tang, Zhou Yang, and Wenyi Tang. 2022. Detecting offensive language on social networks: An end-to-end detection method based on graph attention networks. *arXiv preprint*.

Luke Munn. 2019. Alt-right pipeline: Individual journeys to extremism online.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. *Preprint*, arXiv:2403.12403.

Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2017. "like sheep among wolves": Characterizing hateful users on twitter. *arXiv preprint arXiv:1801.00317*.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. *arXiv preprint arXiv:2206.09917*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*, pages 29–55. Springer International Publishing, Cham.

Kiran K. Thekumparampil, Chong Wang, Sewoong Oh, and Li-Jia Li. 2018. Attention-based graph neural network for semi-supervised learning. *Preprint*, arXiv:1803.03735.

Daniel Robert Thomas and Laila A Wahedi. 2023. Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences*, 120(24):e2214080120.

Elise Thomas. 2019. Aspi explains: 8chan.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. A review of challenges in machine learning based automated hate speech detection. *Preprint*, arXiv:2209.05294.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Barry Wellman and Milena Gulia. 1999. Virtual communities as communities. *Communities in cyberspace*, pages 167–194.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014. International World Wide Web Conferences Steering Committee.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291.

A Prompts and Generated Responses

```
Which of the following utterances may be
perceived as hate speech?

1. Good morning boys and girls!

2. Muslim terrorists did this and
   Obama brought them here to run our
   government. Gotta get rid of Obammy,
   Omar, and all these other fucking
   Muslims.

3. Is there a more debased creature in
   our modern pop-cultural cesspool than
   (((Amy Schumer)))? Probably not.

4. We must secure the existence of
   our people and a future for white
   children!.

5. If white people are going to continue
   existing - we must support our race.

6. When the youngest white generation
   becomes adult they will be 5% of the
   world population, down from 1/3 in
   1900.

7. Did you know that most dog breeds were
   bred by whites? We're awesome and
   they're awesome!

8. White lives matter!

9. Blue lives matter!

10. Black lives matter!
```

GPT-4 The utterances that may be perceived as hate speech are those that explicitly or implicitly promote hate, discrimination, or hostility towards particular groups based on their ethnicity, religion, or other inherent characteristics. Based on the examples you've provided, here's an analysis of which might be considered hate speech:

2. **"Muslim terrorists did this and Obama brought them here to run our government. Gotta get rid of Obammy, Omar, and all these other fucking Muslims"** - This statement is clearly hate speech. It uses derogatory language targeting Muslims and makes unfounded, harmful generalizations about an entire religious group.

3. **"Is there a more debased creature in our modern pop-cultural cesspool than (((Amy Schumer)))? Probably not."** - This statement might be perceived as hate speech, particularly because the use of triple parentheses is often a coded way to identify someone as Jewish in an anti-Semitic context.

4. **"We must secure the existence of our people and a future for white children!"** - This phrase is widely recognized as a white supremacist slogan, known as the "Fourteen Words." It is used to promote the idea of white racial superiority and is considered hate speech.

5. **"If white people are going to continue existing - we must support our race"** - This could be interpreted as promoting racial supremacy or separatism, which might be viewed as hate speech depending on the context in which it is used.

The other statements, while potentially controversial or promoting particular views, do not explicitly convey hate against a group based on inherent characteristics in the same way that the identified statements do.

Gemini 1.5 Pro Analysis of Potentially Hateful Utterances: Here's a breakdown of the provided utterances and their potential to be perceived as hate speech:

1. **Good morning boys and girls!**
   - This is a simple greeting and does not contain any hateful content.

2. **Muslim terrorists did this and Obama brought them here to run our government. Gotta get rid of Obammy, Omar and all these other fucking Muslims.**
   - This statement employs harmful stereotypes against Muslims, associating them with terrorism and expressing a desire to remove them from the government. This is a clear example of hate speech.

3. **Is there a more debased creature in our modern pop-cultural cesspool than (((Amy Schumer)))? Probably not.**

- The use of triple parentheses around a name is a known anti-semitic symbol online. This statement is likely intended to target Amy Schumer for her Jewish heritage and is therefore hate speech.

4. **We must secure the existence of our people and a future for white children!**

5. **If white people are going to continue existing - we must support our race**

6. **When the youngest white generation becomes adult they will be 5% of the world population, down from 1/3 in 1900**

    - These three statements, while not explicitly hateful, promote a white supremacist ideology. They focus on the preservation of white people and imply a fear of their decline, potentially inciting hostility towards other racial groups.

7. **Did you know that most dog breeds were bred by whites? We're awesome and they're awesome!**

    - This statement, while expressing pride in an accomplishment associated with white people, does not directly attack or demean any other group. It is unlikely to be considered hate speech.

8. **White lives matter!**

9. **Blue lives matter!**

10. **Black lives matter!**

    - These statements, on the surface, simply affirm the value of different groups of people. However, depending on the context and intent, they can be interpreted in different ways.
    - *White lives matter* is often used as a response to the "Black Lives Matter" movement, implying a denial of the systemic racism faced by Black people. In such cases, it can be considered hate speech.
    - *Blue lives matter* expresses support for law enforcement. While not inherently hateful, it can be used to dismiss or oppose calls for police accountability and racial justice.
    - *Black lives matter* aims to highlight the systemic racism and violence faced by Black communities. It is a call for justice and equality, not a statement of superiority or hatred towards other groups.

It's crucial to consider the context and intent behind these statements to determine if they constitute hate speech.

B Networks Statistics

Tables 6 and 7 provide more information on each of the datasets and justifies the focus on the largest connected component in each.

C Annotation Guidelines

Each annotator was assigned 50 Parler accounts, ensuring that each account is being annotated by three students. For each account, annotators were given the user name, self-description of the user's account, and 30 posts published by the user. The 30 posts are sampled out of the following: (i) 15 posts with the highest hate speech prediction values[10]; (ii) Its 10 most recent posts; (ii) Five random posts published by the account. The annotators showed a reasonable agreement level with an average Cohen's Kappa of 0.36. The full guidelines for the Parler dataset annotation process can be found here `https://mega.nz/file/D0pjBDDR#9iMK0uC72w41BjSj_eDCmpbaJ6YGU8sZ8g0dvDv0pcU`

D Responsible NLP Research Checklist

A1 Did you describe the limitations of your work? **Yes**

A2 Did you discuss any potential risks of your work? **Yes**

A3 Do the abstract and introduction summarize the paper's main claims? **Yes**

B Did you use or create scientific artifacts? **Yes**

[10]The prediction is according to the BERT model introduced by Israeli and Tsur (2022)

Dataset	Threshold	Precision	Recall	F1	ROC AUC
Echo	1	0.264	1.000	0.417	0.506
	3	0.266	1.000	0.420	0.511
	10	0.277	1.000	0.434	0.538
	50	0.358	0.950	0.520	0.673
	100	0.433	0.820	0.567	0.721
Gab	1	0.252	1.000	0.402	0.506
	3	0.257	1.000	0.409	0.519
	10	0.281	0.984	0.437	0.573
	50	0.380	0.559	0.452	0.628
	100	0.437	0.351	0.389	0.600
Parler	1	0.308	0.883	0.457	0.608
	3	0.341	0.824	0.482	0.644
	10	0.348	0.689	0.462	0.627
	50	0.457	0.387	0.420	0.616
	100	0.607	0.293	0.395	0.615

Table 5: Performance metrics for Echo, Gab, and Parler datasets

Dataset	#Posts	#Users	#Edges	#Connected Components	#Singletons
Echo (Twitter)	18M	7.07K	21.4K	9075	2919
Gab	19.42M	61.36K	2.63M	40K	10.13K
Parler (this work)	115M	3.08M	11.14M	5.45M	2.43M

Table 6: Datasets statistics for the full network, excluding users without any posts.

B1 Did you cite the creators of artifacts you used? **Yes**

B2 Did you discuss the license or terms for use and / or distribution of any artifacts? **Yes, we discuss the limitations and ethical considerations of our work**

B3 Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)? **Yes**

B4 Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it? **No, since our work involves harmful or offensive content.**

B5 Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.? **Yes**

B6 Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? **Yes**

C Did you run computational experiments? **Yes**

C1 Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used? **Yes**

C2 Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?] **Yes**

C3 Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether

	Raw Data					Annotated Data	
Dataset	#Posts	#Users	#Edges	Clustering Coefficient	Optimal Gamma	#Users	% Hate
Echo (Twitter)	9.8M	3.7K	20.7K	0.19	2.8	532	26.1%
Gab	19.28M	51.2K	2.47M	0.402	4.06	982	24.5%
Parler (this work)	60.7M	643K	11.4M	0.224	2.14	881	25.2%

Table 7: Datasets statistics for the largest (weakly) connected component. We treated the graphs as undirected in order to calculate the Clustering Coefficient and the Optimal Gamma (assuming a Power-law distribution)

you are reporting the max, mean, etc. or just a single run? **Yes**

C4 If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)? **Yes**

D Did you use human annotators (e.g., crowdworkers) or research with human participants? **Yes**

D1 Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.? **Yes**

D2 Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.? **Yes**

D3 Did you discuss whether and how consent was obtained from people whose data you're using/curating **Yes**

D4 Was the data collection protocol approved (or determined exempt) by an ethics review board? **N/A**

D5 Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data? **Yes**

E Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing? **Yes** [E1] Did you include information about your use of AI assistants? **We used LLMs as classifiers, see Table 1 and references through the paper.**