Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Prevalence, Substance and Responses to Hate Speech Against LGBTQ Communities on TikTok

Jordi Guillem Condom Tibau[1], **Angelina Voggenreiter**[2], **elena pavan**[1], **Jürgen Pfeffer**[2]
[1]University of Trento
[2]School of Social Sciences and Technology, Technical University of Munich
jordi.condomitibau@unitn.it, angelina.voggenreiter@tum.de, elena.pavan@unitn.it, juergen.pfeffer@tum.de

Abstract

Despite ongoing efforts, online hate speech remains a pervasive issue on social media, particularly affecting vulnerable groups such as LGBTQ communities. While there is extensive debate around how best to address this problem, counter speech is emerging as a promising solution. However, existing research has primarily focused on detecting hateful content, often overlooking broader aspects such as the specific topics of discrimination and the spread of countermeasures online. This study examines the prevalence of hate speech and counter speech in LGBTQ online spaces on TikTok, analysing day-to-day interactions to identify recurring themes and targets. Results reveal that hate speech is widespread: at least 3.5% of messages contain hateful content, spread by approximately 4% of users, and one in three videos attracts hate comments or replies, primarily targeting LGBTQ topics explicitly. Gender identity emerges as a major focus, with transgender and non-binary individuals being frequent targets. Although much hate engagement goes unanswered, when responses occur, they are often in the form of counter speech, especially when LGBTQ-related topics are targeted. These findings improve our understanding of the nature and extent of online hate speech against LGBTQ communities, confirm counter speech as an employed response, and provide a foundation for further research aimed at developing strategies to promote safer, more inclusive social media environments.

1 Introduction

In digital spaces, hate speech (HS) frequently targets minority groups based on appearance, religion, gender, or sexual orientation. Online anonymity, invisibility, and rapid reach amplify the prevalence and intensity of HS by fostering depersonalization and deindividuation (Ștefăniță and Buf 2021; Baider 2020). This anonymity is associated with higher levels of HS, particularly against racialized and gender minorities (Mondal, Silva, and Benevenuto 2017), allowing users to spread hateful messages with little accountability (Brown 2018). The detrimental effects of HS are well-documented, with short-term consequences such as mood swings, anger, and fear, and long-term impacts like anxiety, depression, stress, and alcoholism (Brown 2015; Hawdon, Oksanen, and Räsänen 2017). The prevalence of HS can desensitize victims and the public, reducing reporting and intervention efforts, while overexposure may further normalise HS within public discourse (Soral, Bilewicz, and Winiewski 2018; Harel, Jameson, and Maoz 2020).

LGBTQ[1] communities frequently face extreme discrimination and abuse both offline and online (Díaz-Torres et al. 2020). Although sexual orientation and gender identity are fundamental aspects of personal identity that must be respected and protected (Thurlow 2001), LGBTQ communities worldwide continue to face violence, inequity, and extreme forms of harassment (Barrientos et al. 2010). As a result, many turn to social media for information, disclosure, self-expression, and social connection (Han et al. 2019; Delmonaco and Haimson 2023; Haimson 2018; Duguay 2019), as well as for social support, identity exploration (Ellison et al. 2016; Semaan 2019), activism (Hekma 2016), and fostering intimate and sexual relationships (Ferris and Duguay 2020). The Internet thus serves as a safe harbour (Lucero 2017) where LGBTQ communities can overcome intersectional oppression.

However, social media can be a double-edged sword (Fisher, Tao, and Ford 2024), as vulnerable groups often face exclusion and harassment (Walker and DeVito 2020; Marciano and Antebi-Gruszka 2022). In particular, LGBTQ communities are more likely to experience online violence and HS (Abreu and Kenny 2018; Lingiardi et al. 2020), with transgender individuals being more likely to be targeted than members of other LGBTQ communities (Walters et al. 2020). This is also the case for activists, who often face a higher volume of abuse (Ștefăniță and Buf 2021). Yet, discrimination against LGBTQ communities is intrinsically sociotechnical. The algorithmic nature of social media platforms can automate and reinforce pre-existing biases and inequalities (Hoffmann 2019; Garcia 2016). Binary classification models rooted in cis-heteronormative assumptions frequently result in discrimination against LGBTQ users (Bivens and Haimson 2016; Blackwell et al. 2017). Additionally, transgender and non-binary individuals face disproportionate content removal, with posts mislabelled as

[1]In this paper, we use the acronym LGBTQ (lesbian, gay, bisexual, trans*, queer) to indicate every subjectivity whose identity, experiences and practices do not conform with the cis-heterosexual matrix (Butler 2011). Additionally, we use the term LGBTQ communities to refer to both individuals and groups belonging to the various communities composing the acronym.

"adult", "toxic", or "offensive" (Haimson et al. 2021; Dorn et al. 2024). Platforms like YouTube have been shown to flag and demonetise LGBTQ content under algorithms biased toward "family-friendly" standards (Wilkinson and Berry 2020). Such biases in content moderation not only fail to provide protection but also actively enable wrongful censorship, shadowbanning, and filtering of LGBTQ communities (Rauchberg 2022; Delmonaco et al. 2024; Duguay, Burgess, and Suzor 2020; Dias Oliva, Antonialli, and Gomes 2021).

Existing studies on HS against LGBTQ communities online often concentrate on specific events or contents, constructing and exploring datasets that are skewed towards HS instances or LGBTQ-related issues. For instance, researchers have examined homophobic stereotypes during the 2022 monkeypox outbreak (Carratalá 2023), Facebook news posts on LGBTQ topics (Silva and Silva 2021), Pride Month tweets collected using both neutral and derogatory LGBT-related keywords (Locatelli, Damo, and Nozza 2023), and debates preceding the implementation of the Trans Law in Spain (Fernández 2024). Although these studies are essential for understanding HS against LGBTQ communities, their reliance on event-specific or keyword-based datasets often fails to capture the online experiences and adversities that LGBTQ communities face every day.

Above and beyond addressing specifically online HS instances, recent research has also turned towards examining the various types of responses that users can put forward. Generally speaking, strategies to counter HS fall into four categories: disruption, inaction, education, and counter speech (CS) (Citron and Norton 2011). Each has its strengths and limitations. Inaction may reduce immediate conflict but risks signalling tolerance of HS, while education aims to build long-term awareness through media literacy and public campaigns (Baider 2023). Disruption, such as blocking or removing content, offers short-term mitigation but is less effective against covert HS and may raise free speech concerns or push content to less-regulated platforms (Mathew et al. 2019; Baider 2023). Another existing approach is flagging, which allows users to report offensive content but is often underutilised by victims (Hubbard 2020). Concerns about current HS mitigation strategies are growing, as highlighted by the European Commission's evaluation of the Code of Conduct on illegal online HS (Reynders 2022). Some argue that CS could be a key solution, as debate is often seen as preferable to censorship, even for extreme content (Bartlett and Krasodomski-Jones 2016), and potentially more effective long-term than disruption (Richards and Calvert 2000). CS is defined as a crowdsourced response through alternative and counter narratives (Baider 2023; Bartlett and Krasodomski-Jones 2016). Alternative narratives promote tolerance, while counter-narratives challenge prejudice and encourage critical thinking (Braddock and Horgan 2016; McGowan 2009). More specifically, counter narratives are defined as "communicative actions aimed at refuting HS through thoughtful and cogent reasons, and true and fact-bound arguments" (Schieb and Preuss 2016). CS respects free speech while addressing the attitudes behind HS and engaging witnesses to reduce its impact (Buerger 2021), potentially shifting societal views and fostering meaningful discourse (Benesch et al. 2016). Despite indications of its potential, its true impact remains uncertain due to limited presence of large-scale studies (Gagliardone et al. 2015).

Against this backdrop, this study tackles the dramatically long-lasting problem of online HS against LGBTQ communities by focusing on three key points that aim to advance the understanding and mitigation of this issue.

First, we shift the focus from exclusively examining HS instances to analysing both its prevalence – that is, how widespread HS may become amongst the interactions that unfold online around LGBTQ individuals, groups and contents – and its substance – that is, the contents it consists of. Indeed, while the exposure of LGBTQ communities to online hate is well-documented, how and how much these instances of HS taint their daily online experiences is something that remains insufficiently understood. Second, we link this focus on the prevalence and the substance of HS against LGBTQ communities to CS, an increasingly emphasised and promising countermeasure which is still underinvestigated, especially with respect to LGBTQ communities. Third, we conduct this analysis on TikTok, a platform widely popular especially among young audiences but relatively underexplored in academic research. Extant studies have analysed HS on TikTok, including right-wing populist posts (González-Aguilar, Segado-Boj, and Makhortykh 2023), hashtags like #StopAsianHate (Jacques et al. 2023), or video replies from users with disabilities responding to hate (García-Prieto, Bonilla-del Río, and Figuereo-Benítez 2024). However, research examining anti-LGBTQ content remains scant and mainly focuses on posts rather than interactions through comments and replies (O'Connor 2021).

Along these lines, we seek to contribute to amplify the reach of the current state of the art on online HS against LGBTQ communities by addressing the following research questions:

- **RQ1:** What is the prevalence of HS against LGBTQ communities on TikTok?
- **RQ2:** What are the specific themes around which LGBTQ related HS is concentrated?
- **RQ3:** What is the prevalence of CS in response to HS, and how does such prevalence vary across hate themes?

To address these questions, we collected a large dataset of comments and replies from posts by LGBTQ influencers and activists on TikTok, individuals who are actively engaged with the communities and are often seen as safe spaces for their members. We use this dataset to predict which messages qualify as HS and CS, and then apply topic modelling to analyse their content. For HS messages, we examine their replies to assess CS prevalence and nature. We then compare our findings with existing literature, both qualitative studies and quantitative analyses with other platforms and hate targets, to determine similarities, differences and whether prior conclusions still hold in a large-scale quantitative analysis of TikTok.

2 Related Work

There are many evolving definitions of HS. Commonly, HS includes any expression, whether verbal, non-verbal, symbolic, or communicative, that aims to demean individuals based on group membership (Simpson 2013). HS can manifest in explicit and implicit forms. Explicit HS is unambiguous and consists of expressions whose literal meanings are hateful (ElSherief et al. 2021). In contrast, implicit HS is more elusive, indirect, and coded, using subtle cues such as derogatory metaphors (Musolff 2015), sarcasm or humour (Askanius 2021), veiled inferences, conspiracy theories (Baider 2022), and dog-whistling (Bhat and Klein 2020). Despite legal frameworks like the 2019 UN Action Plan and the 2016 EU Code of Conduct aimed at regulating online communication, they often fall short in addressing implicit HS, and those intent on spreading hate have developed strategies to evade detection and accountability while remaining unpunished (Fortuna and Nunes 2018; Hietanen and Eddebo 2023). With the increase of online media and user-generated content, HS has become more pervasive online, making manual moderation impractical. As a result, much research has focused on HS detection using machine learning (ML) (Yu, Blanco, and Hong 2022; Sap et al. 2020; Abarna, Sheeba, and Pradeep Devaneyan 2023; Koch et al. 2024; Geet d'Sa, Illina, and Fohr 2020), lexicon-based approaches (Gitari et al. 2015), and multimodal techniques (Perifanos and Goutsos 2021; Lippe et al. 2020).

Several studies have estimated HS prevalence on online platforms, but the variability in methodologies and findings underscores the need for more rigorous, standardized research. For instance, around 0.64% to 3.2% of tweets during the COVID-19 crisis contained anti-Asian HS (He et al. 2021; Alshalan et al. 2020). A random sample from the Twitter streaming API showed less than 1% of tweets contained HS, with gender and sexual orientation among the primary targets (Mondal, Silva, and Benevenuto 2017). Research on TikTok is more limited, but around 1.6% of interactions on posts by politicians and celebrities involved harassment (Hinduja and Patchin 2023), and about one-third of comments on discussions about the 2023 Middle East conflict were xenophobic (González-Esteban et al. 2024).

Research on hate targeting LGBTQ communities is limited, often grouped under broader categories of HS or offensive language, which reduces focus on hate specifically directed at LGBTQ communities (Kumar, Singh, and Kumar 2021; Negi Advocate 2023). While some studies address homophobia, transphobia, or queerphobia (Chakravarthi et al. 2022; Chakravarthi 2024; Chakravarthi et al. 2024; Manukonda and Kodali 2024), other forms of LGBTQ hate remain underexplored. Some existing datasets include subsets related to LGBTQ hate (Mollas et al. 2022; Ljubešić, Fišer, and Erjavec 2019; Mathew et al. 2022; Sap et al. 2020; Chung et al. 2019), but often lack sufficient granularity within the community. Moreover, much research neglects the complexity of LGBTQ hate, especially in online spaces (Lingiardi et al. 2020; Sánchez-Sánchez, Ruiz-Muñoz, and Sánchez-Sánchez 2024). Few studies explore the specific themes driving LGBTQ HS or identify triggers for it on a large scale (Fernández 2024), and approaches like multi-class classification or topic modelling often lack a focus on the community (Yigezu et al. 2023; Mollas et al. 2022; Shekhar, Saini et al. 2021; Bano et al. 2023).

Studies on the spread of hate against LGBTQ communities online reveal high levels of HS, with 17% to 58% of responses to news about the 2022 monkeypox outbreak in Spanish newspapers on Twitter being homophobic, depending on the outlet (Carratalá 2023). On Brazilian cybermedia, about one in four comments on nine LGBT-related news posts were HS (Silva and Silva 2021). Additionally, an analysis of LGBT-related tweets across multiple languages found that 36% of the sampled English-language tweets expressed negative sentiment (Locatelli, Damo, and Nozza 2023). These studies highlight the variation in the prevalence of HS based on factors such as the event, target group, and methodology, making comparisons between platforms challenging. Despite differences, estimates suggest that online abuse accounts for between 0.001% and 1% of content on major platforms, although these figures are speculative due to platform opacity (Vidgen, Margetts, and Harris 2019). Prior research on the substance of HS against LGBTQ communities highlights recurring themes such as bathroom use, healthcare, and slurs (Casey et al. 2019). Other prominent focuses include intentional misgendering, deadnaming, accusations of mental illness, or claims that LGBTQ communities threaten traditional values or the family unit. Common online narratives further depict LGBTQ communities as predatory, in need of treatment, or a danger to children (Dhiman 2023). On Twitter, frequent topics also included biological determinism, and claims of paedophilia and mental health disorder (Fernández 2024). Although TikTok remains under-researched on this topic, its large, predominantly youth-based audience underscores the importance of studying HS targeting communities such as LGBTQ. To our knowledge, such research has yet to be conducted at scale, especially regarding the substance and prevalence of HS in everyday interactions.

While there is extensive literature on HS, research on CS remains comparatively underexplored, underscoring the urgent need to expand our understanding of it. Existing studies on CS detection primarily employ ML techniques (Mathew et al. 2018, 2019; Garland et al. 2020; He et al. 2021), with some considering conversational context as a predictive feature (Yu, Blanco, and Hong 2022). However, CS datasets, particularly those focused on LGBTQ online HS, are limited (Bonaldi et al. 2022; Fanton et al. 2021; Chung et al. 2019). Even though there is some research on CS detection, studies on the prevalence of CS actions online are less common (Garland et al. 2020). For example, estimates suggest that about 0.55% of all Twitter replies, from both HS and non-HS tweets, are CS (He et al. 2021), while others suggest that three out of four responses to HS are CS (Mathew et al. 2018).

3 Data

TikTok Data

This study leverages data from two sources: TikTok and existing literature. While the primary focus is TikTok data, its

lack of labels necessitated using existing literature to train HS and CS detection models. Although cross-platform analysis would be valuable, we focus on TikTok for several reasons. Despite claims of actively combating HS, both our findings and prior research indicate its widespread diffusion on the platform. For instance, despite TikTok's prohibition of specific content and behaviours such as claiming individuals have a mental illness if they identify as transgender, deadnaming or misgendering (TikTok 2024), such instances persist in our dataset. The European Commission's 7th evaluation of the Code of Conduct estimated TikTok's HS removal rate at only 60.2% during 2022, decreasing from the previous 80.1% in 2021 (Reynders 2022). Furthermore, TikTok's reports on HS and removal efforts mainly focus on user and post data, often overlooking comments and replies (TikTok 2023). While TikTok acknowledges the challenges of combating HS, it is evident that their current efforts fall short, relying heavily on users to report hateful content and utilise safety tools to filter what they see. Additionally, given TikTok's popularity among young people, protecting them from exposure to hate, especially that which targets their identity, is critically important and emphasises the urgent need for deeper research into the platform's dynamics.

We sourced the TikTok data through the TikTok data scraping API of EnsembleData (EnsembleData 2025). We collected data from 89 TikTok accounts related to LGBTQ communities using queries on TikTok and Google with the keywords LGBTQ/queer/transgender + TikTok + activists/creators/trailblazers/influencers, including the TikTok keyword only in the Google search. For each account, up to the last hundred videos (or fewer, if less were available) and up to a hundred comments for each video (or fewer, if less were available) with all their respective replies were requested and collected. Data collection occurred in two phases: April 29th, 2024 (videos and comments) and May 21st, 2024 (replies). The resulting dataset comprises 345,310 comments and 317,059 replies posted by 410,442 different users in 8,449 TikTok posts. The data ranges from February 2020 to May 2024, with 88% of the data being from 2023 and 2024. In what follows, we use "users" to refer to accounts that commented or replied, while "influencers" are the accounts from which the data was obtained. "Comments" are messages on TikTok posts, and " replies" are messages responding to comments. Only English comments and replies were included in the study.

In our study, we prioritised ethical responsibility, especially regarding privacy and data protection. User handles and video IDs were retained for quality control, such as verifying data accuracy, but we did not attempt to deanonymize users or access personal information beyond what was publicly available. While full anonymization is not feasible, and given the potential for misuse of the dataset, we chose not to make it publicly available to protect individuals. However, we will release the code and a sample dataset for methodological transparency, prioritising the safety of LGBTQ communities [2].

[2]https://github.com/JordiCondom/Prevalence-Substance-Responses-Hate-Speech-Against-LGBTQ-TikTok.git

Labelling Data

A major challenge in this project was that the large amount of collected data was completely unlabelled. To develop models for predicting HS and CS, we utilised pre-existing labelled datasets from the literature, focusing on LGBTQ hate when multiple targets were identified. Although no TikTok-specific dataset was available, we utilised data from various platforms, including Twitter, Facebook, Reddit, YouTube, and hate sites like Gab. This highlights the pressing need to develop datasets for under-researched platforms like TikTok, enabling researchers to work on better understanding them. Differences in content structure and form across platforms can hinder model generalisation from one platform to another, but we addressed this by using data from various social media sites along with a semi-supervised learning approach on the unlabelled TikTok data. As detailed in the Methods section, this combination is expected to enable our model to generalise effectively for predicting HS against LGBTQ communities in TikTok. When collecting the literature labelled data, we aimed to include as many LGBTQ communities as possible, resulting in 30,471 labelled texts: 19,119 non-HS and 11,352 HS. Table 1 outlines the literature datasets and number of instances used from each one, the labels given by the dataset authors to such instances and the online platform they were sourced from, if any. Although no dataset was a subset of another, the intersection was not null, so repeated texts were removed.

The same process was followed for CS, but adapting the dataset to the difficulties of predicting it. At first, attempting to predict whether a single sentence in isolation was CS proved insufficient. To improve prediction, we included conversational context by using both a comment and its corresponding reply as input to the model, a method demonstrated to be effective in previous studies (Yu, Blanco, and Hong 2022). A key challenge was ensuring the model could distinguish CS from non-CS replies to the same HS comment, avoiding the assumption that all responses to HS are CS. For that reason, we used the DIALOCONAN (Bonaldi et al. 2022) and Multitarget CONAN (Fanton et al. 2021) datasets, which contain dialogues labelled with HS and CS instances, including a subset related to LGBTQ. These dialogues have a length of two in Multitarget CONAN (labelled as HS-CS pairs), and a minimum length of four and maximum of eight in DIALOCONAN (alternating HS-CS, starting with HS). By reorganising these dialogues meaningfully and chronologically, we obtained 4,213 labelled pairs: 2,211 HS-CS, 1,001 HS-HS, and 1,001 CS-HS, with 617 pairs from Multitarget CONAN and 3,596 from DIALOCONAN. No suitable social media datasets containing CS within dialogues were identified, in contrast to the availability of datasets for HS, underscoring the critical need for further research in this area, particularly the development of comprehensive CS datasets from social media sites. To address a potential limited generalisability of the detection model to social media texts, we trained our model in a semi-supervised manner on the TikTok dataset, mirroring our approach with HS, which will be detailed in the Methods section.

We acknowledge the limitation that using these datasets for HS and CS detection inherently aligns the model def-

Dataset	Used instances	Label	Platform(s)
Kennedy et al. (2020)	27803	Gay, Lesbian, Bisexual, Non-binary, Transgender, Other Sexuality/Gender	Twitter, YouTube, Reddit
Banerjee and Nguyen (2023)	10000	Queerphobia	YouTube
Ljubešić, Fišer, and Erjavec (2019)	4337	LGBTQ	Facebook
Sap et al. (2020)	2770	Gay, Lesbian, Bisexual, Asexual, Non-binary, Transgender	Twitter, Reddit, Gab, Stormfront
Mathew et al. (2022)	1940	Homosexual, Asexual, Bisexual	Twitter, Gab
Röttger et al. (2021)	1014	Gay people, Transgender people	None
Chung et al. (2019)	465	LGBT+	None
Mollas et al. (2022)	94	Sexual Orientation	HateBusters, Reddit

Table 1: Number of instances used from each HS dataset, together with the label given by the dataset authors to such instances and the platform(s), if any, the instances were sourced from.

initions of HS and CS, and by extension this study, with those established by the respective researchers. While for HS we include expressions intended to demean individuals based on group membership, the datasets used for CS consist exclusively of counter narratives. All datasets used were licensed for free use for research purposes with citation. We acknowledge and comply with the terms of these licenses.

4 Methods

The workflow for semi-supervised training of HS and CS detection, followed by topic modelling, is outlined in Figure 1, with each step detailed in this section.

Hate Speech Prediction

This section details the training of the model that was employed to answer RQ1 by predicting whether comments and replies contained HS, and RQ2 by identifying the texts to which topic modelling should be applied. Since the TikTok dataset lacked HS labels, supervised training was infeasible. To overcome this, we used existing labelled datasets specific to HS against LGBTQ communities from various online platforms. To ensure the detection model generalised to TikTok data, we employed a semi-supervised learning approach since relying solely on such datasets did not ensure generalisation. We first trained a model on labelled data, then used it to predict HS probabilities for the TikTok dataset. The most confident predictions (top and bottom 5% or those exceeding a .95 probability threshold, whichever yielded a smaller dataset) were treated as pseudo-labels and incorporated with the original labelled data for retraining from scratch. A probability threshold was necessary because the top 5% of comments classified as HS reached low-confidence probabilities.

To handle the unique challenges of social media text like spelling mistakes, emojis, or very short messages, preprocessing involved removing usernames, links, punctuation, and emojis, as well as lowercasing the text and correcting spelling errors. While this approach could remove some contextual cues, such as deliberate typos or emojis used to evade detection, this study prioritised analysing HS textual content that could be used for the subsequent topic modelling, resulting in a conservative estimate, or lower bound, of the total HS present in the dataset.

For HS prediction, we fine-tuned the existing HateBERT model (Caselli et al. 2021), which was pretrained on hate comments from banned Reddit communities, and its tokenizer. Using both labelled LGBTQ HS data and pseudo-labelled TikTok data, we adapted the model to identify both general HS and LGBTQ specific hate on TikTok, such as misgendering and deadnaming, which might otherwise be overlooked. We verified that no parts of our labelled dataset were used in HateBert's original training. The model takes a text (TikTok comment/reply) as input and outputs a HS probability. We used the standard binary cross-entropy loss function for binary classification. The final model achieved an F1 score of .89, and a precision and recall of .89 on the test set with a 20-80 test-train split.

Counter Speech Prediction

This section describes the model used to address RQ3 by predicting whether replies to HS comments were CS. Since no pre-trained CS model was available, we fine-tuned a standard BERT model for sequence classification (Devlin et al. 2019) with sequences comprising comment-reply pairs and binary labels (CS or not), along with its tokenizer. The text pre-processing followed the same steps used for HS prediction. The model was trained as a binary classification task using cross-entropy loss. After semi-supervised training including comment-reply pairs from the TikTok dataset, the model achieved an F1 score of .92, precision of .93 and recall of .91 on the test set, with a 20-80 test-train split.

Topic Modelling

Given the HS predictions, we applied topic modelling to hate messages to address RQ2. Then, given the hate topics and the CS predictions, we observed how CS prevalence varied across themes and answered the second part of RQ3. To categorise hate messages by topic we used BERTopic (Grootendorst 2022). We optimized hyperparameters using three metrics: C_V coherence (Röder, Both, and Hinneburg 2015), number of topics, and outlier cluster size (Cluster -1 of BERTopic), aiming to maximise coherence while minimising outliers and balancing the number of clusters, with these last two being decided by the model. The

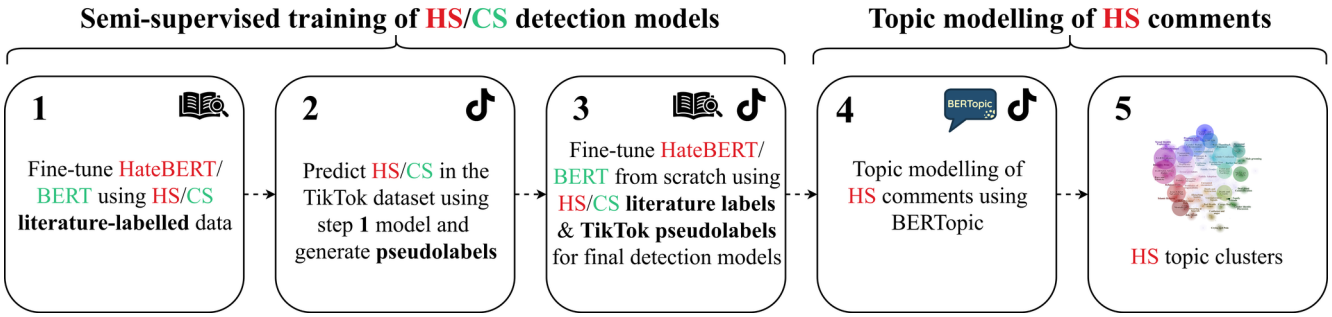

Figure 1: Workflow diagram illustrating the processes for the semi-supervised training of HS and CS detection models, along with the subsequent topic modelling.

text pre-processing for the inputs followed the same procedures as for HS and CS prediction. To enhance and diversify keyword-based topic representation, we used three distinct models: KeyBERTInspired, PartOfSpeech (focusing on nouns and adjectives), and a combination of KeyBERTInspired with Maximal Marginal Relevance. We also applied zero-shot labelling to the generated topic representations using a large language model (LLM), specifically OpenHermes-2.5-Mistral-7B (HuggingFace 2023). This approach offered an alternative interpretation of the topics, and a method to represent clusters in the embedding space, discussed in the next section. We acknowledge potential biases (Ntoutsi et al. 2020) and limitations (Ollion et al. 2023) introduced by this annotation and carefully monitored for any undesirable behaviour from the LLM. The prompt used was: *These are the top words of a topic resulting from topic modelling: {top_words}. Provide a concise label for this topic, limited to a maximum of 5 words:*. The {*top_words*} parameter is the set of words from the three topic representations.

Property	**Statistic**
Number of comments	345310
Number of (frac.) HS comments	4748 (1.4%)
Number of (frac.) HS comments with CS reply	1327 (28%)
Number of replies	317059
Number of (frac.) HS replies	18450 ($\sim$ 6%)
Number of replies to HS comments	5380
Number of (frac.) CS replies to HS comments	4053 (75.33%)
Number of users	410442
Number of (frac.) hate users	16719 (4.07%)
Number of videos	8449
Number of (frac.) videos with HS	2978 (35.25%)

Table 2: Summary of TikTok's dataset HS and CS prediction statistics.

5 Results

Hate and Counter Speech Prevalence

Table 2 summarises the results of HS and CS predictions. Addressing RQ1, 3.5% of comments and replies were classified as HS. However, there is a notable disparity between them: 1.4% of comments versus 6% of replies contained HS expressions. This suggests replies are four times more likely to contain HS, possibly due to the higher visibility or moderation of comments compared to replies. Additionally, 4% of users posted at least one HS message, and about one-third of videos included at least one HS instance.

Comparing HS prevalence across studies is challenging due to differing detection methods, dataset constructions, and prevalence metrics, as well as variations in reporting behaviours among hate targets, platform-specific moderation policies, and removal biases. These complexities emphasise the need for more consistent research approaches to support targeted communities online. However, comparing our results with studies from other platforms and hate targets remains crucial for contextualizing and understanding their broader implications.

We found more LGBTQ hate on TikTok than Anti-Asian hate on Twitter during the COVID-19 crisis, which ranged from 0.64% to 3.2% (He et al. 2021; Alshalan et al. 2020). Random samples from the Twitter streaming API reported less than 1% of HS tweets, including those targeting gender and sexual orientation (Mondal, Silva, and Benevenuto 2017), a lower rate than found here. Additionally, only 0.2% to 0.3% of Twitter users are estimated to spread HS (Vidgen, Margetts, and Harris 2019), significantly less than the proportion we identified. Platforms like Facebook, YouTube, and Reddit show even lower rates of abusive content: 0.001%, 0.001%, and 0.0001%, respectively (Vidgen, Margetts, and Harris 2019). The differences between our results and these rates are unsurprising given the well-documented abuse faced by LGBTQ communities, but the relative scale compared to such baseline remains striking.

On TikTok, studies have found that 19% of posts from right-wing populist parties contain HS (González-Aguilar, Segado-Boj, and Makhortykh 2023), and over 50% of videos using the #StopAsianHate hashtag included abuse targeting Asian individuals (Jacques et al. 2023). Although useful for comparison, these studies focus on specific events or hashtags, limiting their application to understanding the everyday dynamics of HS. More comparable data comes from a

study that found 1.6% of interactions on TikTok profiles of politicians and celebrities involved harassment (Hinduja and Patchin 2023). Notably, LGBTQ influencers and activists face more than double this rate, underscoring the disproportionate hostility endured by the community.

Regarding CS and in response to RQ3, 36% of HS comments received a reply, 75% of which were classified as CS. Out of all HS comments, 28% received at least one CS reply. In other words, while most HS comments go unanswered, those that receive a reply are often met with CS, highlighting significant support and counteraction. However, the community seems to selectively engage with certain themes of hate, leaving others unaddressed, at least with CS. Comparatively, a study that estimated 0.64% of anti-Asian hate on Twitter during COVID-19, also estimated 0.55% of replies to be CS (He et al. 2021). This is lower than what we have observed, where around 1.3% of total replies, including those to both HS and non-HS comments, were classified as CS. Also, on Twitter, it was estimated that around three out of four replies to HS were CS (Mathew et al. 2018), which aligns closely with the patterns observed in this study. These results reflect that although LGBTQ communities receive significant HS compared to other hate targets in different platforms, there is also a strong community support in the form of CS.

Hate Topics

Figure 2 presents the most frequent hate topics (out of a total of 130), with the top 15 highlighted. The topic modelling achieved a coherence score of .45, which is relatively satisfactory given the brevity and limited context of social media comments. Optimized HDBSCAN clustering parameters included a minimum cluster size of 8 and the Euclidean distance metric. In the visualisation, each point represents a comment from a non-outlier cluster in the embedding space, with clusters shown in different colours. Neutral pink tones represent comments outside the most frequent hate topics. Larger circles indicate clusters with more hate messages, while proximity in the embedding space reflects semantic and topical similarity. Topic labels were generated using zero-shot LLM labelling, with one name being pixelated for anonymity. It is important to clarify that topic interpretations relied not only on these labels but also on BERTopic cluster representations and comment examples within each cluster. For instance, the cluster labelled as *Barber Services* targeted transgender or non-binary individuals' body hair, underscoring the need for multi-faceted interpretations to fully understand the hate topics.

One focus of this study was to shift the perspective from exploring the causes of HS to examining its substance, what it consists of. In this context, and answering RQ2, we found that the most prominent themes of HS explicitly target LGBTQ topics, including aspects of identity and sexuality. Transgender and non-binary individuals face significant HS related to misgendering, deadnaming, bathroom access, sports participation, surgical procedures, portrayal as merely wearing costumes, and body-related hate, among others. These discourses reflect recurring themes in the literature on HS targeting LGBTQ communities, both online and offline (Casey et al. 2019; Dhiman 2023). Attacks on LGBTQ identities are sometimes intersected with racist HS or discrimination based on physical disabilities. Political debates and processes, particularly those concerning LGBTQ rights, also emerge as significant themes, along with family-related hate, especially around raising children. Other notable forms of HS include general slurs (e.g., "sickening", "gross", "you are a joke"), death threats, claims of mental illness (especially directed at transgender individuals), grooming accusations, and derogatory jokes, as well as portrayals of LGBTQ communities as predatory (Fernández 2024). Religious individuals also frequently target influencers, urging them to repent, turn to God, or claiming they will not be forgiven. Notably, despite TikTok's prohibition of practices such as labelling transgender individuals as mentally ill, deadnaming, or misgendering, these behaviours remain prevalent in daily online environments.

HS Theme	Prevalence	CS
Gender identity	41%	39%
Other LGBTQ explicit hate	13%	12%
General slurs	16%	8%
Religion	4%	16%
Political processes	5%	11%

Table 3: Prevalence of the most prominent HS themes in comments and the proportion of CS responses addressing each one.

To gain a deeper understanding of the scope of the identified hate topics, we manually grouped clusters based on their representations, labels, and text examples, and calculated the percentage each cluster contributed to the total hate. Table 3 represents the most prominent hate themes, their prevalence in our dataset of HS instances, and how much of the generated CS addressed such topic. For example, hate messages from religious individuals made up 4% of the total hate, while general slurs accounted for 16% of non-outlier hate messages. Although these slurs do not explicitly target gender identity or sexual orientation, this does not mean they are not motivated by these factors. In contrast, HS specifically targeting gender identity emerged as the most prevalent theme, including misgendering, deadnaming, body-related insults, and issues related to bathroom access and sports participation, which together comprised at least 41% of hate messages. Additionally, other forms of LGBTQ-related hate, not covered in the previous category, such as hate targeting sexual orientation or political processes related to LGBTQ rights, made up at least 18% of HS. Specifically, hate related to political processes accounted for 5% of hate comments. These percentages are considered conservative lower bounds, as they include only clusters where we were confident in the classification, excluding those we were uncertain about.

Thus far, we have observed that CS is unevenly distributed across hate comments, with the majority receiving none. To address the second part of RQ3, we analysed CS responses across the most prominent hate themes. General slurs, which account for 16% of total hate comments, receive only 8% of

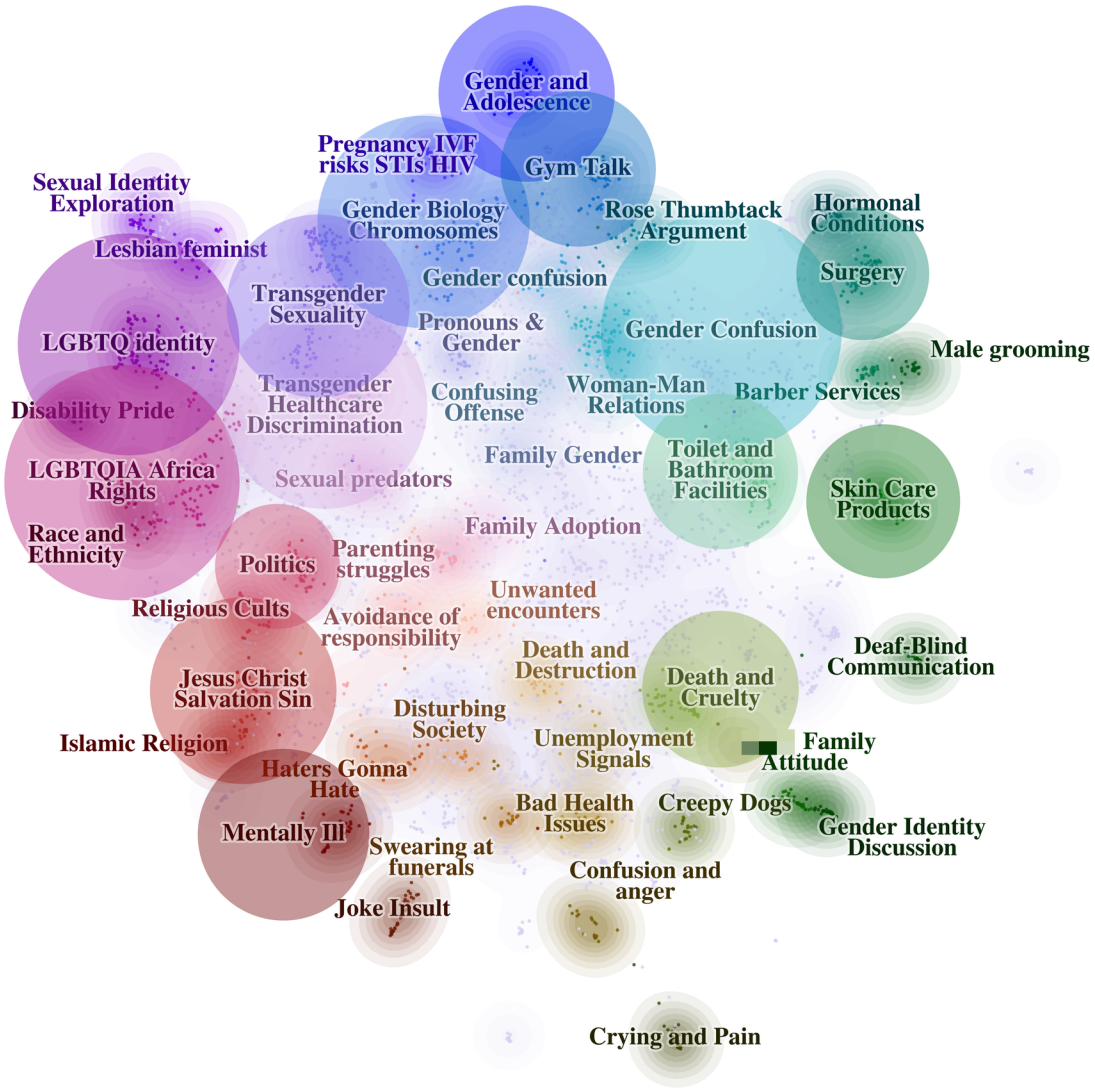

Figure 2: Representation of the embedding space of hate comments with their respective hate topics.

CS responses, with less than one in five comments attracting CS. In contrast, explicit LGBTQ-related hate topics receive a significantly higher level of community response in the form of CS. The most prominent hate theme, gender identity, garnered 39% of all CS responses, followed by other forms of LGBTQ-related hate not explicitly linked to gender identity, which received 12% of CS messages. Interestingly, despite their smaller proportions, hate messages from religious individuals and political processes receive substantial counteraction in the form of CS, accounting for 16% and 11% of CS responses, respectively. This highlights that while CS is unevenly distributed across hate comments, it is widely used as a response to certain hate themes targeting LGBTQ communities, particularly those explicitly focused on LGBTQ topics. Given its use by LGBTQ communities, this underscores the need for further research to explore its potential as a strategy to mitigate HS.

6 Conclusion

Despite TikTok's stated efforts to reduce HS, our study confirms that the platform remains far from free of hate. In this large-scale analysis of LGBTQ hate prevalence and substance in daily interactions, we found that at least 3.5% of content is classified as HS, with replies containing four times more hate than comments. This content originates from 4.07% of users. While general slurs make up a minority of the hate, at least 41% targets gender identity, particularly transgender and non-binary individuals, with recurring themes including misgendering, deadnaming, bathroom access, sports participation, body image, and surgery. Additionally, at least 18% of the hate is explicitly related to other LGBTQ topics. These findings highlight not only that HS remains pervasive, but that LGBTQ communities explicitly receive large amounts of hate targeting their identity, potentially leading to significant adverse effects. However,

the community actively counters much of this hate: 36% of hate comments receive a reply, with 75% of those replies being CS, resulting in approximately one-third of hate messages receiving at least one CS response. This suggests that LGBTQ communities have boundaries in terms of which hate they will respond to with inaction or other strategies, and which hate they will attempt to respond to with CS. Notably, LGBTQ-specific HS attracts more CS responses than general slurs. Certain themes, despite their lower prevalence, trigger a substantial community response in the form of CS, such as political hate and hate from religious individuals. These results highlight the role of CS as a proactive response to hate directed at LGBTQ communities, underscoring the importance of continued exploration to fully understand its potential to mitigate HS.

7 Limitations

This study constitutes an entry point to tackle the problem of HS against LGBTQ communities. As such, it did not aim to provide a comprehensive mapping of all instances of HS that may have targeted these actors. Rather, it sought to provide systematic evidence of what HS instances are made of and how TikTok users are responding to them. While contributing to further the study of LGBTQ-oriented HS, our study provides an entry point and is inevitably characterised by a set of limitations. First, and perhaps more importantly, our study adopts a large-scale quantitative approach for which the inherently diverse LGBTQ communities are treated as a homogeneous collective entity and are approached through the accounts of some of its very prominent members. More should be done in the future to do justice to the internal diversity of the LGBTQ collectivity and the HS experiences that all its members, including individuals with lower visibility, live and resist in their everyday lives. From a more technical perspective, instances of HS we identified and tackled are less numerous than in practice as a result of several methodological choices and factors: the prioritisation of precision over recall to minimize false positives for the subsequent topic modelling, the exclusion of non-textual (e.g., emoji-based) and non-explicit forms of HS, and the possibility to only work on publicly available data that survive deletion, user reporting and platform flagging. Additionally, our model was trained solely on existing datasets, whereas it appears crucial to develop and incorporate labelled TikTok data to improve its performance. In addition, relying on pre-existing datasets for HS and CS detection means that our models, and by extension this study, adopt the definitions of HS and CS established by the original dataset creators, potentially overlooking forms of HS or CS that are absent from those datasets. In the specific case of CS, further research within the field is needed to develop new datasets that draw from social media platforms and include labels beyond HS and CS. All of these elements should be addressed in future work to build a stronger foundation to detect and understand the potential of response strategies and, particularly, of CS.

References

Abarna, S.; Sheeba, J.; and Pradeep Devaneyan, S. 2023. A novel ensemble model for identification and classification of cyber harassment on social media platform. *Journal of Intelligent & Fuzzy Systems*, 45(1): 13–36.

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

Alshalan, R.; Al-Khalifa, H.; Alsaeed, D.; Al-Baity, H.; and Alshalan, S. 2020. Detection of hate speech in covid-19–related tweets in the arab region: Deep learning and topic modeling approach. *Journal of Medical Internet Research*, 22(12): e22609.

Askanius, T. 2021. On frogs, monkeys, and execution memes: Exploring the humor-hate nexus at the intersection of neo-Nazi and alt-right movements in Sweden. *Television & new media*, 22(2): 147–165.

Baider, F. 2020. Pragmatics lost? Overview, synthesis and proposition in defining online hate speech. *Pragmatics and Society*, 11(2): 196–218.

Baider, F. 2022. Covert hate speech, conspiracy theory and antisemitism: Linguistic analysis versus legal judgement. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 35(6): 2347–2371.

Baider, F. 2023. Accountability Issues, Online Covert Hate Speech, and the Efficacy of Counter-Speech. *Politics and Governance*, 11(2): 249–260.

Banerjee, S.; and Nguyen, H. 2023. Dataset for identification of queerphobia. *Journal of Student Research*, 12(1).

Bano, H.; Akbar, W.; Aslam, N.; and Bilal, M. 2023. Identification and Classification of Extremist by Topic Modeling Sentiment Analysis. *VFAST Transactions on Software Engineering*, 11(2): 235–248.

Barrientos, J.; Silva, J.; Catalán, S.; Gómez, F.; and Longueira, J. 2010. Discrimination and victimization: parade for lesbian, gay, bisexual, and transgender (LGBT) pride, in Chile. *Journal of homosexuality*, 57(6): 760–775.

Bartlett, J.; and Krasodomski-Jones, A. 2016. Counter-speech on Facebook. *Demos*.

Benesch, S.; Ruths, D.; Dillon, K. P.; Saleem, H. M.; and Wright, L. 2016. Counterspeech on twitter: A field study. Dangerous Speech Project. Available at: https://www.dangerousspeech.org/libraries/counterspeech-on-twitter-a-field-study. Accessed: 2025-04-07.

Bhat, P.; and Klein, O. 2020. Covert hate speech: White nationalists and dog whistle communication on twitter. *Twitter, the public sphere, and the chaos of online deliberation*, 151–172.

Bivens, R.; and Haimson, O. L. 2016. Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society*, 2(4): 2056305116672486.

Blackwell, L.; Dimond, J.; Schoenebeck, S.; and Lampe, C. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–19.

Bonaldi, H.; Dellantonio, S.; Tekiroglu, S. S.; and Guerini, M. 2022. Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8031–8049. Association for Computational Linguistics.

Braddock, K.; and Horgan, J. 2016. Towards a guide for constructing and disseminating counternarratives to reduce support for terrorism. *Studies in Conflict & Terrorism*, 39(5): 381–404.

Brown, A. 2015. *Hate Speech Law: A Philosophical Examination*. Taylor & Francis.

Brown, A. 2018. What is So Special About Online (as Compared to Offline) Hate Speech? *Ethnicities*, 18(3): 297–326.

Buerger, C. 2021. Counterspeech: A literature review. *Available at SSRN 4066882*.

Butler, J. 2011. *Bodies that matter: On the discursive limits of sex*. routledge.

Carratalá, A. 2023. The viralization of stigma online: Hate speech against gay men in connection with the monkeypox outbreak. *Hate Speech in Social Media*.

Caselli, T.; Basile, V.; Mitrović, J.; and Granitzer, M. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 17–25. Online: Association for Computational Linguistics.

Casey, L. S.; Reisner, S. L.; Findling, M. G.; Blendon, R. J.; Benson, J. M.; Sayde, J. M.; and Miller, C. 2019. Discrimination in the United States: Experiences of lesbian, gay, bisexual, transgender, and queer Americans. *Health services research*, 54: 1454–1466.

Chakravarthi, B. R. 2024. Detection of homophobia and transphobia in YouTube comments. *International Journal of Data Science and Analytics*, 18(1): 49–68.

Chakravarthi, B. R.; Kumaresan, P.; Priyadharshini, R.; Buitelaar, P.; Hegde, A.; Shashirekha, H.; Rajiakodi, S.; García, M. Á.; Jiménez-Zafra, S. M.; García-Díaz, J.; et al. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, 124–132.

Chakravarthi, B. R.; Priyadharshini, R.; Durairaj, T.; McCrae, J. P.; Buitelaar, P.; Kumaresan, P.; and Ponnusamy, R. 2022. 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*, 369–377.

Chung, Y.-L.; Kuzmenko, E.; Tekiroglu, S. S.; and Guerini, M. 2019. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2819–2829. Florence, Italy: Association for Computational Linguistics.

Citron, D. K.; and Norton, H. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91: 1435.

Delmonaco, D.; and Haimson, O. L. 2023. "Nothing that I was specifically looking for": LGBTQ+ youth and intentional sexual health information seeking. *Journal of LGBT Youth*, 20(4): 818–835.

Delmonaco, D.; Mayworm, S.; Thach, H.; Guberman, J.; Augusta, A.; and Haimson, O. 2024. "What are you doing, TikTok?" : How Marginalized Social Media Users Perceive, Theorize, and "Prove" Shadowbanning. *Proceedings of the ACM on Human-Computer Interaction*, 8: 1–39.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Dhiman, B. 2023. *Impact of Social Media Platforms on LGBTQA+ Community: A Critical Review*.

Dias Oliva, T.; Antonialli, D. M.; and Gomes, A. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, 25: 700–732.

Díaz-Torres, M. J.; Morán-Méndez, P. A.; Villasenor-Pineda, L.; Montes, M.; Aguilera, J.; and Meneses-Lerín, L. 2020. Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 132–136.

Dorn, R.; Kezar, L.; Morstatter, F.; and Lerman, K. 2024. Harmful Speech Detection by Language Models Exhibits Gender-Queer Dialect Bias. arXiv:2406.00020.

Duguay, S. 2019. "Running the Numbers": Modes of Microcelebrity Labor in Queer Women's Self-Representation on Instagram and Vine. *Social Media + Society*, 5(4): 2056305119894002.

Duguay, S.; Burgess, J.; and Suzor, N. 2020. Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence*, 26(2): 237–252.

Ellison, N. B.; Blackwell, L.; Lampe, C.; and Trieu, P. 2016. "The question exists, but you don't exist with it": Strategic anonymity in the social lives of adolescents. *Social Media+ Society*, 2(4): 2056305116670673.

ElSherief, M.; Ziems, C.; Muchlinski, D.; Anupindi, V.; Seybolt, J.; Choudhury, M. D.; and Yang, D. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. arXiv:2109.05322.

EnsembleData. 2025. Social Media data scraping through simple APIs. ensembledata.com/apis/docs. Accessed: 2025-04-06.

Fanton, M.; Bonaldi, H.; Tekiroğlu, S. S.; and Guerini, M. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3226–3240. Association for Computational Linguistics.

Fernández, F. J. S. 2024. Online Homophobia: Hate Speech and Conspiracy Theories towards LGBTQI+ people on Twitter in Spain. *Culture e Studi del Sociale*, 9(1): 39–56.

Ferris, L.; and Duguay, S. 2020. Tinder's lesbian digital imaginary: Investigating (im) permeable boundaries of sexual identity on a popular dating app. *New Media & Society*, 22(3): 489–506.

Fisher, C. B.; Tao, X.; and Ford, M. 2024. Social media: A double-edged sword for LGBTQ+ youth. *Computers in Human Behavior*, 156: 108194.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-04-06.

Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4): 1–30.

Gagliardone, I.; Gal, D.; Alves, T.; and Martinez, G. 2015. *Countering online hate speech*. Unesco Publishing.

Garcia, M. 2016. Racist in the Machine. *World Policy Journal*, 33(4): 111–117.

García-Prieto, V.; Bonilla-del Río, M.; and Figuereo-Benítez, J. C. 2024. Disability, hate speech and social media: video replies to haters on TikTok [Discapacidad, discursos de odio y redes sociales: video-respuestas a los haters en TikTok]. *Revista Latina de Comunicación Social*, (82): 1–21.

Garland, J.; Ghazi-Zahedi, K.; Young, J.-G.; Hébert-Dufresne, L.; and Galesic, M. 2020. Countering hate on social media: Large scale classification of hate and counter speech. arXiv:2006.01974.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Geet d'Sa, A.; Illina, I.; and Fohr, D. 2020. Classification of Hate Speech Using Deep Neural Networks. *Revue d'Information Scientifique & Technique* , 25(01).

Gitari, N. D.; Zuping, Z.; Damien, H.; and Long, J. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4): 215–230.

González-Aguilar, J. M.; Segado-Boj, F.; and Makhortykh, M. 2023. Populist right parties on TikTok: Spectacularization, personalization, and hate speech. *Media and communication*, 11(2): 232–240.

González-Esteban, J.-L.; Lopez-Rico, C. M.; Morales-Pino, L.; and Sabater-Quinto, F. 2024. Intensification of Hate Speech, Based on the Conversation Generated on TikTok during the Escalation of the War in the Middle East in 2023. *Social Sciences*, 13(1): 49.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794.

Haimson, O. L. 2018. *The social complexities of transgender identity disclosure on social media*. University of California, Irvine.

Haimson, O. L.; Delmonaco, D.; Nie, P.; and Wegner, A. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–35.

Han, X.; Han, W.; Qu, J.; Li, B.; and Zhu, Q. 2019. What happens online stays online? —— Social media dependency, online support behavior and offline effects for LGBT. *Computers in Human Behavior*, 93: 91–98.

Harel, T. O.; Jameson, J. K.; and Maoz, I. 2020. The normalization of hatred: Identity, affective polarization, and dehumanization on Facebook in the context of intractable political conflict. *Social Media+ Society*, 6(2): 2056305120913983.

Hawdon, J.; Oksanen, A.; and Räsänen, P. 2017. Exposure to Online Hate in Four Nations: A Cross-National Consideration. *Deviant Behavior*, 38(3): 254–266.

He, B.; Ziems, C.; Soni, S.; Ramakrishnan, N.; Yang, D.; and Kumar, S. 2021. Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 90–94. IEEE.

Hekma, G. 2016. Caught in a web? The Internet and deterritorialization of LGBT activism. In *The Ashgate research companion to lesbian and gay activism*, 225–242. Routledge.

Hietanen, M.; and Eddebo, J. 2023. Towards a definition of hate speech—with a focus on online contexts. *Journal of communication Inquiry*, 47(4): 440–458.

Hinduja, S.; and Patchin, J. W. 2023. Harassment in TikTok Comments: A Pilot Test of the TikTok Research API. Cyberbullying Research Center. https://cyberbullying.org/Harassment-in-TikTok-Comments-Research-API-Report.pdf. Accessed: 2025-04-07.

Hoffmann, A. L. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7): 900–915.

Hubbard, L. 2020. Online Hate Crime Report: Challenging online homophobia, biphobia and transphobia. London: Galop, the LGBT+ anti-violence charity. https://www.report-it.org.uk/files/online-crime-2020_0.pdf. Accessed: 2025-04-07.

HuggingFace. 2023. OpenHermes-2.5-Mistral-7B. https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B. Accessed: 2025-04-06.

Jacques, E. T.; Basch, C. H.; Fera, J.; and Jones II, V. 2023. # StopAsianHate: A content analysis of TikTok videos focused on racial discrimination against Asians and Asian Americans during the COVID-19 pandemic. *Dialogues in Health*, 2: 100089.

Kennedy, C. J.; Bacon, G.; Sahn, A.; and von Vacano, C. 2020. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. arXiv:2009.10277.

Koch, L.; Ghawi, R.; Pfeffer, J.; and Steinert, J. I. 2024. Online Misogyny Against Female Candidates in the 2022 Brazilian Elections: A Threat to Women's Political Representation? arXiv:2403.07523.

Kumar, G.; Singh, J. P.; and Kumar, A. 2021. A deep multi-modal neural network for the identification of hate speech from social media. In *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society: 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society, I3E 2021, Galway, Ireland, September 1–3, 2021, Proceedings 20*, 670–680. Springer.

Lingiardi, V.; Carone, N.; Semeraro, G.; Musto, C.; D'Amico, M.; and and, S. B. 2020. Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, 39(7): 711–721.

Lippe, P.; Holla, N.; Chandra, S.; Rajamanickam, S.; Antoniou, G.; Shutova, E.; and Yannakoudakis, H. 2020. A Multimodal Framework for the Detection of Hateful Memes. arXiv:2012.12871.

Ljubešić, N.; Fišer, D.; and Erjavec, T. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. arXiv:1906.02045.

Locatelli, D.; Damo, G.; and Nozza, D. 2023. A cross-lingual study of homotransphobia on twitter. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 16–24.

Lucero, L. 2017. Safe spaces in online places: Social media and LGBTQ youth. *Multicultural Education Review*, 9(2): 117–128.

Manukonda, D.; and Kodali, R. 2024. byteLLM@ LT-EDI-2024: Homophobia/Transphobia Detection in Social Media Comments-Custom Subword Tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, 157–163.

Marciano, A.; and Antebi-Gruszka, N. 2022. Offline and online discrimination and mental distress among lesbian, gay, and bisexual individuals: The moderating effect of LGBTQ Facebook use. *Media Psychology*, 25(1): 27–50.

Mathew, B.; Kumar, N.; Ravina; Goyal, P.; and Mukherjee, A. 2018. Analyzing the hate and counter speech accounts on Twitter. arXiv:1812.02712.

Mathew, B.; Saha, P.; Tharad, H.; Rajgaria, S.; Singhania, P.; Maity, S. K.; Goyal, P.; and Mukherjee, A. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, 369–380.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2022. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. arXiv:2012.10289.

McGowan, M. K. 2009. Oppressive speech. *Australasian Journal of Philosophy*, 87(3): 389–407.

Mollas, I.; Chrysopoulou, Z.; Karlos, S.; and Tsoumakas, G. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Mondal, M.; Silva, L. A.; and Benevenuto, F. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*, 85–94.

Musolff, A. 2015. Dehumanizing metaphors in UK immigrant debates in press and online media. *Journal of Language Aggression and Conflict*, 3(1): 41–56.

Negi Advocate, C. 2023. An Overview of Worldwide Cyberbullying and Cyberviolence Against Women, Teenagers, LGBTQ on Social Media: Facebook, Instagram, Telegram, WhatsApp, Snapchat, YouTube, LinkedIn and Twitter. *Boston College International and Comparative Law Review, Forthcoming*.

Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3): e1356.

Ollion, E.; Shen, R.; Macanovic, A.; and Chatelain, A. 2023. ChatGPT for text annotation? mind the hype. *SocArXiv preprint*, 32.

O'Connor, C. 2021. Hatescape: An in-depth analysis of extremism and hate speech on TikTok. *Institute for Strategic Dialogue*, 24.

Perifanos, K.; and Goutsos, D. 2021. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7): 34.

Rauchberg, J. S. 2022. # Shadowbanned: Queer, Trans, and Disabled creator responses to algorithmic oppression on TikTok. In *LGBTQ digital cultures*, 196–209. Routledge.

Reynders, D. 2022. 7th evaluation of the Code of Conduct. European Commision. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_7109. Accessed: 2025-04-07.

Richards, R. D.; and Calvert, C. 2000. Counterspeech 2000: A new look at the old remedy for bad speech. *BYU L. Rev.*, 553.

Röder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; and Pierrehumbert, J. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sánchez-Sánchez, A. M.; Ruiz-Muñoz, D.; and Sánchez-Sánchez, F. J. 2024. Mapping homophobia and transphobia on social media. *Sexuality Research and Social Policy*, 21(1): 210–226.

Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N. A.; and Choi, Y. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *ACL*.

Schieb, C.; and Preuss, M. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ica annual conference, at fukuoka, japan*, 1–23.

Semaan, B. 2019. 'Routine Infrastructuring'as' Building Everyday Resilience with Technology' When Disruption Becomes Ordinary. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24.

Shekhar, S.; Saini, A.; et al. 2021. Utilizing Topic Modelling to Identify Abusive Comments on YouTube. In *2021 International Conference on Intelligent Technologies (CONIT)*, 1–5. IEEE.

Silva, M. P. d.; and Silva, L. S. d. 2021. Hate speech dissemination in news comments: analysis of news about LGBT universe on Facebook cybermedia from Mato Grosso do Sul. *Intercom: Revista Brasileira de Ciências da Comunicação*, 44: 137–155.

Simpson, R. M. 2013. Dignity, harm, and hate speech. *Law and Philosophy*, 32(6): 701–728.

Soral, W.; Bilewicz, M.; and Winiewski, M. 2018. Exposure to Hate Speech Increases Prejudice through Desensitization. *Aggressive Behavior*, 44(2): 136–146.

Thurlow, C. 2001. Naming the "outsider within": homophobic pejoratives and the verbal abuse of lesbian, gay and bisexual high-school pupils. *Journal of Adolescence*, 24(1): 25–38.

TikTok. 2023. Community Guidelines Enforcement Report. https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-3/. Accessed: 2024-08-25.

TikTok. 2024. Hate Speech and Hateful Behavior. https://www.tiktok.com/community-guidelines/en/safety-civility/#2. Accessed: 2024-08-25.

Vidgen, B.; Margetts, H.; and Harris, A. 2019. How much online abuse is there. *Alan Turing Institute*, 11.

Walker, A. M.; and DeVito, M. A. 2020. "'More gay' fits in better": Intracommunity Power Dynamics and Harms in Online LGBTQ+ Spaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Walters, M. A.; Paterson, J.; Brown, R.; and McDonnell, L. 2020. Hate crimes against trans people: assessing emotions, behaviors, and attitudes toward criminal justice agencies. *Journal of interpersonal violence*, 35(21-22): 4583–4613.

Wilkinson, W. W.; and Berry, S. D. 2020. Together they are Troy and Chase: Who supports demonetization of gay content on YouTube? *Psychology of Popular Media*, 9(2): 224.

Yigezu, M. G.; Kolesnikova, O.; Sidorov, G.; and Gelbukh, A. F. 2023. Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification. In *IberLEF@SEPLN*.

Yu, X.; Blanco, E.; and Hong, L. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. arXiv:2206.06423.

Ștefăniță, O.; and Buf, D.-M. 2021. Hate Speech in Social Media and Its Effects on the LGBT Community: A Review of the Current Research. *Romanian Journal of Communication and Public Relations*, 23(1): 47–55.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Methods.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes, see Limitations.
   (f) Did you discuss any potential negative societal impacts of your work? Yes.
   (g) Did you discuss any potential misuse of your work? Yes, see Data.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Data.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, although only the code with mock data is provided for methodological transparency. The motivation to do so is explained in Data.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Methods.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, because the study did not require any extra compute resources.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Methods.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes.

(b) Did you mention the license of the assets? No, although we do not explicitly specify the license of the assets, we confirm our compliance with their terms. All datasets used were freely available for research purposes, with proper citation provided. In cases where the license required authors to be informed, this was duly carried out.

(c) Did you include any new assets in the supplemental material or as a URL? Yes, a link to the code with a mock dataset is included.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Data.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Data.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA