Hope vs. Hate: Understanding User Interactions with LGBTQ+ News Content in Mainstream US News Media through the Lens of Hope Speech

Jonathan Pofcher♣ **Christopher M. Homan**♣
Randall L. Sell♢ **Ashiqur R. KhudaBukhsh**♣[*]
♣Rochester Institute of Technology
♢Drexel University
jep7453@rit.edu, cmhvcs@rit.edu, randy@drexel.edu, axkvse@rit.edu

Abstract

This paper makes three contributions. First, via a substantial corpus of 1,419,047 comments posted on 3,161 YouTube news videos of major US cable news outlets, we analyze how users engage with LGBTQ+ news content. Our analyses focus both on positive and negative content. In particular, we construct a *hope speech* classifier that detects positive (*hope speech*), negative, neutral, and irrelevant content. Second, in consultation with a public health expert specializing on LGBTQ+ health, we conduct an annotation study with a balanced and diverse political representation and release a dataset of 3,750 instances with crowd-sourced labels and detailed annotator demographic information. Finally, beyond providing a vital resource for the LGBTQ+ community, our annotation study and subsequent in-the-wild assessments reveal (1) strong association between rater political beliefs and how they rate content relevant to a marginalized community, (2) models trained on individual political beliefs exhibit considerable in-the-wild disagreement, and (3) zero-shot large language models (LLMs) align more with liberal raters.

Trigger Warning: this paper contains offensive material that some may find upsetting.

1 Introduction

From crowdsourced storymapping projects sharing stories of love, loss, and a sense of belonging (Kirby et al., 2021) to safe, anonymous spaces to seek resources (McInroy et al., 2019) and dating platforms (Blackwell et al., 2015) – the internet and modern technologies play a positive role in the health and well-being of the LGBTQ+ community in various ways. However, cyberbullying (Abreu and Kenny, 2018), exposure to dehumanization through news media (Mendelsohn et al., 2020), and more recently, homophobic biases in large language models (LLMs) (Dutta et al., 2024a, 2025a) - are some of the modern technology perils the community grapples with.

How do mainstream US cable news outlets cover the LGBTQ+ community and how do users interact with such news content? Via a substantial corpus of more than 130 million YouTube comments on more than 300K YouTube news videos uploaded on three major US cable news outlets, this paper investigates how LGBTQ+ discussions are situated in mainstream US political discourse. While prior literature has analyzed LGBTQ+ discourse in print media through the lens of a dehumanization framework (Mendelsohn et al., 2020), such analyses consulted a single source of news (The New York Times) and did not consider reader responses to LGBTQ+ news. In this work, we consider three major US cable news networks representing diverse political views (Fox News, CNN, and MSNBC) with a key goal to understand how users interact and engage with LGBTQ+ news content. In doing so, we not only focus on identifying negative discussions about the LGBTQ+ community, one of our key contributions is operationalizing the detection of *hope speech* championing the disadvantaged minority in the broader spirit of counter-speech literature (Benesch, 2014; Mathew et al., 2019; Palakodety et al., 2020b; Chakravarthi, 2020; Hengle et al., 2024).

While hate speech detection and mitigation strategies have been extensively researched (see, e.g., Fortuna and Nunes, 2018), relatively little attention has gone into the opposite: *hope speech*. Defined as comments and posts that inspire optimism and diffuse hostility in online spaces (Palakodety et al., 2020b; Chakravarthi, 2020), hope speech is crucial for groups who face a disproportionate amount of hate speech. In the current US cultural climate, where issues of race, gender, and sexuality are central to public debate (Hartman, 2019), there is a need for tools that can detect both hope speech and negative content as

---

[*] Ashiqur R. KhudaBukhsh is the corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19862–19888
November 4-9, 2025 ©2025 Association for Computational Linguistics

YouTube Comment	Label
This is why we have 500+ million guns in america. We need to start using them. If a man walked in my daughters bathroom, that would be his last day on earth. People arent gonna keep tolerating this. Theres about to be rainbow street justice youre gonna start seeing. When comes to my, or anyones kids, and i frankly dont care what your laws. Never did in the first place, but when it comes to my kid. . . . . . ill wake up and chose violence EVERY time...	Negative
Hey if using the term f[redacted]g is hate speech, then I stand guilty as charged. I hate what they stand for and I hate the corruption they're bringing into society, and I especially hate that they are preying on the minds of children in an attempt to make them believe that their sick, perverted lifestyle is normal...	Negative
LISTEN PEOPLE: EVEN IF YOU DONT SUPPORT IT, YOU CAN RESPECT US AT LEAST. ALL WE WANT IS BASIC HUMAN DECENCY. LITERALLY JUST THAT. me and my girlfriend got spat at on a bus once, and then called homophobic slurs. we just want respect, being gay doesn't change the fact that I'm a human...	Positive
Reading these comments are making me cry. I'm 13 and my Dad is transgender. 99% of Doctors accept it. Why are you one to argue with Science? Please, if you are a bigot, don't reply to this comment. It will make me cry. I love my Dad, and you ignorant people who don't accept science won't change that....	Positive

Table 1: Illustrative examples of positive and negative YouTube comments identified by our classifier in the wild.

listed in Table 1 and increase understanding of relevant online conversations. According to the FBI, one in five hate crimes in the US targets the LGBTQ+ community (Federal Bureau of Investigation, 2024), and globally, homosexual activity remains punishable by death in six UN member states while only 37 countries recognize same-sex marriage (Mendos and Rohaizad, 2024). Understanding and improving online discourse around LGBTQ+ issues is thus a vital step toward protecting a vulnerable community.

Contributions. Our contributions are as follows.
(1) In consultation with a health expert specializing in LGBTQ+ health for more than two decades, we curate a novel dataset of 3,750 instances with labels: *neutral*, *irrelevant*, *positive* (hope speech), and *negative*[1]. Each instance is labeled by three raters (one Republican, one Democrat, and one independent) ensuring diverse and balanced political perspectives. 98 raters self-identified as being part of the LGBTQ+ community.
(2) We analyze the association between rater political beliefs and how they rate content relevant to the LGBTQ+ community.
(3) We analyze the alignment of LLMs with political beliefs in connection with LGBTQ+ content and how rater biases affect model fine-tuning.
(4) We provide novel insights into audience engagement patterns surrounding LGBTQ+ content, enhancing our understanding of LGBTQ+ discussions within mainstream US political discourse. To our knowledge, no such study exists at our scale.

2 Related Work

Our work is closely related with the *hope speech* literature (Palakodety et al., 2020b,a; Chakravarthi, 2020; Yoo et al., 2021) and the broader literature of counter speech (Benesch, 2014; Mathew et al., 2019; Hengle et al., 2024; Saha et al., 2022; Gupta et al., 2023; Saha et al., 2024). Our work contributes to this literature through our unique focus on LGBTQ+ discourse in mainstream US politics and our investigation on rater subjectivity and political leanings.

Political polarization in the US is widely studied across multiple disciplines in rich and diverse settings that include congressional voting patterns on policy issues (Poole and Rosenthal, 1984), mate selection (Huber and Malhotra, 2017), allocation of scholarship funds (Iyengar and Westwood, 2015), and annotating offensive content (Sap et al., 2022; Weerasooriya et al., 2023). Prior research showed systematic differences in offensive speech annotation based on annotators' beliefs (Sap et al., 2022; Weerasooriya et al., 2023) and experiences (Patton et al., 2019). While associations between political leanings and annotation of negative content has been studied in US politics (Sap et al., 2022; Weerasooriya et al., 2023) and hot-button issues like reproductive rights and gun control (Weerasooriya et al., 2023), our work extends this in two ways: first by examining perceptions of positive rather than negative content, and second by focusing specifically on LGBTQ+ discourse. Our study thus contributes to the broader literature of annotator subjectivity (Pavlick and Kwiatkowski, 2019; Sap et al., 2019; Al Kuwatly et al., 2020; Larimore et al., 2021; Sap et al., 2022; Goyal et al., 2022; Pei and Jurgens, 2023; Weerasooriya et al., 2023; Homan et al., 2024; Prabhakaran et al., 2024).

Our annotation study design is grounded in the prior literature (Sap et al., 2022; Weerasooriya et al., 2023) and draws from Weerasooriya et al., 2023 and Crowl et al., 2025 in seeking diverse and

---

[1] The dataset is available at `https://github.com/Social-Insights-Lab/LGBTQplus_HopeSpeech/`.

balanced political perspectives including from the independents. Our work also touches upon political bias in LLMs as we observe that zero-shot classification of LGBTQ+ content of several models are more aligned with liberal raters (Feng et al., 2023; Bang et al., 2024). To address class imbalance, we construct an ensemble active learning pipeline much akin to Palakodety et al., 2020a and Khorramrouz et al., 2023 leveraging well-known active learning strategies (e.g., certainty sampling (Sindhwani et al., 2009) and margin sampling (Scheffer et al., 2001)).

3 Dataset and Annotation Study Design

Figure 1 presents a schematic diagram of our data collection process. which will be described in more detail in the following section.

3.1 YouTube News Videos

We consider a dataset of 333,032 YouTube videos uploaded between 04/10/2008 to 08/31/2024 by the official YouTube channels of three major US cable news networks: Fox News, CNN, and MSNBC. First introduced in KhudaBukhsh et al., 2021, this dataset represents a reliable snapshot of US political discourse encompassing diverse political perspectives over a substantial period of time. This dataset has found prior use in studying election-related discourse (KhudaBukhsh et al., 2022; Mittal et al., 2024), health-related discourse (Yoo and KhudaBukhsh, 2023), and rater subjectivity (Weerasooriya et al., 2023; Pandita et al., 2024; Dutta et al., 2025b). We extend this dataset using the same data collection method to include videos up to August 2024.

3.2 Identifying LGBTQ+ Relevant Videos

We prune our initial set of videos using a two-step process. First, for each video, we pass the video title and description to `GPT-4o mini` prompting it to classify the video as specfcially relevant to the LGBTQ+ community or not (Appendix contains prompt details in § D). This step prunes our initial set of 333,032 videos down to 5,435 videos. Next, for each of these 5,435 videos, we use the same prompt and pass the video title and description to `GPT-4o`. While the `GPT-4o` model performed better and ended up with fewer false positives, we use this two-step process because it was considerably more expensive to run `GPT-4o` on all 333,032 videos. Overall, we obtain 3,161 relevant videos, denoted $\mathcal{V}_{LGBTQ+}$. A random sample of 300 videos from this pipeline (100 from each channel) yielded 2 false positives for *related* and no false negatives for *unrelated*, indicating that our pruning steps are robust[2].

3.3 Constructing $\mathcal{D}_{hope}$

For each video in $\mathcal{V}_{LGBTQ+}$, we collect user comments using publicly available YouTube API yielding 1,419,047 comments. From these comments, we construct $\mathcal{D}_{hope}$, a dataset that categorizes content based on its stance toward the LGBTQ+ community. In consultation with the health expert specializing on LGBTQ+ health research for more than two decades, we identify four broad categories: *positive* (*hope speech*); *negative*; *irrelevant*; and *neutral*. We define *hope speech* as content that expresses support, advocacy, or acceptance for the LGBTQ+ community. *Negative* content is defined as content that expresses opposition, discrimination, or hostility towards the LGBTQ+ community. *Neutral* content is defined as content that is relevant to LGBTQ+ issues but does not belong to any of the previous two categories (*hope speech* and *negative*). *Irrelevant* content is defined as content that is not related to LGBTQ+ issues. Detailed annotation guidelines are present in the Appendix (§ E).

In brief, this process consists of three steps:

1. We first construct a seed set of 1,950 instances using a collaborative human-LLM annotation framework. This seed set is annotated by crowd-sourced annotators for the aforementioned four categories (details are described in § A.1). This step yields $\mathcal{D}_{seed}$ consisting of 649 positives, 150 negatives, 87 neutrals, and 660 irrelevants and a validation set $\mathcal{D}_{eval}$ of 250 instances set aside for all performance evaluation.
2. Next, we conduct minority class certain sampling (Sindhwani et al., 2009) of additional 900 instances to address class imbalance (details are described in § A.2). This step yields $\mathcal{D}_{certainty}$ consisting of 680 positives, 682 negatives, 227 neutrals, and 764 irrelevants.
3. Finally, we conduct margin sampling (Scheffer et al., 2001) (described in § A.2) that yields $\mathcal{D}_{margin}$ consisting of 824 positives, 947 negatives, 314 neutrals, and 1,020 irrelevants.

3.3.1 Annotating $\mathcal{D}_{hope}$

We use Prolific for our annotation study, and hosted the survey on Qualtrics. Our study was reviewed

---

[2] Appendix contains examples (Table A18) of LLM explanations for a true positive and a false positive.

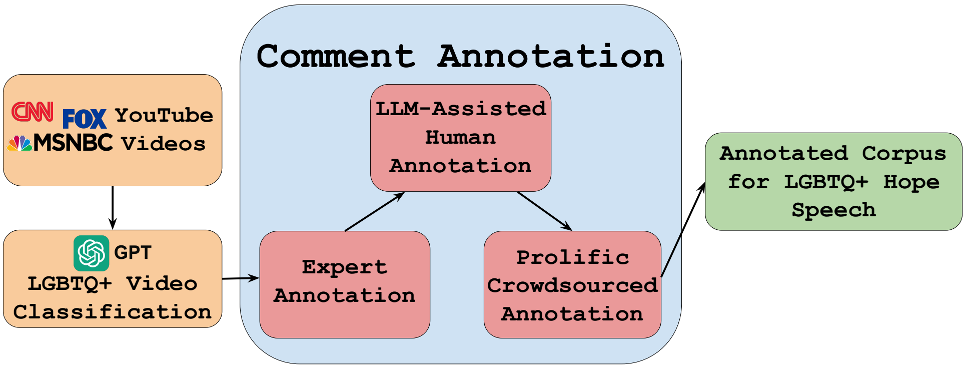


Figure 1: Corpus annotation pipeline

Dataset	Size	Sampling Strategy	Positive	Negative	Neutral	Irrelevant
$\mathcal{D}_{seed}$	1,950	Initial crowd-sourced dataset	649 (33.3%)	150 (7.7%)	87 (4.5%)	660 (33.8%)
$\mathcal{D}_{certainty}$	2,850	Certainty sampling	680 (23.9%)	682 (23.9%)	227 (8.0%)	764 (26.8%)
$\mathcal{D}_{margin}$	3,750	Margin sampling (Scheffer et al., 2001), a well-known variant of uncertainty sampling	824 (22.0%)	947 (25.3%)	314 (8.4%)	1,020 (27.2%)
$\mathcal{D}_{eval}$	250	Random set of comments taken from $\mathcal{D}_{seed}$ held out from training for evaluation	824 (22.0%)	947 (25.3%)	314 (8.4%)	1,020 (27.2%)

Table 2: Evolution of $\mathcal{D}_{hope}$ through active learning stages, showing dataset size, sampling methodology, and label distribution.

by our Institutional Review Board and was deemed as exempt. The annotation study follows a similar annotation design as Weerasooriya et al., 2023 and Crowl et al., 2025 in considering raters evenly distributed across three political leanings (Republican, Democrat, and Independent). We split our corpus into a set of 65 batches, each containing an equal amount of comments from each channel. Each batch is labeled by three annotators (one Republican, one Democrat, and one Independent). Since we do not want to expose raters to objectionable content for a sustained period, following standard practices, we limit our batch size to 30 instances. We also left an unstructured text feedback input for each of the workers. While many raters thanked us for providing an interesting and thought-provoking task, some responded to the survey with homophobia (Appendix contains examples in § I) which might lead to interesting follow-on research.

On our final annotated dataset, we achieve moderate inter-rater agreement. The Fleiss' κ for all four labels is 0.431 and for just two labels (*HopeSpeech* (positive) and *not-HopeSpeech* (neutral, negative, and irrelevant)) is 0.485. For a potentially subjective task like ours this agreement is in line with extensive prior literature. For instance, Guest et al., 2021 reported Fleiss' κ of 0.48 while Sanguinetti et al., 2018 reported category-wise κ = 0.37. We further note that our observed inter-rater agreement is higher than Gomez et al., 2019 (κ = 0.15), Fortuna et al., 2019 (κ = 0.17) and Weerasooriya et al., 2023 (κ = 0.39). Weerasooriya et al., 2023 studied political subjectivity in offensive content annotation and is closest to our study.

4 Classification Results

We consider `Mistral` (Mistral-7B-Instruct-v0.3) (Jiang et al., 2023) and `Llama 3` (Llama-3-8B-Instruct) (Dubey et al., 2024), two well-known open LLMs for fine-tuning. We determine the aggregate label by taking a majority vote, a standard approach to resolving annotator disagreement (see, e.g., Davidson et al., 2017; Wiegand et al., 2019). Any comments that received three separate labels are not used in the aggregate data, accounting for 11.3% of instances. In addition to fine-tuning models on aggregate labels, we also fine-tune models specific to each political affiliation. For those models, we use these 11.3% additional instances[3].

Table 3 summarizes our models' performance on $\mathcal{D}_{eval}$. Collectively, we refer to our annotated corpus as $\mathcal{D}_{hope}$, which evolved through three active learning phases as seen in Table 2: initial seed collection ($\mathcal{D}_{seed}$) comprising 1,950 instances, certainty sampling ($\mathcal{D}_{certainty}$) which expanded the dataset to 2,850 instances, and margin sampling ($\mathcal{D}_{margin}$) which yielded our final dataset of 3,750 instances. $\mathcal{D}_{eval}$ consists of 250 instances randomly selected from $\mathcal{D}_{seed}$ and held out from all training processes, serving as a consistent evaluation set for

---

[3] Table 8 describes the noise audit results of these models.

	`Llama` (Dubey et al., 2024)				`Mistral` (Jiang et al., 2023)			
Metric	$\mathcal{M}(.)$	$\mathcal{M}(\mathcal{D}_{seed})$	$\mathcal{M}(\mathcal{D}_{certainty})$	$\mathcal{M}(\mathcal{D}_{margin})$	$\mathcal{M}(.)$	$\mathcal{M}(\mathcal{D}_{seed})$	$\mathcal{M}(\mathcal{D}_{certainty})$	$\mathcal{M}(\mathcal{D}_{margin})$
Accuracy	0.5520	0.8480	0.8400	0.8480	0.6640	0.8120	**0.8560**	0.8480
Macro-averaged metrics:								
Precision	0.5468	**0.8025**	0.7204	0.7993	0.5654	0.6297	0.7947	0.7960
Recall	0.5547	0.6366	0.6866	**0.6947**	0.5488	0.6001	0.6600	0.6843
F1 Score	0.4768	0.6680	0.6995	**0.7162**	0.5375	0.6017	0.6880	0.7087
Class-specific F1 scores:								
Positive	0.6845	0.8703	0.8727	0.8661	0.8312	0.8448	**0.8851**	0.8722
Negative	0.3009	0.5806	0.5789	**0.6842**	0.4545	0.5714	0.5882	0.6486
Neutral	0.2553	0.2759	**0.4103**	0.3871	0.1765	0.0690	0.3333	0.3871
Irrelevant	0.6667	**0.9453**	0.9360	0.9275	0.6879	0.9216	**0.9453**	0.9268

Table 3: Comparison of `Llama` and `Mistral` model results on aggregate labels over five independent training runs on $\mathcal{D}_{eval}$. $\mathcal{M}(.)$ denotes the zero-shot performance. $\mathcal{M}(\mathcal{D})$ denotes model performance when fine-tuned on $\mathcal{D}$. Table 2 summarizes dataset statistics.

all models. We consider both zero-shot and fine-tuned models (Appendix contains prompt details in § D). For fine-tuning, we use a sequence classification head. We split the training data, after removing the validation set, into an 80/20 test/train split and train for 5 epochs, considering the best model as the one with the highest Macro-F1. We find (1) the fine-tuned models considerably outperform zero-shot models; (2) the models' performance improve gradually across different stages of the active learning pipeline; and (3) the best-performing model is `Llama 3` trained on $\mathcal{D}_{margin}$.

For the sake of completeness, we compare performance against several off-the-shelf classifiers on adjacent tasks such as hate speech detection, sentiment classification, and stance detection (results in Table A1). As sentiment classifier baselines, we consider a `DistilBERT` model (Sanh et al., 2019) fine-tuned on SST-2 (Socher et al., 2013), which provides only positive/negative predictions. For fair comparison, we map positive sentiment to *hope speech* and negative sentiment to *negative* content, while removing any comments with *neutral* or *irrelevant* content. It achieves a Macro F1 of 0.601 and Positive F1 of 0.520, significantly lower than our best `Llama`-based model (Macro F1: 0.864, Positive F1: 0.685), even with the removal of neutral/irrelevant label noise. This gap stems from sentiment classifiers failing to capture intent in LGBTQ+ discourse. For instance, a sentiment classifier will predict *I will be extremely happy if Congress passes a resolution and takes away voting rights from the LGBTQ+ people* as positive.

We further compare our approach with toxicity and hate speech detectors. We use `Detoxify` (Hanu and Unitary team, 2020) for toxic comment detection and cardiffnlp's `hate-roBERTa` (Barbieri et al., 2020) for hate speech detection. For fair comparison, we map toxic and hateful content to negative speech and hope, neutral, or irrelevant content to non–toxic and non-hateful speech. Therefore, we report their Negative F1 scores (i.e., their ability to correctly identify harmful content) rather than Positive F1. These models achieved low Negative F1 scores (0.129 and 0.258 respectively), as they focus on overt toxicity, while in LGBTQ+ discourse disapproval can often manifest without explicit hate speech.

Finally, our task bears many similarities to that of determining stance on the LGBTQ+ community. However, off-the-shelf stance detectors are typically developed for specific topics (e.g., abortion or climate change), and there are few systems available that allow for input of a custom topic beyond using LLMs to specify in the prompt. In practice, these LLM-based approaches for stance detection end up resembling the zero-shot evaluations we have already reported.

Error analysis. Consistent with the existing literature (Joshi et al., 2017; Farabi et al., 2024), we observe that our models often struggled in detecting sarcasm. Appendix (§ N) discusses challenging examples.

5 Substantive Findings

5.1 User Engagement Findings

$\mathcal{V}_{LGBTQ+}$ represents a set of 3,161 news videos relevant to LGBTQ+ issues. We first construct a control set, $\mathcal{V}_{control}$. Each video v in $\mathcal{V}_{LGBTQ+}$ is matched with a video that is (1) not relevant to LGBTQ+ issues; (2) is uploaded by the same news outlet; (3) has roughly same video duration as v; and (4) has the publishing time closest to v.

Let for a given video v, v_{likes}, $v_{dislikes}$, v_{views}, and $v_{comments}$ denote the total number of likes, dislikes, views and comments on v, respectively. Following López-Navarrete et al., 2021; Le et al., 2022;

Braun et al., 2017; KhudaBukhsh et al., 2021, we calculate engagement rate and dislike ratio for a given video as follows.

$$engagementRate(v) = \frac{v_{likes} + v_{dislikes} + v_{comments}}{v_{views}}$$

$$dislikeRate(v) = \frac{v_{dislikes}}{v_{dislikes} + v_{likes}}$$

Outlet	**Type**	**Dislike Rate**		**Engagement Rate**	
		Mean	**Med.**	**Mean**	**Med.**
MSNBC	$\mathcal{V}_{control}$	22.7%	13.5%	2.69%	2.47%
	$\mathcal{V}_{LGBTQ+}$	24.1%	18.8%	2.60%	2.30%
FOX	$\mathcal{V}_{control}$	18.7%	10.6%	3.40%	3.29%
	$\mathcal{V}_{LGBTQ+}$	29.7%	20.8%	2.63%	2.52%
CNN	$\mathcal{V}_{control}$	27.3%	21.2%	1.69%	1.31%
	$\mathcal{V}_{LGBTQ+}$	35.5%	32.4%	1.66%	1.30%

Table 4: Dislike and Engagement Rates by outlet and video type.

Table 4 compares the engagement rate and dislike rates of videos in $\mathcal{V}_{LGBTQ+}$ and $\mathcal{V}_{control}$. YouTube stopped allowing users to access dislikes starting in 2021, so our dislike rate data only includes videos up until that point[4]. We observe that LGBTQ+ videos consistently received higher dislike rates across all three channels compared to non-LGBTQ+ content. This difference is most pronounced for FOX News and CNN, with smaller but still noticeable differences for MSNBC. Overall engagement rates for LGBTQ+ videos are generally lower than for non-LGBTQ+ content, though the magnitude of this difference varied by news outlet. FOX News showed the largest decrease in engagement for LGBTQ+ content, while CNN and MSNBC exhibited differences small enough to likely be insignificant. The higher dislike ratios and lower engagement rate for LGBTQ+ content across all news outlets, regardless of perceived political leaning, point to a persistent societal resistance to LGBTQ+ topics. Although, this resistance is shown to be more pronounced for the most conservative channel, FOX News. Appendix K and L contain additional contrastive analyses with other hot-button issues and a deeper look into heavily disliked videos.

5.2 Annotation Study Findings

Table 5 suggests partisan differences in the annotation of *hope speech* for the LGBTQ+ community. Even given the same comment text and annotation guidelines, Republicans and Democrats or Independents may have different conclusions as to what is hope speech or negative towards the LGBTQ+ community. The agreement between Democrats and Independents is higher than the agreement between either group and Republicans.

Our survey also includes a question[5] asking the raters if they considered themselves a member of the LGBTQ+ community. Out of 375 raters, 98 (31 Democrats; 29 Republicans; and 38 Independents) identified themselves as being LGBTQ+. The Cohen's κ agreement between LGBTQ+ and non-LGBTQ+ annotators is 0.449. This agreement fits between the low values found when comparing Republicans to others, and the high value found comparing Independents and Democrats and is close to the overall Fleiss' κ of 0.431. We find this result particularly interesting as it indicates that even for a topic highly relevant to gender and sexual orientation, rater disagreement is perhaps impacted more by political leanings than sexual orientation. Even if we collapse the labels into *HopeSpeech* and *not-HopeSpeech*, the qualitative findings on political disagreement remain unchanged (see, Appendix).

To better understand intra-group annotation agreement, we collect additional annotations for a subset of 30 batches where each instance was labeled by two annotators of the same political affiliation. Independent annotators show the highest internal consistency (κ = 0.500), suggesting more unified perspectives within this group than the Democrats (κ = 0.446) and the Republicans (κ = 0.368). The Republicans show the lowest intra-group agreement (even lower than their inter-group agreement with Democrats or Independents) suggesting more varied perspectives within the group.

To analyze the intersection of LGBTQ+ identity and political beliefs, we next conduct two-proportion z-tests comparing label distributions between LGBTQ+ and non-LGBTQ+ annotators within each political affiliation. The analysis has two key major takeaways (details in Table A17). First, LGBTQ+ annotators across all political affiliations were consistently less likely to label content as irrelevant compared to their non-LGBTQ+ counterparts (Democrats: -4.1%, $p < 0.05$; Republicans: -6.0%, $p < 0.001$; Independents: -3.5%, $p < 0.05$), suggesting they more readily recognize content relevant to LGBTQ+ issues. Table A13 shows a comment that a LGBTQ+ annotator correctly identified as LGBTQ+-related, while the two other annotators seemed to have not enough context to make the connection. Second, the strongest differences appear among Republican annotators, where LGBTQ+ Republicans were significantly more likely to label content as negative and less likely to label it as

---

[4] sourced from Dutta et al., 2024b

[5] Raters could opt out of answering this question.

	Dem	*Rep*	*Ind*
Dem	-	0.408	0.477
Rep	0.408	-	0.396
Ind	0.477	0.396	-

Table 5: Human rater agreement across political affiliations. A cell $\langle i, j \rangle$ presents the Cohen's κ agreement between raters with political affiliation i and j. *Dem*, *Rep*, and *Ind* denote Democrat, Republican, and Independent, respectively.

positive compared to non-LGBTQ+ Republicans. These findings suggest that lived experience as an LGBTQ+ person may affect content interpretation, with the strongest differences emerging for ideologically charged content among conservative annotators.

In addition to our main model fine-tuned on the aggregate labels, we also fine-tune models specifically on the labels of each political affiliation (denoted by $\mathcal{M}_{rep}$, $\mathcal{M}_{dem}$, and $\mathcal{M}_{ind}$). $\mathcal{M}_{rep}$ was trained only on the 3,750 comments labeled by Republicans, $\mathcal{M}_{dem}$ only on the 3,750 labeled by Democrats, and $\mathcal{M}_{ind}$ on the 3,750 labeled by Independents. Using each of these models, we classify a set of 50k unseen comments from each channel. Following (Kahneman et al., 2021; Weerasooriya et al., 2023), we conduct a *noise audit* of these models as shown in Table 8. A *noise audit* involves examining agreement between different model variants on identical inputs, revealing systematic differences in model predictions. To verify whether political biases are introduced during fine-tuning, we first examine agreement between pre-trained and fine-tuned models. We observe notably low agreement scores: Cohen's κ of 0.29 for both $\mathcal{M}_{ind}$ and $\mathcal{M}_{dem}$ with pre-trained `Llama`, and 0.25 for $\mathcal{M}_{rep}$ with pre-trained `Llama`. These low agreement scores suggest substantial shifts in model behavior during fine-tuning. While we observe the in-the-wild agreement across models is higher than the agreement across human raters with different political beliefs, the impact of the political affiliations still shone through. Once again, $\mathcal{M}_{dem}$ and $\mathcal{M}_{ind}$ share higher agreement with each other, while they both have lower agreement with $\mathcal{M}_{rep}$ to a similar degree. This demonstrates that the human biases due to political affiliation in labeled data can perpetuate to fine-tuned models. Tables 6 and 7 present illustrative examples showing how trans-exclusionary stance of human Republican raters may get perpetuated to $\mathcal{M}_{rep}$ affecting marginalized voices.

Comment	*Rep*	*Dem & Ind*
Good Job America - Keep it UP- GO WOKE GO BROKE - Transheuser-Busch deserves IT	Positive	Negative
Thank you so much for sharing. Your story highlights how messed up the treatment is for Trans people.	Negative	Positive

Table 6: Disagreements in labels across human raters on Prolific. *Rep*, *Dem*, and *Ind* denote Republican, Democrat, and Independent, respectively.

Comment	$\mathcal{M}_{rep}$	$\mathcal{M}_{dem}$, $\mathcal{M}_{ind}$
Since when was having a phobia such a bad thing? You wouldnt say this about some scared of heights of or spiders... there freaks amd theres nothing to be ashamed of when having a transphobia	Positive	Negative
damn! first time I've seen Tucker actually lose. I love you Tucker, but the military transgender ban is BS. it's not a huge expense. Like Sean Patrick said in the end before Tucker rage quits "They're earning those benefits by putting their life on the line for America and have done more for this country than you (Tucker) and I ever will"...	Negative	Positive

Table 7: Illustrative examples where $\mathcal{M}_{rep}$ does not agree with $\mathcal{M}_{dem}$ and $\mathcal{M}_{ind}$.

We also calculate the performance metrics of zero-shot classification against the annotations of each political affiliation (see, Table 9). The results show similar performance with Democrat and Independent labels, but each metrics was lower for the Republican labels. Similar trend was present across the three different models we tested; `GPT-4o-mini`, `Llama`, and `Mistral`. This suggests that the biases present in these models may align closer with the biases found in our human Independent and Democrat raters, than those of a Republican.

In addition, we conduct a parallel analysis training separate models on annotations from LGBTQ+ and non-LGBTQ+ annotators. For fair comparison, in this experiment, we use a reduced dataset where comments had labels from both LGBTQ+ and non-LGBTQ+ annotators. Consistent with our findings from human annotators, we observe that the non-LGBTQ+ model more frequently classified content as "irrelevant" compared to the LGBTQ+ model. Table A13 shows an example where the LGBTQ+ model identified as relevant content that non-LGBTQ+ annotators marked as irrelevant.

Key takeaways: (1) Annotation of *hope speech* is associated with political leanings that drive more

	$\mathcal{M}_{dem}$	$\mathcal{M}_{rep}$	$\mathcal{M}_{ind}$
$\mathcal{M}_{dem}$	-	0.572	0.670
$\mathcal{M}_{rep}$	0.572	-	0.573
$\mathcal{M}_{ind}$	0.670	0.573	-

Table 8: In-the-wild Cohen's κ agreement of models trained on data annotated by raters with same political affiliation.

disagreement than gender and sexual identity on a task relevant to gender and sexual identity; (2) LGBTQ+ annotators across all political affiliations were consistently less likely to label content as irrelevant and showed stronger recognition of LGBTQ+-related content, with the most pronounced differences appearing among Republican annotators; (3) political biases in human annotations carried over to fine-tuned language models, with models trained on Democrats and Independents showing higher agreement compared to the Republican trained model.

5.3 In-the-wild *hope speech* findings

In order to look at the hope and hate speech makeup of LGBTQ+ discussion in-the-wild, we consider our best-performing model and classify 50k randomly sampled, unseen comments from each news outlet. The results, as shown in Figure 2 with numerical results in Table A19, reveal distinct patterns. In particular, FOX News has the highest proportion of negative comments (24.15%) and the lowest proportion of *hope speech* (2.3%) among the three outlets. This aligns with the outlet's reputation for conservative viewpoints. In contrast, CNN shows a more balanced distribution, with the highest percentage of *hope speech* (7.66%) and a negative comment rate (17.2%) closer to the overall average. MSNBC, however, has a much lower proportion of *hope speech* than CNN (4.55%) and the lowest proportion of negative comments (9.36%).

Figure 2 reports the ratio of *hope speech* and *hope speech* and *negative*s found in the wild. Dubbed *positivity ratio*, this ratio focuses on comments expressing polarized stance towards the LGBTQ+ community, excluding *irrelevant* and *neutral* comments. We find that MSNBC has the highest *positivity ratio* at 32.7%, followed by CNN at 30.81%, both above the overall average of 22.25%. In stark contrast, FOX News shows a considerably lower *positivity ratio* of 8.7%. These findings further emphasize the divergence in audience sentiment across channels, with MSNBC and CNN fostering more balanced or slightly supportive discussions around LGBTQ+ content, while FOX News comments lean heavily towards negative sentiment. To better understand the nature of these discussions beyond labels, Appendix M contains a lexical analysis of the language used in different comment categories.

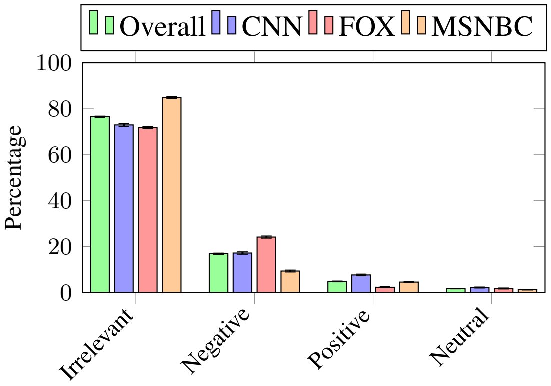


Figure 2: Label breakdown by news outlets with 95% confidence intervals. 50k comments from each news outlet found in-the-wild were classified by our best performing model. Numerical values are listed in Table A19.

Perhaps the most important results come from looking at the trends as a whole. For every channel, there were more negative comments than *hope speech*. This consistent imbalance, suggests a broader societal tendency towards critical or oppositional engagement with LGBTQ+ topics in online spaces. It implies that, regardless of the audience, viewers are more inclined to express disapproval or criticism than support or affirmation when commenting on LGBTQ+-related content. [6] Another noteworthy trend is the uniformly low percentage of neutral comments across all channels, ranging from 1.22% to 2.22%. This scarcity of neutral perspectives, coupled with the positive-negative imbalance, points to a highly polarized discourse surrounding LGBTQ+ issues. It suggests that those who engage in these comment sections tend to hold and express strong opinions, whether positive or negative, rather than maintaining a neutral stance.

To confirm the accuracy of our model when generalizing to in-the-wild data, we conducted a manual inspection of 100 comments labelled as Positive. From those 100, we found 16 false positives, mainly consisting of comments that we would label as Negative according to our guidelines, such as:

LGBTQHDJFNEIEKBEJDJDNDIDBDJ-

---

[6] Our Limitations and Appendix contain an exploratory study into Reddit showing that there are certain social web pockets which may be more supportive than negative, warranting additional investigation into more varied social media.

Metric	GPT-4o-mini			Meta-Llama-3-8B			Mistral-7B-v0.3		
	Dem.	**Rep.**	**Ind.**	**Dem.**	**Rep.**	**Ind.**	**Dem.**	**Rep.**	**Ind.**
Accuracy	**0.5963**	0.5520	0.5864	0.5143	0.4809	0.5091	0.5354	0.4954	0.5260
Macro-averaged metrics:									
Precision	**0.6146**	0.5664	0.6057	0.4534	0.4101	0.4484	0.4316	0.4019	0.4236
Recall	**0.5839**	0.5391	0.5717	0.3803	0.3497	0.3736	0.4178	0.3849	0.4061
F1 Score	**0.5749**	0.5320	0.5636	0.3716	0.3401	0.3653	0.4080	0.3790	0.3990
Class-specific F1 scores:									
Positive	**0.7236**	0.6464	0.7113	0.5801	0.5102	0.5609	0.6586	0.5884	0.6469
Negative	**0.6509**	0.6171	0.6330	0.5846	0.5670	0.5781	0.6103	0.5571	0.5886
Neutral	**0.3406**	0.3233	0.3264	0.2096	0.1718	0.1943	0.2604	0.2537	0.2453
Irrelevant	**0.5843**	0.5412	0.5840	0.4840	0.4513	0.4930	0.5108	0.4957	0.5141

Table 9: Zero-shot model results across political affiliation specific labels evaluated on the full human annotated corpus of 3,750 comments.

> IDJDIGWJOQPQPLAIDHWNRBHXKS-OFORBRIDKWNDIKFKDNRUWODUE
> Get it together bigot!

This example shows a comment that is Negative towards the LGBTQ+ community in a more subtle or sarcastic way. An error analysis in Appendix N further discusses some limitations that we identified through challenging comments it labeled incorrectly.

6 Conclusions

This paper presents a comprehensive picture of LGBTQ+ discussions in mainstream US political discourse. Overseen by a health expert specializing in LGBTQ+ health for decades, we conduct a detailed annotation study that reveals political biases of raters are associated with how they rate supportive content for the LGBTQ+ community. Our study shows that such biases may perpetuate into models affecting marginalized voices. Our in-the-wild assessment of LGBTQ+ discussions reveal that negative comments about the community considerably outweigh positive discourse indicating a technological gap to ensure safer spaces for marginalized communities.

7 Ethics Statement

Our study was reviewed by our Institutional Review Board and was deemed as exempt. Our study is overseen by a health expert with decades of research on LGBTQ+ health. We investigate publicly available data collected using public APIs.

We do not collect any PII (Personally Identifiable Information) about the raters and compensate them above the minimum wage. Since content moderation can be potentially gruesome and affect the mental health of the raters (Solon, 2017), we maintain a small batch size (30 YouTube comments).

While our goal is to broaden our understanding of LGBTQ+ discussions in mainstream US political news and our content classifier can assist human moderators to identify supportive content for the LGBTQ+ community, any content filtering system can be tweaked for malicious purposes. For instance, an inverse filter can be made that filters out *hopeSpeech* posts while filtering in *not-hopeSpeech* ones.

Our substantive findings rely on fine-tuned LLM. Studies indicate that these models have a wide range of biases that reflect the texts on which they were originally trained, and which may percolate to downstream tasks (Bender et al., 2021).

8 Limitations

While our study offers valuable insights into hope speech detection for LGBTQ+ topics in US political discourse, we recognize there are limitations that shape the scope and applicability of the findings.

Our focus on YouTube comments for major news channels, while providing a rich dataset, is only a fraction of the online discource related to LGBTQ+. Platforms like Twitter, Reddit, and others have their own unique demographics, content, and user interactions. What we observed on YouTube may not be true for the rest of the Internet, potentially limiting our generalizability across social media. We conducted an exploratory investigation on Reddit, sampling 15,000 comments on LGBTQ+-relevant posts (2016-2024) from r/politics, r/republican, and r/democrat, for a total of 45,000 comments. Overall, the positivity ratio remains below 0.5 (0.40), meaning that negative comments outnumber hope speech on Reddit, as on YouTube. However the subreddit-level breakdown is more nuanced, the r/politics and r/democrat subreddits show a slight tilt toward hope speech, whereas r/republican is strongly negative. This differs from YouTube, where even communities one might expect to have more positivity, such as the CNN comment section, still had negative comments outnumbering hope speech. These results, seen in Table A20, underline the fact that future work across additional communities and media is necessary.

Additionally, our study is grounded in US politics, where societal attitudes towards the LGBTQ+ have been affected by historical and legal context. However, LGBTQ+ rights and acceptance may vary dramatically across the world. In some countries, open support for LGBTQ+ rights might be more common and less contentious, while in others, it could be far more risky or even illegal to express such support. This global variation in LGBTQ+ rights and societal attitudes means that the patterns of hope speech and the very definition of what constitutes supportive language could differ significantly across cultural and national boundaries. Recent research (Samir et al., 2024) demonstrates these cross-cultural differences concretely, finding significant variations in how LGBTQ+ individuals are portrayed across different language Wikipedias, with some versions systematically emphasizing negative biographical information or omitting positive achievements based on local cultural attitudes.

While our study did not specifically aim at participatory AI (Harrington et al., 2019; Birhane et al., 2022) as our key goal was to study the interplay of rater politics and how they perceive LGBTQ+ discussions, a considerable fraction of our raters self-reported as being part of the community. Future studies can solely focus on participatory AI involving more raters from the LGBTQ+ community and extend this research to vicarious interactions (Weerasooriya et al., 2023).

We also acknowledge that many challenges faced by individual groups within the LGBTQ+ community are unique. For instance, the trans community faces several additional challenges such as participation in competitive sports or access to gender reassignment treatment. Future studies solely focusing on the trans community will add further value to prior literature focusing on specific exclusionary behavior (see, e.g., Lu and Jurgens, 2022).

In terms of political affiliations, while we attempted to capture the broad ideological difference present in the US, this may oversimplify their nuanced reality. Our categorization of political affiliations into the groups of "Democrat," "Republican," and "Independent" may not fully capture these nuances. In reality, political views exist on a spectrum rather than in discrete categories.

While we have attempted to make the definition of hope speech clear, operationalizable, and indicate how it differs from positive sentiment, we acknowledge room for further nuance. Appendix E lists a nuanced definition of *hope speech* which we considered in our initial annotation phase conducted by the author. However, our initial pilot revealed that this fine-grained definition resulted in poor quality annotation and low annotator agreement. Keeping rater instructions simple to reduce cognitive load on raters is a recommended practice in the crowd-sourcing literature (Finnerty et al., 2013). To reduce cognitive load, we adopted a simpler four-way scheme, *Positive*, *Negative*, *Neutral*, *Irrelevant*, and worded our annotation guidelines as advised by the subject-matter expert using the more familiar term *Positive* in place of *Hope Speech*. Although this worked well for the purposes of our study, a finer-grained taxonomy and clearer operational guidelines merit deeper investigation.

Finally, we acknowledge a lack of intersectionality analysis in our study. We have focused on LGBTQ+ identity and political affiliation, overlooking the possibility of intersection with factors such as race, ethnicity, age, or socioeconomic status. The interplay of these identities can have significant influence on how an individual perceives the LGBTQ+ community and speech surrounding it. Future work could be improved by capturing these complex intersection and identifying possible impact on LGBTQ+ and hope speech perception.

9 Acknowledgements

We acknowledge Research Computing at the Rochester Institute of Technology for providing computational resources. We thank Pranita Menavlikar and Pavan Turlapati for conducting preliminary investigations that informed this work, and Logan Crowl for assisting with the implementation of the crowdsourced experiments, along with the human annotators who took part in the study. We also thank Dr. Naveen Sharma and Dr. Christian Newman for their support and feedback as part of Jonathan Pofcher's thesis committee. KhudaBukhsh were partly supported by a gift from Lenovo.

References

Roberto L Abreu and Maureen C Kenny. 2018. Cyberbullying and lgbtq youth: A systematic literature review and recommendations for prevention and intervention. *Journal of Child & Adolescent Trauma*, 11:81–97.

Saharsh Agarwal, Uttara M Ananthakrishnan, and Catherine E Tucker. 2022. Deplatforming and the control of misinformation: Evidence from parler. *Available at SSRN*.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms*, pages 184–190.

Josh Attenberg, Prem Melville, and Foster Provost. 2010. A unified approach to active dual supervision for labeling features and examples. In *Proceedings of the 2010th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECMLPKDD'10, page 40–55, Berlin, Heidelberg. Springer-Verlag.

Fan Bai, Alan Ritter, and Wei Xu. 2021. Pre-train or annotate? domain adaptation with a constrained budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5002–5015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159, Bangkok, Thailand. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *ACM FaccT*, pages 610–623.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. *EAAMO 2022*, pages 1–8.

Courtney Blackwell, Jeremy Birnholtz, and Charles Abbott. 2015. Seeing and being seen: Co-situation and impression formation using grindr, a location-aware gay dating app. *New media & society*, 17(7):1117–1136.

Peter Braun, Alfredo Cuzzocrea, Lam MV Doan, Suyoung Kim, Carson K Leung, Jose Francisco A Matundan, and Rashpal Robby Singh. 2017. Enhanced prediction of user-preferred youtube videos based on cleaned viewing pattern history. *Procedia Computer Science*, 112:2230–2239.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José García-Díaz. 2022a. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 378–388, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Logan Crowl, Sujan Dutta, Ashiqur R KhudaBukhsh, Edson Severnini, and Daniel S Nagin. 2025. Measuring criticism of the police in the local news media using large language models. *Proceedings of the National Academy of Sciences*, 122(9):e2418821122.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Arka Dutta, Adel Khorramrouz, Sujan Dutta, and Ashiqur R. KhudaBukhsh. 2024a. Down the toxicity rabbit hole: A framework to bias audit large language models with key emphasis on racism, antisemitism, and misogyny. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7242–7250. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Arka Dutta, Aman Priyanshu, and Ashiqur R. KhudaBukhsh. 2025a. All You Need Is S P A C E: When Jailbreaking Meets Bias Audit and Reveals What Lies Beneath the Guardrails (Student Abstract). In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 29353–29355. AAAI Press.

Sujan Dutta, Deepak Pandita, Tharindu Cyril Weerasooriya, Marcos Zampieri, Christopher M. Homan, and Ashiqur R. KhudaBukhsh. 2025b. ARTICLE: annotator reliability through in-context learning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 14230–14237. AAAI Press.

Sujan Dutta, Mallikarjuna Tupakula, Sumeet Kumar, and Ashiqur R KhudaBukhsh. 2024b. Anonymous dissent in the digital age: A youtube dislikes dataset. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 89–99. Springer.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, pages 8020–8028. ijcai.org.

Federal Bureau of Investigation. 2024. Fbi crime data explorer.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Ailbhe Finnerty et al. 2013. Keep it simple: Reward and task design in crowdsourcing. In *SIG CHItaly*, pages 1–4.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Paula Fortuna, João Rocha, Juan Soler Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled portuguese hate speech dataset. pages 94–104.

Daniel García-Baena, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, and Rafael Valencia-García. 2023. Hope speech detection in spanish: The lgbt case. *Language Resources and Evaluation*, pages 1–28.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring hate speech detection in multimodal publications. *Preprint*, arXiv:1910.03814.

Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An investigation of large language models for real-world hate speech detection. *Preprint*, arXiv:2401.03346.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 5792–5809. Association for Computational Linguistics.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, Maria Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118:e2116310118.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *CSCW*, 3:1–25.

Andrew Hartman. 2019. *A war for the soul of America: A history of the culture wars*. University of Chicago Press.

Amey Hengle, Aswini Padhi, Sahajpreet Singh, Anil Bandhakavi, Md. Shad Akhtar, and Tanmoy Chakraborty. 2024. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with RLAIF. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024*, pages 6716–6733. Association for Computational Linguistics.

Christopher Homan, Gregory Serapio-Garcia, Lora Aroyo, Mark Díaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. 2024. Intersectionality in ai safety: Using multilevel models to understand diverse perceptions of safety in conversational ai. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 131–141.

Gregory A Huber and Neil Malhotra. 2017. Political homophily in social relationships: Evidence from online dating behavior. *The Journal of Politics*, 79(1):269–283.

Shanto Iyengar and Sean J Westwood. 2015. Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. 2021. *Noise: A flaw in human judgment*. Little, Brown.

Adel Khorramrouz, Sujan Dutta, and Ashiqur R. KhudaBukhsh. 2023. For women, life, freedom: A participatory ai-based social web analysis of a watershed moment in iran's gender struggles. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6013–6021. ijcai.org.

Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom M. Mitchell. 2021. We don't speak the same language: Interpreting polarization through machine translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 14893–14901. AAAI Press.

Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom M. Mitchell. 2022. Fringe news networks: Dynamics of US news viewership following the 2020 presidential election. In *WebSci '22: 14th ACM Web Science Conference 2022*, pages 269–278. ACM.

Emma Kirby, Ash Watson, Brendan Churchill, Brady Robards, and Lucas LaRochelle. 2021. Queering the map: Stories of love, loss and (be) longing within a digital cartographic archive. *Media, Culture & Society*, 43(6):1043–1060.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90.

Truong Le, Minh-Vuong Nguyen-Thi, Huy Le, Quoc-Thang Vo, Tung Le, and Huy Tien Nguyen. 2022. Entube: A dataset for youtube video engagement analytics. *Research Square*.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alberto-Jesús López-Navarrete, Marga Cabrera-Méndez, and Rebeca Díez-Somavilla. 2021. Formula for measuring the engagement of the viewer on youtube : exploratory research on the main spanish youtubers.

Christina Lu and David Jurgens. 2022. The subtle language of exclusion: Identifying the toxic speech of trans-exclusionary radical feminists. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 79–91, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Lauren B McInroy, Rebecca J McCloskey, Shelley L Craig, and Andrew D Eaton. 2019. Lgbtq+ youths' community engagement and resource seeking online versus offline. *Journal of Technology in Human Services*, 37(4):315–333.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3.

Lucas Ramon Mendos and Dhia Rezki Rohaizad. 2024. Laws on Us: A global overview of legal progress and backtracking on sexual orientation, gender identity, gender expression, and sex characteristics. Report, ILGA World, Geneva. Contributing authors: Kellyn Botha and Alan Cueli and Alejandro Hilarión Moncada and Hanlu Hu and Curro Peña Díaz.

Shubh Mittal, Tisha Chawla, and Ashiqur R. KhudaBukhsh. 2024. You must be a trump supporter: Political identity projections on the social web. In *Social Networks Analysis and Mining - 16th International Conference, ASONAM 2024*, volume 15211 of *Lecture Notes in Computer Science*, pages 391–404. Springer.

Karsten Müller and Carlo Schwarz. 2020. Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Anh Ngo and Hanh Thi Hong Tran. 2023. Zootopi at hope2023@iberlef: Is zero-shot chatgpt the future of hope speech detection? In *IberLEF@SEPLN*.

Shriphani Palakodety, Ashiqur Khudabukhsh, and Jaime Carbonell. 2020a. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:454–462.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020b. Hope speech detection: A computational analysis of the voice of peace. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1881–1889. IOS Press.

Deepak Pandita, Tharindu Cyril Weerasooriya, Sujan Dutta, Sarah Luger, Tharindu Ranasinghe, Ashiqur R. KhudaBukhsh, Marcos Zampieri, and Christopher Homan. 2024. Rater cohesion and quality from a vicarious perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 5149–5162. Association for Computational Linguistics.

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating twitter data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265. Association for Computational Linguistics.

Keith T Poole and Howard Rosenthal. 1984. The polarization of american politics. *The journal of politics*, 46(4):1061–1079.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Díaz, Ding Wang, and Gregory Serapio-García. 2024. Grasp: A disagreement analysis framework to assess group associations in perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3473–3492.

Punyajoy Saha, Abhilash Datta, Abhik Jana, and Animesh Mukherjee. 2024. Crowdcounter: A benchmark type-specific multi-target counterspeech dataset. *CoRR*, abs/2410.01400.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5157–5163. ijcai.org.

Farhan Samir, Chan Young Park, Anjalie Field, Vered Shwartz, and Yulia Tsvetkov. 2024. Locating information gaps and narrative inconsistencies across languages: A case study of lgbt people portrayals on wikipedia. *Preprint*, arXiv:2410.04282.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer.

Vikas Sindhwani, Prem Melville, and Richard D Lawrence. 2009. Uncertainty sampling and transductive experimental design for active dual supervision. In *ICML*, pages 953–960.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Olivia Solon. 2017. Facebook is hiring moderators. But is the job too gruesome to handle? The Guardian.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Tharindu Cyril Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur R. KhudaBukhsh. 2023. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 11648–11668. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608.

Clay H. Yoo and Ashiqur R. KhudaBukhsh. 2023. Auditing and robustifying COVID-19 misinformation datasets via anticontent sampling. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*, pages 15260–15268. AAAI Press.

Clay H. Yoo, Shriphani Palakodety, Rupak Sarkar, and Ashiqur KhudaBukhsh. 2021. Empathy and hope: Resource transfer to model inter-country social media dynamics. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 125–134, Online. Association for Computational Linguistics.

A Curating $\mathcal{D}_{hope}$

A.1 Initial Human and LLM Annotation

Initially, one of the authors label a set of 1,500 comments (500 randomly sampled comments from three news outlets hosting $\mathcal{V}_{LGBTQ+}$ videos) as being hope speech directed towards the LGBTQ+ community or not. See Appendix A for the guidelines used in this process. This step only identified 80 positives (5.33%) i.e., LGBTQ+ hope speech suggesting that hope speech is a rare class. In order to curate a balanced dataset in a time- and cost-effective way, we turn to leveraging human-LLM collaborative annotation in the following way (Wang et al., 2024). Using the initial labelled set of 1,500 comments, we fine-tune `Llama2-13B` (training details are in the Appendix). While this fine-tuned classifier was not highly accurate, it allowed us to identify hope speech more efficiently. We use this model to classify sets of 20,000 unseen comments from each channel. The same author as before went through the positives and verified any true positives until we obtain a total of 975 verified LBGTQ+ hope speech comments; 325 from each channel. These comments combined with a set of 975 random negatives labeled by the model, balanced for each channel, made up the 1,950 comment corpus that would be labelled by our crowd-sourced annotators as our seed set.

We devote a separate subsection to describe our crowd-sourced annotation study design. In what follows, we outline our active learning steps to expand our seed set.

A.2 Active Learning

We observe that our initial fine-tuned models performed considerably worse on the minority classes of *Negative* and *Neutral*, while they had relatively high performance on the *Positive* and *Irrelevant* classes. As a remedial measure, we employ two different active learning sampling strategies.

The first was minority certainty sampling. Prior literature shows that this method can be useful for reducing class imbalance (Palakodety et al., 2020a; Attenberg et al., 2010). We consider our best-performing model at this point (a fine-tuned `Llama`) and have it classify an unseen set of 50k randomly sampled comments from each channel. We took the 150 comments of the highest certainty with labels *Negative* and *Neutral*, for each channel, totaling 900 instances. We used the same crowd-sourced annotation design described in Section 3.3.1 to have these labelled. This allowed us to help close the gap in imbalance for these two labels. Our final dataset at this step was 2,850 instances. Looking at the distribution of comments with consensus among the annotators, we had 680 Positives, 682 Negatives, 227 Neutrals, and 764 Irrelevants. We found that this improved our model's performance, especially with the Neutral class, and were interested in seeing if continue sampling could improve it more.

Next we tried margin sampling. Using the now highest performing model, Llama trained on the certainty sampled data, we labelled 50,000 more comments from each channel. We looked at the margin of difference between the probability of the chosen label and the next highest one. We wanted to collect samples where this margin was lowest for each combination of labels, giving us a balanced look at where the model was uncertain. In total, we collected 25 comments for the 12 combinations of labels for each channel, adding up to 900 samples. Once again, we used the same crowd-sourced annotation design described in Section 3.3.1 to have these labelled. This was our final dataset of 3,750 instances, which consisted of 824 Positive, 947 Negative, 314 Neutral, and 1020 Irrelevant consensus labels.

B Comparison to Off-The-Shelf Classifiers

Table A1 compares the performance of our *hope speech* classifier against traditional baselines for sentiment classification, hate speech detection, and stance detection.

Classifier	**Macro F1**	**Pos. F1**	**Neg. F1**
Traditional Sentiment (DistilBERT-sst2)	0.578	0.402	–
Traditional Sentiment (Positive/Negative only)	0.601	0.52	0.682
Toxicity-based (Detoxify-toxic-bert)	0.503	–	0.129
Hate Speech (cardiffnlp hate-roBERTa)	0.605	–	0.258
GPT-4o (Classification Prompt)	0.666	0.862	0.590
Our best model (Fine-tuned Llama)	0.716	0.866	0.684

Table A1: Performance comparison for off-the-shelf classifiers on relevant tasks. For classifiers optimized to detect supportive content, the Positive F1 is reported on a binary formulation mapping hope speech (Positive) versus non-hope speech (Negative); for toxicity and hate speech detectors the Negative F1 is shown on a binary formulation mapping hate speech (Negative) versus non-hate speech (Positive, Neutral, Irrelevant). Blank cells indicate metrics that do not apply.

C Additional Related Work

C.1 Additional Literature on Counter Speech

There exists a large body of research done into the automatic detection of hate-speech online. This is mainly relevant in the case of social media platforms such as Facebook or YouTube, where such hateful content can be very harmful to both user's online and in-person experiences, especially those belonging to marginalized or minority groups(Müller and Schwarz, 2020). By identifying this hate-speech, either before or after it is posted, it can be censored or removed and the user's that posted such content can be warned or banned. In extreme cases, entire social media platforms can be taken down by those hosting them(Agarwal et al., 2022). More recent research has shown that censorship and suppression of offenders is not always the best way to deal with such situations. These approaches can often be unsuccessful in limiting the spread of hateful ideas and may even backfire, both in the context of social media and in society as a whole (Benesch, 2014).

Counter-speech has been proposed as a method of mitigating the effects of and preventing hate-speech, without the need to limit free speech through censorship. Counter-speech is a somewhat vague term and current literature defines it in a multitude of ways and acknowledges that there are many different types and strategies of counter-speech that can be found online, each achieving varying levels of success in different communities. It was found that empathy-based approaches could cause posters of xenophobic hate-speech to delete their comments and be less likely to post similar content in the near feature (Hangartner et al., 2021). However, it has also been shown that strategies such as utilizing a 'Positive Tone' or affiliating oneself with the marginalized group are often not effective and may even garner replies by other users stating that such comments will not change the opinions of the hate speakers (Mathew et al., 2019). Another negative aspect of relying on counter-speech to moderate online communities is that it requires the existence of hate-speech to be employed. By most working definitions of counter-speech, it must be in reply to a hateful comment. This limits its effectiveness as it can only be used as a response and not proactively.

C.2 Additional Literature on Hope Speech

Addressing the drawbacks of counter-speech and the complexities of censorship, the emerging field of hope speech research presents a promising approach. It focuses on promoting positive, inclusive dialogues, aiming to transform and de-escalate online communities into environments of support and mutual respect. The term hope speech started to be used a few years ago by authors looking for user posted content that aimed to diffuse conflict, for example, in the context of the tensions between Indian and Pakistan (Palakodety et al., 2020b). Before this, similar work had been done, but used other terms such as "help-speech" to refer to the positive discourse (Palakodety et al., 2020a).

Since then, there have been multiple shared tasks in hope speech detection with over 50 teams presenting their findings (Chakravarthi et al., 2022a) (Chakravarthi and Muralidaran, 2021). These shared tasks utilized datasets specifically created with hope-speech detection in mind, such as HopeEDI: a multi-lingual collection of almost 60,000 YouTube comments manually labeled as containing hope-speech or not (Chakravarthi, 2020). This dataset makes a point to contain hope speech directed at LGBTQ+ communities, among many other topics. Our study is different as it focuses directly on the LGBTQ+ community, but also views it through the lens of the US political divide.

One related study that didn't explicitly focus on hope speech but provided valuable advancement to the field is the study of dehumanization. Dehumanization involves perceiving or treating people as less than human and often leads to extreme intergroup bias and hate speech. Mendelsohn, Tsvetkov, and Jurafsky developed a computational linguistic framework for analyzing dehumanizing language and applied it to discussions of LGBTQ people in the New York Times from 1986 to 2015 (Mendelsohn et al., 2020). They found increasingly humanizing descriptions of LGBTQ people over time. Notably, different words with similar meanings, such as "gay" and "homosexual," had different levels of dehumanization, with "homosexual" being more associated with dehumanizing attitudes.

While much of the previous literature on hope speech detection mentions the LGBTQ+, as the task gravitates towards marginalized groups, there are gaps in research focusing directly on it. García-Baena et al. created the dataset Spanish-HopeEDI. It contains 1,650 annotated Spanish LGBTQ+-related tweets. A multitude of traditional machine-learning algorithms were tested on identifying hope-speech on them, leading to very impressive results (García-Baena et al., 2023). One related shared task similar to the previous ones was conducted focusing on this topic, however it was for hate-speech detection instead. For this, a dataset that specifically labelled LGBTQ+-related comments as homophobic or transphobic was used (Chakravarthi et al., 2022b).

Another important question for hope speech detection is understanding how subjective interpretations of language can impact the effectiveness of content moderation. A related study investigates the disagreement between human and machine moderators on what constitutes offensive speech in political discourse (Weerasooriya et al., 2023). This research highlights the challenges of subjectivity in moderation, revealing that both human and machine moderators often disagree on what is considered offensive, especially when political leanings are involved. These findings are relevant for hope speech detection as they underscore the need for nuanced understanding and handling of subjective content, ensuring that genuine hope speech is accurately identified amidst varying interpretations. In the same way that there can be disagreement on what is offensive, there can be disagreement on what is supportive or hopeful.

The recent introduction and consistent improvement of LLMs in the field of natural language processing has opened many new doors, and the task of hope speech detection is no different. However, at this point in time there has not been a lot of research into how effective LLMs are at hope-speech detection, never mind LGBTQ+ hope speech superficially. At the time of writing, only one paper has looked into this and attempted zero-shot ChatGPT prompting to label a Hope Speech dataset (Ngo and Tran, 2023). They found that it resulted in poor performance for English texts, but was very successful for Spanish text. More work has been done surrounding LLM performance into hate-speech detection. The first extensive study into this found that prompting strategies had a significant impact on results (Guo et al., 2024). They found using a Chain-of-Thought Reasoning Prompt had the highest results in all metrics and overall was effective in detecting hate-speech.

D Prompts Used

```
Based on the following video title and
description, can you predict if this video
is likely to be specifically relevant to the
LGBTQ+ community? Answer in a single word
`yes' or `no'. Then print out an empty line
after which give your explanation.

Video title: {title}

Video description: {description}
```

Figure 3: Prompt for LGBTQ+ Video Classification

```
Based on the following comment, categorize
it regarding its stance towards the LGBTQ+
community. Choose the most appropriate
category:

A. Positive
B. Negative
C. Neutral
D. Irrelevant

Comment: {text}

Respond only with the letter of your choice
(A, B, C, or D).

Answer:
```

Figure 4: Prompt used for classification

E Annotation Guidelines

A comment is marked as **Hope Speech** if it meets any of the following criteria:

1. Advocates for LGBTQ+ well-being, rights, or acceptance. *E.g.,* "How can we support LGBTQ+ youth in America?"

2. Urges support for LGBTQ+ rights or anti-discrimination efforts. *E.g.,* "Politicians need to be against the Don't Say Gay Bill."

3. Pushes for LGBTQ+ equal rights (marriage, anti-discrimination, gender recognition). *E.g.,* "Trans rights are human rights, and it's time for our laws to reflect that."

4. Denounces anti-LGBTQ+ violence, discrimination, or hate speech. *E.g.,* "Homophobes need to be stopped."

5. Shows sympathy for LGBTQ+ struggles and solidarity. *E.g.,* "I'm straight, but love is love!"

6. Indirectly supportive and inclusive of LGBTQ+ community. *E.g.,* "Everyone deserves to love who they love, without fear or judgment."

A comment is marked as **Non-Hope Speech** if it meets any of the following criteria:

1. Expresses violent intent or supports discriminatory practices against LGBTQ+. *E.g.,* "If my kid has a gay teacher, they better watch out."
2. Calls for actions/policies harmful to LGBTQ+ rights or well-being. *E.g.,* "Marriage should only be between a man and a woman."
3. Diverts from LGBTQ+ issues to unrelated topics, diminishing their importance. *E.g.,* "I'm all for same-sex marriage, but should we worry about world hunger before this stuff?"
4. Spreads misinformation or stereotypes about LGBTQ+ community. *E.g.,* "The alphabet mob is brainwashing our kids!"
5. Demonstrates sarcastic or mocking support for LGBTQ+. *E.g.,* "Everyone should be able to identify as anything they want. I identify as an attack helicopter!"
6. Unrelated to LGBTQ+ issues. *E.g.,* "Abortion is a sin!"

F Two-Label Agreement

	Dem	*Rep*	*Ind*
Dem	-	0.447	0.556
Rep	0.447	-	0.432
Ind	0.556	0.432	-

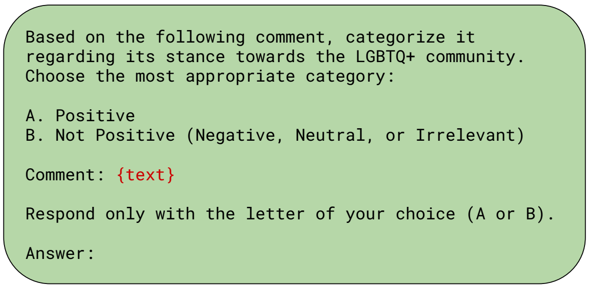
Table A2: Two-Label Political Affiliation Agreement (Human Annotators) Values are Fleiss' κ

	Non-LGBT
LGBT	0.498

Table A3: Two-Label LGBTQ+ Agreement (Human Annotators) Values are Fleiss' κ

G Two-Label Classification Prompt

The two-label classification prompt is presented in Figure 5.



```
Based on the following comment, categorize it
regarding its stance towards the LGBTQ+ community.
Choose the most appropriate category:

A. Positive
B. Not Positive (Negative, Neutral, or Irrelevant)

Comment: {text}

Respond only with the letter of your choice (A or B).

Answer:
```

Figure 5: Prompt for 2-Label Classification

Examples of Negative Survey Feedback
"You can identify as anything you want However how you were born cannot be changed or altered. Surgery and clothing does not change what you were born as. Time for common sense to reign again. 1,000 years from now when somebody digs up bones you will never hear them say oh look we just found a tr[redacted]nny!!!!"
"Comments speaking truth about the alphabet mafia are not negative just because the alphabet mafia does not like being confronted with truth."

Table A4: Examples of survey feedback expressing homophobic or negative views.

H Crowd-sourced Study Compensation

We compensate raters \$0.10 for each instance. This yields a compensation of \$3 for each 30 comment batch. We observe a median task completion time of 11:53 minutes, implying an hourly wage estimate of \$15.16, more than the minimum wage. Rater compensation is grounded in prior literature (Leonardelli et al., 2021; Bai et al., 2021; Weerasooriya et al., 2023). No participants complained about the compensation in their feedback.

I Survey Feedback

Table A4 lists a few examples of homophobic survey feedback. Table A5 lists a few examples of survey feedback expressing a positive stance towards our research task or the LGBTQ+ community.

J Annotator Demographics

All annotators were living in the United States. Complete demographics and a breakdown of LGBTQ+ demographics by political affiliation can be found in Table A6 and A7

Examples of Positive Survey Feedback
"Reading so much hateful, ignorant rhetoric was tough! Thanks for all you're doing to spread peace and love in the world."
"This study was very eye opening. Thank you for letting me participate in this wonderful study. Have a wonderful day."

Table A5: Examples of survey feedback expressing support for the research and LGBTQ+ community.

Category	Count	Percentage
Political Affiliation		
Independent	125	33.33%
Democrat	125	33.33%
Republican	125	33.33%
LGBTQ+ Community Identity		
Yes	98	26.13%
No	277	73.87%
Age Range		
18-24	45	12.00%
25-34	103	27.47%
35-44	104	27.73%
45-54	74	19.73%
55-64	34	9.07%
65 or older	15	4.00%
Self Description		
Male	153	40.80%
Female	209	55.73%
Nonbinary/third gender	10	2.67%
Self-describe	1	0.27%
Prefer not to say	2	0.53%

Table A6: Demographic Breakdown of Survey Respondents (Total Responses: 375)

K Comparison Across Controversial Topics

In an attempt to expand upon the findings in Table 4, we analyze dislike ratios across multiple hot-button issues in Table A8. While many of the topics elicit more polarized engagement than control content, LGBTQ+ content shows distinct patterns: it ties for the most polarizing content on MSNBC (24.1% mean dislike ratio), is considerably more polarizing on Fox News (29.7%, significantly exceeding traditional controversial topics like gun control at 22.8% and climate change at 21.7%), and maintains the third-highest negative engagement on CNN (35.5%) after immigration (40.8%) and gun-related content (37.9%). These findings demonstrate that LGBTQ+ topics generate uniquely high negative engagement across all platforms, regardless of the channel's political leaning. Although this is more pronounced on FOX, and less pronounced on CNN, which receives high dislike rates compared to the other channels across all topics.

Political Affiliation	LGBTQ+ Count (%)	Not LGBTQ+ Count (%)
Independent	38 (30.4%)	87 (69.6%)
Democrat	31 (24.8%)	94 (75.2%)
Republican	29 (23.2%)	96 (76.8%)
Total	98 (26.1%)	277 (73.9%)

Table A7: Distribution of LGBTQ+ Identity Across Political Affiliations

L Analysis of Video Content by Positivity Ratio

In order to examine the tone and content of videos that elicit highly positive or negative responses towards the LGBTQ+ community, we find the videos with the highest and lowest positivity ratio for each channel in Table A9 Our analysis shows distinct patterns in how different channels cover LGBTQ+ topics. Fox News videos maintain an opposing stance regardless of positivity ratio, with only one positive video appearing to feature a balanced perspective. CNN and MSNBC's most positive content focuses on pop culture and celebrity figures (e.g., Elton John, Randy Rainbow, Maddow). Gender identity and transgender issues appear disproportionately in lowest-positivity content across channels (3/5 CNN, 5/5 FOX, 1/5 MSNBC videos). One interesting finding is that a very specific news story, coverage of Brittney Griner's release, consistently generates lower positivity ratios, appearing in 3 of the least positive videos.

M Lexical Analysis of Comments by Category

In order to examine the type of discussion happening in these comments, we performed a lexical analysis in different comment categories based on aggregate crowd-sourced labels. For each category comparison (e.g., positive vs. negative), we computed log odds ratios with Laplace smoothing to identify words that appear disproportionately in one category versus another. A positive log odds ratio indicates the word is more characteristic of the first category, while a negative ratio indicates association with the second category. The magnitude of the ratio reflects the strength of this association. This analysis reveals patterns in how language is

	MSNBC		FOX		CNN	
Topic	**Mean**	**Med.**	**Mean**	**Med.**	**Mean**	**Med.**
Control	21.5%	13.5%	19.4%	10.9%	27.1%	20.3%
LGBTQ+	24.1%	18.8%	29.7%	20.8%	35.5%	32.4%
Gun	24.1%	16.3%	22.8%	15.5%	37.9%	33.3%
Climate	17.0%	10.6%	21.7%	11.4%	34.6%	32.0%
Abortion	20.3%	12.4%	17.6%	8.0%	29.0%	23.4%
Immigration	19.5%	14.2%	16.9%	8.0%	40.8%	38.7%

Table A8: Dislike ratio comparison of YouTube videos across controversial topics by news outlet. For each topic, we show the mean and median (Med.) dislike ratios.

Channel	Category	Representative Videos
CNN	Most Positive	1. Biden's surprise brings Elton John to tears at White House 2. Pete Buttigieg reacts to Martha Alito's Pride flag comment 3. Supreme Court says federal law protects LGBTQ workers from discrimination 4. Church member defends Pastor Worley's Anti-Gay Rant 5. GOP Rep. attends gay son's wedding after opposing protections
	Least Positive	1. Chicago police slam decision in Jussie Smollett case 2. Zoey Tur on Transgender Rights 3. Watch the moment Brittney Griner lands on US soil 4. FILE: JOAN RIVERS JOKES MRS OBAMA "IS A TRANS" 5. See confrontation between Ben Carson and lawmaker over transgender rights
FOX	Most Positive	1. Santorum answers gay soldier's DADT question 2. Franklin Graham: Christians should prepare for persecution after gay marriage ruling 3. Huckabee: 'This is where we've gone to la la land' 4. Trey Gowdy, Kayleigh McEnany rip liberals for attacking Caitlyn Jenner 5. 'OUTRAGE': Biden honors 'Trans Day of Visibility' on Easter Sunday
	Least Positive	1. Democratic Socialist convention erupts over pronouns 2. Gutfeld: How did this nutcase get a security clearance? 3. 'Turned into a circus': Democrats storm out of gender reassignment hearing 4. Target insider sounds alarm: 'Terrified' of Bud Light-style backlash 5. Gutfeld: Biden official accused of second luggage heist
MSNBC	Most Positive	1. Internet Sensation Randy Rainbow On Florida's 'Don't Say Gay' Bill 2. Maddow: We Feared Susan's Covid Would Kill Her 3. Federal Judge Rules Anti-HIV Medicine Is Unconstitutional 4. Drag queens describe a Pride Month like no other 5. See GOP confronted over 'straights only' discrimination
	Least Positive	1. Jen Psaki Leaves White House 2. Jotaka Eaddy: Criticizing Brittney Griner's Return Is 'Un-American' 3. John Heilemann: What's the body count on Critical Race Theory? 4. Brittney Griner Freed As Part Of Russian Prisoner Swap 5. For Facts Sake: Florida's Parental Rights In Education Bill

Table A9: Most and least positive LGBTQ+ videos by channel, ranked by user engagement positivity ratio. Videos selected from those with at least 100 comments and at least 10 positive/negative comments.

utilized across supportive, critical, neutral, and off-topic comments.

M.1 Positive vs. Negative Comments

Table A10 presents the distinctive words for positive and negative comments. On the positive side, one of the most clear patterns is the presence of terms such as *constitution*, *freedoms*, *citizen*, *fundamental*, and *defending*. These words suggest that supportive comments often reference the protection of human rights and civil liberties in the defense of the LGBTQ+.

In contrast, the negative comments are most clearly distinguished by words like *holy*, *sodom*, *gomorrah*, *leviticus*, and *repent*. This pattern suggests that opposition to LGBTQ+ rights often stems from religious frameworks and traditional moral values, sometimes even directly citing the Bible. The stark difference in these foundational frameworks—legal rights versus religious doctrine—highlights a fundamental tension in how different groups conceptualize LGBTQ+ issues.

One additional divergence in vocabulary between these two group may be rather obvious, but can be seen in the tone of words used. Positive comments frequently employ terms of inclusivity and understanding such as *love*, *wonderful*, *cared*, *understanding*. Negative comments are marked by language of moral condemnation such as *perversion*, *insanity*, *immorality*, and *unfair*. This shows how these categories are more than just "stance on LGBTQ+" and how one side is dehumanizing the

LGBTQ+ community, while the other is humanizing it.

Finally, the political dimension of this discourse is evident in the distinctive use of political identifiers. The presence of *rightwing* in positive comments, contrasted with *democrat*) and *democrats* in negative ones, suggests that supporters of LGBTQ+ rights often frame their opposition in terms of rightwing ideology, while those expressing negative sentiments frequently attribute their grievances to Democratic politics.

M.2 Irrelevant vs. On-Topic Comments

In Table A11, we compare vocabulary between irrelevant comments and those that are on-topic. As with neutral comments, it is somewhat difficult to identify a pattern in the distinctively irrelevant words, as it encompasses such a large spectrum. Irrelevant comments often contain references to other politicized topics, with terms like *newsom*, *covid*, *blm*, and *migrants*. This pattern suggests that discussions of LGBTQ+ issues often become entangled with broader political debates and culture war topics.

On-topic comments, conversely, maintain focus on LGBTQ+-specific terminology and issues, with words almost all words being related such as *lgbt*, *sexuality*, and *transgender*. This clear lexical differentiation helps validate our classifications and highlights how political discourse can drift across topics.

Distinctive Words for Positive vs. Negative Comments	
Top 30 words (Positive)	**Top 30 words (Negative)**
constitution (2.254)	looks (–3.035)
species (2.053)	perversion (–2.709)
spread (2.053)	mankind (–2.542)
scary (2.053)	company (–2.447)
affect (2.053)	democrat (–2.342)
condition (2.053)	letting (–2.342)
fully (2.053)	holy (–2.342)
love (2.018)	cake (–2.342)
roman (1.935)	flesh (–2.284)
rightwing (1.935)	logic (–2.224)
persecute (1.935)	insanity (–2.224)
hypocrisy (1.935)	west (–2.224)
wonderful (1.935)	services (–2.224)
neighbors (1.935)	sodom (–2.224)
treat (1.871)	leviticus (–2.224)
experiences (1.802)	aint (–2.159)
cared (1.802)	masculine (–2.090)
akon (1.802)	professor (–2.090)
activity (1.802)	sacred (–2.090)
valid (1.802)	gomorrah (–2.090)
citizen (1.802)	muslim (–2.090)
persecuted (1.802)	space (–2.090)
freedoms (1.802)	unfair (–2.090)
understanding (1.802)	gain (–2.090)
existence (1.802)	lust (–2.090)
treated (1.765)	libraries (–2.090)
beings (1.727)	financial (–2.090)
defending (1.647)	repent (–2.090)
fundamental (1.647)	democrats (–2.090)
testicles (1.647)	immorality (–2.090)

Table A10: Distinctive words for Positive versus Negative labeled comments. Numbers in parentheses are the log odds ratios.

Distinctive Words for Irrelevant vs. On-Topic Comments	
Top 30 words (Irrelevant)	**Top 30 words (On-Topic)**
yahuah (4.513)	transgender (–3.014)
yahushua (3.889)	lgbt (–2.913)
yahuahs (3.579)	gays (–2.799)
xxxxxxxxxxxxxxxxxxxxxxxxx (2.973)	homosexual (–2.770)
survey (2.973)	gay (–2.726)
newsom (2.973)	lgbtq (–2.716)
cookies (2.790)	compete (–2.470)
hoping (2.790)	heterosexual (–2.345)
covid (2.790)	homophobia (–2.285)
misty (2.790)	sexuality (–2.237)
obedience (2.790)	lesbian (–1.997)
brittany (2.790)	bigoted (–1.864)
susan (2.790)	gender (–1.774)
commandment (2.790)	shooter (–1.764)
teenagers (2.790)	dysphoria (–1.764)
gavin (2.790)	surgery (–1.764)
rapists (2.790)	visibility (–1.764)
slaves (2.685)	trans (–1.745)
rico (2.567)	transgenders (–1.737)
disapproval (2.567)	sports (–1.709)
material (2.567)	flesh (–1.709)
biracial (2.567)	community (–1.681)
sources (2.567)	celebrate (–1.652)
thanksgiving (2.567)	students (–1.652)
gift (2.567)	basic (–1.652)
blm (2.567)	cis (–1.652)
60s (2.567)	consider (–1.652)
migrants (2.567)	womens (–1.640)
139 (2.567)	weird (–1.592)
mob (2.567)	transphobic (–1.592)

Table A11: Distinctive words for Irrelevant versus on-topic (Positive+Negative+Neutral) comments. Numbers in parentheses are the log odds ratios.

YouTube Comment	Model Prediction
Bestiality rights are human rights. Vote biden to continue social progress.	Positive
Power and respect to her and all other biologically female athletes. STAY LOUD	Positive

Table A12: Two example challenging comments our classifier mistakenly identified as Positive.

Comment	LGBT	Non-LGBT
Imagine how much faster HE would be in the water, if he didn't have his junk slowing him down!	Negative	Irrelevant

Table A13: Example human annotations showing differences between LGBT and non-LGBT annotators.

N Error Analysis

While our model performed well in identifying the rare class of LGBTQ+ hope speech in the wild, there were cases where the beliefs expressed in comments were too nuanced to be classified correctly. Table A12 showcases two such examples. Both comments are generally supportive on a surface level. One is promoting human rights and social progress, the other is giving respect to female athletes. However, in context it is clear the support is not directed to the LGBTQ+. The first comment is equating the social progress of the LGBTQ+ community with bestiality, a common homophobic cliché. The second comment is specifically singling out "biological female athletes" to support, taking a clear stance on the issue of transgender people in sports.

Examples like these illustrate the challenges a task like this entails. Our model fails to fully understand the sarcasm or superficial support that less straight-forward comments may contain. Enhancing the model's ability to understand context and implicit meanings, possibly through the use of more advanced language models or improved tuning merits deeper exploration.

O Licenses

Meta Llama 3 is used under the Meta Llama 3 Community License. Mistral is used under the Mistral AI Non-Production License.

Comment	$\mathcal{M}_{LGBTQ+}$	$\mathcal{M}_{non\text{-}LGBTQ+}$
Now why did God destroy Sodom and Gomorrah? Hmmmmm.	Negative	Irrelevant

Table A14: Example predictions from models trained on LGBTQ+ and non-LGBTQ+ annotated data (denoted by $\mathcal{M}_{LGBTQ+}$ and $\mathcal{M}_{non\text{-}LGBTQ+}$, respectively).

P Computational Resources

Models were trained using both 7B (Mistral) and 8B (Llama) parameter versions. Training was done on a university-ran high-performance computing cluster using a single A100 GPU. Experiments totalled approximately 120 GPU hours.

Q Packages

Package	Version
PyTorch	2.3.0-rc12
Transformers	4.35.2
Datasets	2.8.0
NumPy	1.26.3
Scikit-learn	1.4.0
PEFT	0.5.0

Table A15: Software packages and versions

R Training Setup and Hyperparameters

Hyperparameter	Value
Learning Rate	2e-4
Batch Size	8
Number of Epochs	5
Weight Decay	0.01
LoRA rank (r)	8
LoRA alpha	32
LoRA dropout	0.1

Table A16: Hyperparameters used during our fine-tuning

We used the PEFT library for parameter-efficient fine-tuning with LoRA. Model selection was done using the best macro F1 score on the evaluation dataset.

S LGBTQ+ Label Distribution Comparison

Table A17 contrasts the label distributions obtained from LGBTQ+ and non-LGBTQ+ raters across different political affiliations.

T LGBTQ+ Video Classification Pipeline Examples

Table A18 shows an example true positive and a false positive found in our results. The first shows a case in which the video ended up not being LGBTQ+ relevant, except for the fact that the person featured is a director of trans care, but the video does not focus on that fact. It is interesting that `GPT-4o` actually acknowledges this in its explanation, however did end up still flagging it as relevant. The second is an example that identifies the *bathroom bill* as relevant to LGBTQ+ issues.

U In-the-wild Label Distribution

V Subreddit Label Distribution

Political Affiliation	Label Type	LGBTQ+	Non-LGBTQ+	Difference	p-value
Democrat	Positive	25.3%	24.3%	+1.0%	0.543
	Negative	29.6%	29.1%	+0.6%	0.753
	Neutral	17.7%	15.2%	+2.5%	0.078
	Irrelevant	27.4%	31.5%	-4.1%	0.023*
Republican	Positive	21.7%	26.1%	-4.3%	0.012*
	Negative	36.8%	27.7%	+9.1%	<0.001***
	Neutral	16.5%	15.3%	+1.2%	0.405
	Irrelevant	24.9%	30.8%	-6.0%	0.001**
Independent	Positive	27.7%	24.7%	+3.0%	0.059
	Negative	30.1%	28.6%	+1.5%	0.370
	Neutral	14.1%	15.1%	-1.0%	0.438
	Irrelevant	28.1%	31.6%	-3.5%	0.038*

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table A17: Label distribution by political affiliation and LGBTQ+ identity. Values show the percentage of comments assigned each label type, with z-tests used to compute significance of differences.

Video title	Video description	`GPT-4o` explanation	Classification
Attacks On Abortion Rights Already Forcing Women To Dangerous 'Back Alley' Alternatives	Dr. Bhavik Kumar, medical director of primary and trans care at Planned Parenthood Gulf Coast, talks with Rachel Maddow about the desperation and hardship the Texas abortion ban is causing people seeking abortion services to which they have a right, and explains that the longer term restrictions on reproductive rights are already forcing people seeking abortions into dangerous circumstances.	The video features Dr. Bhavik Kumar, who is the medical director of primary and trans care at Planned Parenthood Gulf Coast. Although the main topic of the video is about the attacks on abortion rights in Texas, the mention of "trans care" in Dr. Kumar's professional role suggests that the topic might be relevant to transgender people, who are a part of the LGBTQ+ community. However, since the description doesn't explicitly state this, it's not a certainty, but a possibility.	False Positive
Montel Williams slams North Carolina's 'bathroom bill'	Activist talks problematic articles within the controversial legislation on 'The Kelly File'	The video seems to be relevant to the LGBTQ+ community based on the reference to North Carolina's 'bathroom bill,' a controversial legislation that affected the rights of transgender people, which makes it pertinent to their interests and concerns. Montel Williams being noted as an activist, alongside mentioning 'problematic articles' in the bill, suggests a discussion about these LGBTQ+ issues.	True Positive

Table A18: Videos identified as LGBTQ+ relevant by our `GPT-4` pipeline. The top example shows a case in which the video ended up not being LGBTQ+ relevant, except for the fact that the person featured is a director of trans care, but the video does not focus on that fact. It is interesting that `GPT-4o` actually acknowledges this in its explanation, however did end up still flagging it as relevant. The second is an example that identifies the *bathroom bill* as relevant to LGBTQ+ issues.

Label	Overall	CNN	FOX	MSNBC
Irrelevant	76.53% ± 0.025%	72.96% ± 0.06%	71.78% ± 0.045%	84.87% ± 0.045%
Negative	16.90% ± 0.025%	17.20% ± 0.05%	24.15% ± 0.045%	9.36% ± 0.04%
Positive	4.84% ± 0.015%	7.66% ± 0.035%	2.30% ± 0.02%	4.55% ± 0.02%
Neutral	1.73% ± 0.01%	2.19% ± 0.02%	1.77% ± 0.02%	1.22% ± 0.01%
Positivity Ratio	22.25% ± 0.065%	30.81% ± 0.12%	8.70% ± 0.07%	32.70% ± 0.15%

Table A19: Label breakdown and positivity rates by channel with 95% confidence intervals. 50k comments from each channel found in-the-wild were classified by our best performing model.

Label	Overall	r/democrat	r/republican	r/politics
Irrelevant	80.61%	85.87%	77.89%	78.06%
Negative	9.01%	5.04%	14.52%	7.48%
Positive	6.17%	5.88%	3.47%	9.15%
Neutral	4.21%	3.21%	4.11%	5.31%
Positivity Ratio	40.64%	53.83%	19.29%	55.01%

Table A20: Label breakdown by subreddit. 15k comments from each subreddit were classified by our best performing model