TIDEs: A Transgender and Nonbinary Community-Labeled Dataset and Model for Transphobia Identification in Digital Environments

Francesca Lameiro
University of Michigan
Ann Arbor, USA
flameiro@umich.edu

Lavinia Dunagan
University of Michigan
Ann Arbor, USA
laviniad@umich.edu

Dallas Card
University of Michigan
Ann Arbor, USA
dalc@umich.edu

Eric Gilbert
University of Michigan
Ann Arbor, USA
eegg@umich.edu

Oliver Haimson
University of Michigan
Ann Arbor, USA
haimson@umich.edu

Abstract

Transphobic rhetoric is a prevalent problem on social media that existing platform policies fail to meaningfully address. As such, trans people often create or adopt technologies independent from (but deployed within) platforms that help them mitigate the effects of facing transphobia online. In this paper, we introduce *TIDEs* (Transphobia Identification in Digital Environments), a dataset and model for detecting transphobic speech to contribute to the growing space of trans technologies for content moderation. We outline *care-centered data practices*, a methodology for constructing and labeling datasets for hate speech classification, which we developed while working closely with trans and nonbinary data annotators. Our fine-tuned DeBERTa model succeeds at detecting several ideologically distinct types of transphobia, achieving an F1 score of 0.81. As a publicly available dataset and model, TIDEs can serve as the base for future trans technologies and research that confronts and addresses the problem of online transphobia. Our results suggest that downstream applications of TIDEs may be deployable for reducing online harm for trans people.

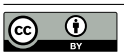
CCS Concepts

• **Human-centered computing** → *Interactive systems and tools.*

Keywords

transgender, content moderation, natural language processing, hate speech detection, trans technologies

ACM Reference Format:
Francesca Lameiro, Lavinia Dunagan, Dallas Card, Eric Gilbert, and Oliver Haimson. 2025. TIDEs: A Transgender and Nonbinary Community-Labeled Dataset and Model for Transphobia Identification in Digital Environments. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3715275.3732095



This work is licensed under a Creative Commons Attribution 4.0 International License.
FAccT '25, Athens, Greece
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732095

1 Introduction

In early January 2025, Meta announced an overhaul of their content moderation practices, which, in addition to ending their fact-checking program and introducing a community notes system, included "getting rid of a number of restrictions on topics like immigration, gender identity and gender" [40]. Despite framing the announcement as an expansion of user freedoms, Meta chose to include several examples of hateful speech that would now be permitted on Facebook, Instagram, and Threads, including calling trans[1] people mentally ill on the basis of their "transgenderism"[2][42]. Additionally, Meta's new moderation guidelines show that posts calling trans and nonbinary people "it" [55] or using violent anti-trans slurs [11] no longer violate their policies. While Meta's previous Hateful Conduct policy explicitly acknowledged that hateful speech online "may promote offline violence" [42], this reference has been removed.

Meta's new policies reflect a shifting political climate that increasingly targets trans people. The past few years have marked a significant increase in the number of anti-trans bills seeking to restrict trans people's access to gender-affirming healthcare, bathrooms, sports, and other facets of public life [38, 53, 61]. The prevalence of anti-trans rhetoric online is on the rise beyond Meta's platforms, with accounts like *Libs of Tiktok* and *EndWokeness* on X (formerly Twitter) amassing millions of followers online while inciting violence against trans people [13]. Trans people are especially vulnerable to online abuse and harassment, which they face at higher rates than cisgender people, including cisgender LGBT people [58], and they may also encounter general anti-trans rhetoric during their day-to-day social media use.

While some social media platforms deploy automated content moderation strategies to address hate speech [28, 31], some anti-trans content is often unmoderated [30]. In 2023, X removed its policy against deadnaming trans people and ceased removing tweets that fall under the category of hateful conduct [57]. Although the policy was later reinstated [9], this case demonstrates that current policies meant to protect trans people online are highly unstable,

[1]We use "trans" to refer to transgender people, including nonbinary transgender people.
[2]The term "transgenderism" is associated with anti-trans conservatism, and used to imply that being trans is an ideology rather than an identity that should be protected [29].

and that platforms may adopt anti-trans policies at any time—as seen with Meta's policy changes in early 2025. Platform policy depends on a wide range of factors that rapidly change and, as such, policy alone is not sufficiently reliable to ensure trans online safety. Some trans users abstain from certain online platforms in an attempt to avoid seeing transphobic posts [18], which may impede them from experiencing social media's positive effects [12, 67].

Because platforms often fail to effectively moderate anti-trans content, trans people have developed various mechanisms and tools to protect themselves from harassment and transphobia online. For example, some users avoid transphobia by liberally blocking transphobic users using blocklists (manually or algorithmically curated lists of users to be blocked automatically) [26, 39]. Though large transphobia blocklists have been developed and used in the past [1, 68], the blocklist is a limited tool that is contingent on a free, publicly-available API, as observed when X blocklists became obsolete after the platform significantly restricted the features in its free API tier [71]. In an online landscape that does not prioritize marginalized users' safety, many trans people protect themselves by creating and adopting tools that allow them to choose whose content they want to see or interact with. As technical, monetary, and platform-governance related constraints continue to limit trans people's ability to remove transphobic users from their feeds, new technologies are needed to foster trans safety online.

To meaningfully address online transphobic rhetoric, we must first be able to detect it. While there are existing tools that are frequently used to measure the toxicity of online speech, they are not currently optimized to identify transphobia in its many different forms. The Perspective API, a widely-deployed commercial tool for hate speech detection, performs only slightly better than chance at identifying trans-exclusionary radical feminist (TERF) transphobic rhetoric [47], and it cannot distinguish between different kinds of speech that target distinct marginalized groups. Past work on transphobia detection in the field of Natural Language Processing (NLP) has combined the tasks of homophobia and transphobia detection [15, 43], yet these are overlapping but distinct prejudices with distinct speech. Lu and Jurgens [47] created two models to recognize TERF transphobia, one which classifies X users and another which classifies X posts. However, the tasks of detecting general transphobia and TERF transphobia specifically have not yet been combined to create a dataset and model that can comprehensively detect different kinds of transphobia.

To further ongoing efforts to mitigate transphobia online, we created the TIDEs (Transphobia Identification in Digital Environments) dataset, which incorporates data from Tumblr, Truth Social, YouTube, and Reddit, and used it to fine-tune models for transphobic detection. Our best-performing model succeeded in detecting transphobic rhetoric, attaining an F1 score of 0.81. In building and creating this dataset, we placed special emphasis on both collecting data that captures the varied kinds of transphobia that trans people face online and on meaningfully engaging with and fairly compensating the data annotators. We worked with four trans and nonbinary data annotators to arrive at a final dataset that was carefully labeled by people who have personal experience with encountering and detecting online transphobia. Our work and care-centered approach have implications for informing research on transphobia detection and hate speech detection by centering care for the needs and experiences of those affected by hateful speech online.

1.1 Summary of contributions

We make the following contributions:

(1) We present *TIDEs*, a dataset and model for transphobia detection that can identify a combination of different kinds of transphobia from a range of different social media sources[3]. We also include our codebook from data labeling as a resource for future research.

TIDEs can support trans people in doing something that they already do on a daily basis when they use the internet: classify transphobic content. Transphobic content marks unsafe online territory that most trans people choose to avoid. Our transphobia classifier lays the groundwork for future tools and applications that address this problem, such as new systems that shield or alert users to transphobic content online or large scale computational analysis of transphobia on social media.

(2) We introduce the concept of *care-centered data practices*, which we define as intentional care practices for working with data annotators, data collection, and data annotation. We employed care-centered data practices to create TIDEs and we detail our processes so that future researchers can adopt our approach.

The care-centered data practices described in this paper were crucial to how we approached the task of creating a transphobia classifier. They shaped how and from which sources we collected data, how we annotated it, and how we engaged and worked with data annotators. Many of the decisions we made during TIDEs's development (including collecting multi-site data, hiring a trans and nonbinary annotation team, and holding annotation consensus meetings that yielded significant discussion) are not commonly employed in dataset construction and hate speech detection tasks. We argue that care-centered data practices enrich dataset development and research on hate speech classifiers within NLP, especially that which seeks to benefit marginalized communities. Approaching this work with care is a crucial step in improving dataset construction and ensuring that researchers involve data workers through fair labor practices that do not reproduce harm.

2 Related Work

Our paper builds on prior work on trans technologies, online safety for trans people, and hate speech classifiers by combining the three, to contribute a new trans technology that is a hate speech classifier that supplements content moderation efforts. Here we discuss the various technologies that have influenced and motivate the need for a robust transphobia classifier.

2.1 Trans Technologies

Trans technologies address issues faced by trans people [33, 36] and are aligned with the queer notions of transition, multiplicity, fluidity, and ambiguity [34]. Trans technologies are often born out of urgent necessity to increase trans people's safety or access to basic resources [36]. As a novel tool for improved transphobia

[3]TIDEs is available at http://www.francescalameiro.com/tides.

detection online, TIDEs can be deployed in various ways, such as removing transphobic content from user feeds or conducting network analyses to understand the spread of transphobic rhetoric. By creating TIDEs, we advance work that creates safer online spaces for trans people; thus, TIDEs qualifies as a kind of trans technology.

2.2 Online Safety for Trans People

2.2.1 Transphobia Online and Content Moderation. Trans people rely on social media to build community [17, 65, 67], find and share gender-affirming care resources [7], and express their identities [7, 32]. However, they also encounter harassment and transphobic speech online [19, 58], which may impact how they use and benefit from social media. Content moderation, widely used on social media, is a strategy to mitigate online harms. Most platforms generally screen user-generated content and remove, censor, or suppress content that they or user-moderators deem illicit [27, 28].

Content moderation currently falls short in the face of pervasive anti-trans content online. Platform efforts to moderate hateful content may simply be insufficient to adequately address it, despite best efforts. Before Meta enacted its new transphobic policies, the company prohibited slurs, dehumanizing speech that equates marginalized people to sexual predators and terrorists, promotion of conversion therapy, targeted misgendering, and other kinds of hateful speech on its platforms [51], but it failed to adequately remove content that violated these policies [30]. This problem is pervasive across many platforms; while X's hate speech policy technically forbids the targeted usage of the word "groomer" against trans people, the platform does little to enforce this rule [14]. Furthermore, hateful content can remain unaddressed even on user-moderated platforms like Reddit, which relies on user-driven content moderation, charging a group of users ("moderators") to moderate each subreddit [66].

Beyond failing to remove harmful posts, platforms sometimes change their policies to actively allow or promote hateful content, as observed by X's decision deeming "cisgender" a slur [59] and Meta's new policies. X not only allows transphobic content to be prominently displayed, but also censors any posts containing the word. Additionally, trans and Black users experience disproportionately high rates of content removal online [35]. Due to poor content moderation practices, marginalized users often sense that their best interests are not considered [22, 35, 49]. As such, Thach and colleagues [69] have introduced the concept of *trans-centered moderation*, an approach that seeks to improve content moderation by centering and addressing the needs of trans people.

2.2.2 Systems for mitigating transphobia. As platforms and official content moderation strategies and systems are often unable to protect trans users and in some cases and have actively harmed them, many trans users use and create third-party systems to protect themselves online [69].

Shinigami Eyes [41] is a browser extension that marks anti-trans and trans-friendly social media accounts, pages, and websites by highlighting them with distinct colors (by default, red for anti-trans and green for trans-friendly). The creator of Shinigami Eyes states that it was "created through a mix of manual labeling and machine learning" [41]. Users can right-click on any online entity to mark it as anti-trans or trans-friendly. Shinigami Eyes alerts users when they come across content that was created by a transphobic person or source, regardless of whether the specific content they are seeing is transphobic. The anonymous developer describes being motivated to combat the increasingly subtle forms of transphobia that have become more prominent with increased trans visibility, including trans-exclusionary radical feminism (TERF) [8, 41]. While transphobic rhetoric from conservatives is usually easy to spot, transphobia from self-proclaimed feminists or queer people may be harder to discern [47]. Despite its benefits, Shinigami Eyes does not prevent the user from coming across transphobic content; rather, it primes them to be alert to the possibility of transphobic rhetoric within another user's post. Additionally, the system's list of user classifications is not publicly available, meaning that third parties cannot validate the dataset or use it to develop new tools. Shinigami Eyes is also focused on identifying transphobic users or entities, rather than individual pieces of content or ideas, a limitation that could be addressed in future work [62].

The Aegis Security System [2, 69] is a content moderation trans technology that helps to shield online communities from harassment. Aegis enables trans Twitch and Discord communities to report users who engage in homophobic and transphobic harassment and hate raids, facilitating content moderation for specific LGBTQ-centered online communities by permanently banning abusive users. Aegis is currently a community-centered, invite-only system, meaning that an individual user cannot use the system to block transphobic users. Rather, a Twitch streamer or Discord moderator must apply to use Aegis in the community that they host. Nonetheless, Aegis is an exciting tool that can positively impact trans and queer online communities.

For several years, blocklists served as tools that enabling individual users to block large numbers of people online. On Tumblr, users shared blocklists containing hundreds of TERF users as a way to avoid contact with TERF communities on the platform [1, 68], and on X, Erin Reed popularized a transphobe blocklist with over 50,000 accounts and a TERF blocklist with over 200,000 accounts [60]. Unlike Shinigami Eyes, blocklists are definitive because they remove a list of targeted users from users' feeds. However, blocklists require APIs, and in the past relied on free or low cost API tiers. When X changed its API pricing model, many user-developed automated blocklist tools became obsolete. There is no longer a free, automated way to block new users on X, highlighting the instability of both platforms themselves and community tools that rely on mainstream platforms.

TIDEs directly builds on the ideas behind blocklists, Aegis, and Shinigami Eyes by empowering users to shield themselves and their online communities from harassment and transphobic speech. Unlike blocklists, TIDEs can moderate content without relying directly on unstable tools developed by platforms. Unlike Aegis, TIDEs can be adopted by individuals, in addition to larger online communities. Unlike Shinigami Eyes, TIDEs focuses on post-level rather than user-level transphobia detection, as suggested by [62]. Though each system has advantages and disadvantages, TIDEs serves as a novel tool for transphobia detection within self- or community-governed content moderation.

2.3 Hate Speech Detection

We outline previous work on hate speech classification in the field of NLP, and some of that work's shortcomings for the particular task of transphobia detection.

2.3.1 Hate speech classifiers. In the past eight years, language models built upon the transformer architecture [70], such as BERT (Bidirectional Encoder Representations from Transformers) [23] and RoBERTa (Robustly Optimized BERT-Pretraining Approach) [45], have been used as powerful base models for a variety of NLP tasks.

Although fine-tuned models are powerful tools with the potential to improve hate speech detection and content moderation online, they have some shortcomings [44]. Racism classifiers, for example, sometimes disproportionately flag Black users as being racist when using reclaiming slurs [64]. Effective hate speech detection must account for marginalized communities reclaiming "offensive" language typically used against them [24]. It is crucial that hate speech detection does not reinforce the very issues it seeks to address by penalizing marginalized users.

2.3.2 Transphobia Classifiers. Within NLP, a limited body of research has focused on hate speech detection classifiers for transphobia. Chakravarthi et al. [15] conducted a shared task on homophobia and transphobia detection in social media comments, using a dataset of mixed Tamil-English YouTube comments labeled as "homophobic," "transphobic," or "non-anti-LGBTQ+ content." This shared task was part of a broader project aiming to develop hate speech detection tools for low-resource languages. It includes only 13 "transphobic"-labeled instances in English. The researchers trained various base models, including BERT [54], RoBERTa [48], and ensemble models comprised of both BERT, RoBERTa, and HateBERT [56]. Kumaresan et al. [43] trained similar, YouTube data based models that detect homophobic and transphobic language in Malayalam and Hindi, in addition to English, Tamil, and Tamil-English. This prior work enables comparison of different base models for transphobia detection, and lays the ground for future fine-tuned models trained with larger datasets that solely focus on detecting transphobia. Researchers have also used machine learning systems to specifically detect TERF speech [47]. Lu and Jurgens [47] used a Twitter dataset to successfully detect both TERF users and TERF rhetoric. This is an important finding, because TERF transphobia is often characterized by subtle language that can remain undetected by toxicity detection tools such as Perspective API.

Prior work demonstrates that hate speech detection tools can detect transphobia on social media. We extend it by creating a robust, multi-domain transphobia dataset and classifier, which includes both TERF rhetoric and other kinds of transphobia, to improve transphobia detection, which can help address transphobic speech online.

3 Method

We describe the process of developing TIDEs, including data collection, data labeling, and model training. We detail the care-centered data practices we employed, which impacted the construction of the dataset and our labeling work with data annotators, and which differ from the methods traditionally employed for hate speech detection tasks.

3.1 Data Collection

To build a dataset that captures a varied sample of online transphobic content, we chose a multi-site approach, gathering text posts from four social media platforms: YouTube, Reddit, Tumblr, and Truth Social. These platforms were chosen on the basis of their popularity and the ease of accessing post data. The dataset contains YouTube comments on popular trans creators' videos, Reddit posts and comments from trans-related and political subreddits, Tumblr posts with transphobic tags, and Truth Social posts containing the word "transgender." The sampling decisions described here were informed by our own understanding of discourses about trans people online, resulting in significantly different data collection strategies for each platform. Next, we outline our data collection methods for each platform.

3.1.1 YouTube. YouTube comments were collected from a dataset of videos uploaded by 39 English-speaking trans YouTube content creators with large followings. We sourced this list of content creators from a study that our colleagues conducted (not yet published), that asked a large sample of trans young people to list trans content creators they follow. We collected 38 comments from each content creator's most popular video, excluding videos that were not explicitly about trans topics.[4] By focusing on videos about trans topics, we were able to collect a greater number of comments where commenters referred specifically to transness and trans people, both in transphobic and not transphobic ways. Comments were sorted and collected by "Most Recent" using the YouTube API. YouTube sorts comments by "Top Comments" by default, generally displaying comments in order of the amount of engagement they receive. On trans content creators' videos, the "Top Comments" tend to be positive comments from fans and other sympathetic viewers. By instead collecting the most recent comments from each creator's most popular trans-related video, we obtained more transphobic comments.

3.1.2 Reddit. Reddit posts and comments were collected from 25 subreddits. Subreddits fell into four categories: trans-related (n = 16), LGBT-related (n = 3), anti-SJW (n = 3)[5], and political discussion (n = 3) (see Table 3). We selected subreddits using two strategies: first by consulting community-curated lists of transphobic subreddits and second by searching for the keywords "trans" and "gender" in an archive of all subreddits. While selecting subreddits to collect data from, we loosely categorized them as transphobic or not transphobic. We do not intend these subreddit categories to prescriptively imply that every user that posts in a specific subreddit must necessarily be transphobic or not, but rather to source a mix of transphobic and not transphobic posts.

[4] For example, if the content creator's most popular video was a cover of a song and their second most popular video was about a gender affirming surgery they underwent, we chose to collect comments from the second video. We define "videos about trans topics" as videos where the content creator is primarily discussing their trans identity, social or medical transition, "coming out," or any other topic linked to their transness.
[5] SJW stands for *social justice warrior*, a derogatory term that is used to describe people who care about social justice. While anti-SJW rhetoric is not necessarily transphobic, trans people have been associated with the term SJW.

Out of the 16 trans-related subreddits included, 10 of them were also categorized as not transphobic and 6 of them were categorized as transphobic. Not transphobic trans-related subreddits like `r/ask_transgender` and `r/transgender` were chosen to represent posts that discuss trans topics, often shared in online communities where transphobia is not tolerated. Conversely, transphobic trans-related subreddits like `r/GenderCritical` and `r/assignedmale` were chosen to represent explicitly transphobic posts, often shared in online communities where transphobia is not only common but also what binds the community together in the first place. Out of the three LGBT-related subreddits included, two of them were categorized as not transphobic and one was categorized as transphobic. `r/ActuallyLesbian` is a transphobic lesbian community subreddit that explicitly excludes trans women, which led to the creation of the alternative, trans-friendly subreddits `r/ActualLesbians` and `r/ActualLesbiansOver25`. All three anti-SJW subreddits included were classified as transphobic because they were chosen from the aforementioned community-curated lists of transphobic subreddits. However, we acknowledge that anti-SJW subreddits are generally less transphobic than trans-related transphobic subreddits because their topic of focus can include but is not necessarily trans people. Lastly, we did not categorize the three political discussion subreddits as transphobic or not transphobic because they facilitate political discussions on many topics and welcome diverse opinions. We chose to collect data from these subreddits to capture general discussion about trans issues on forums that are not specifically focused on trans topics. To capture relevant posts from these subreddits, we searched for trans-related keywords (see Appendix). We chose a mix of general terms related to trans people as well as terms that have become relevant because of anti-trans news trends.

3.1.3 Tumblr. Tumblr posts were collected from the following tags: `#gender critical`, `#lgb`, `#lgbdropthet`, `#terfblr`, `#troons`, and `#radical feminists do interact`. We used Tumblr data to ensure that TERF rhetoric would be represented in the dataset, as the platform has a substantial active TERF community. We did not collect data from the more popular hashtag `#terf` because, at the time of data collection, it was being spammed by bots, and the resulting data was overwhelmingly unusable. Additionally, popular Tumblr tags tend to get used for posts on a broad range of topics, leading to less relevant content for the purposes of collecting TERF data. As such, we chose to begin with `#gender critical`, a term that has essentially become synonymous with TERF as the latter term has gained a negative connotation. From the `#gender critical` tag, we selected a few additional tags that were also used in `#gender critical` posts. On Tumblr, it is common to use many tags on a single post to increase its reach and ensure it appears on the feeds of like-minded users. We collected posts from these TERF-related tags and excluded any duplicates.

3.1.4 Truth Social. Truth Social posts were collected by running a search query for "transgender" using the Truthbrush API [50]. We opted for this simple data collection method because Truth Social, which is owned by Donald Trump's Trump Media & Technology Group, is a smaller platform with a majority right-wing and transphobic user base.

3.2 Data Labeling

The dataset was labeled by the first author and four data annotators who were compensated at $20/hour. All annotators were trans or nonbinary social media users. We choose to disclose the annotators' background and compensation because we view these factors as integral to both the construction of a robust dataset and the well-being of the workers who contributed to it. We sought to adopt data annotation practices that are congruent with the larger project of combating transphobia. Thus, we avoided hiring undercompensated data workers, and instead chose to hire annotators that could become meaningful members of the project team.

Before hiring each annotator, we screened them for their familiarity with online transphobia by having an initial meeting with them. We explained the project to each annotator, addressing any of their questions or concerns and ensuring they understood our goal and the work's potential effects on their well-being. Additionally, we ensured that annotators understood that they could stop working on the dataset at any point without penalty. We discuss the implementation and implication of these practices of care toward annotators in the Discussion section.

Each instance in the dataset was labeled by the first author and one other annotator as "Transphobic" or "Not transphobic." Figure 1 and Table 2 (in Appendix) describe the characteristics of our dataset by platform. The "Not transphobic" class includes both posts about trans and non-trans related topics. After every instance was labeled by two annotators, they held consensus meetings to discuss discrepancies in these annotations. This process surfaced various kinds of rhetoric present in the dataset that were particularly difficult to label as transphobic or not, particularly those that were very borderline transphobic or not. We initially disagreed and then resolved the discrepancies in labeling on 250 Tumblr instances, 66 YouTube instances, 99 Truth Social instances, and 128 Reddit instances.

Take, for example, the instance: "I don't know why people say he is a girl he looks like a guy" **Label**: *Transphobic*. Initially, one annotator labeled this post as transphobic and the other labeled it as not transphobic. The first annotator read this post as being about a trans woman who is being misgendered, while the second annotator read it as being about a trans man who the poster is defending by saying he should not be misgendered because he looks like a man. Without context, we cannot say which reading is correct. However, we ultimately chose to label this instance as transphobic because the poster is flippantly discussing and assuming a trans person's gender identity from their presentation. At the very least, this is a misguided and unnecessary way to discuss a trans person and their appearance, even if it was meant to be affirming.

There were other types of difficult-to-label posts, such as posts that are overtly anti-men or anti-queer (as a word and identity). Neither of these sentiments is necessarily transphobic, but anti-men and anti-queer sentiments are often expressed by transphobic people, in both directly and indirectly transphobic ways. Many transphobes choose to express hatred for "men" as a stand-in for trans women. Some lesbian, gay, or bisexual cisgender people choose to express their hatred for people who identify as "queer" as a stand-in for trans people. While annotators reached a final decision for every instance, challenging posts that fell under topics like these were additionally coded with descriptive sub-codes [63] representing

the rhetorical categories they fell under (e.g., "uninformed", "anti-man", "anti-queer"). We further discuss these more granular labels' purpose and implications in the Discussion section.

3.2.1 Codebook for Transphobic Data Annotation. Before the data annotation process began, we created a preliminary codebook that identified basic examples of common transphobic rhetoric (e.g., misgendering, calling trans people mentally ill). After hiring data annotators, we shared this version of the codebook with them, clarifying that it was not a comprehensive list of all types of transphobia, and encouraged them to contribute to the codebook as they annotated. Since we worked with trans and nonbinary annotators because of their understanding and familiarity with online transphobia, we asked them to exercise their judgment when annotating transphobic instances in the dataset and adding to the codebook. We have included this codebook in Appendix C.

3.3 Model Training

After finalizing the dataset, we split it into train, development, and test sets with an 80:10:10 ratio. We tested two models: a Logistic Regression model and a fine-tuned DeBERTa model [37]. The DeBERTa model was fine-tuned using the AdamW optimizer [46] with a learning rate of 3.19e-05, weight decay of about 0.07, and a batch size of 16. The model was fine-tuned over 10 epochs. For the Logistic Regression baseline, we tokenized the text using `deberta-v3-small`'s tokenizer and produced count vectors representing the resulting lists of tokens. We report the tokens associated with the highest and lowest coefficients in the Appendix, Table 4. We compared these models against the Perspective API as a baseline, using the *toxicity* and *identity attack* metrics.

4 TIDEs: a Dataset and Model for Transphobia Detection in Digital Environments

4.1 Dataset

The final TIDEs dataset (See Table 2, Figure 1) contains a balanced mix of posts across the two different classes, with 54.8% not transphobic posts (n = 1,907) and 45.1% transphobic posts (n = 1,572), for a total of 3,479 posts. Truth Social had the highest proportion of transphobic posts, while Reddit had the highest proportion of not transphobic instances.

Our diversified data collection methods succeeded in representing different transphobic ideologies in the dataset. When collecting data for TIDEs, it was crucial to avoid flattening transphobic data into one cohesive category. If a transphobia classifier is not trained on TERF data, for example, it will fail to detect a significant amount of transphobic rhetoric that greatly impacts trans people. As such, we have deliberately chosen to outline the goals and process of collecting data from each source so that we may move toward a dataset that represents transphobia in its varied and complex forms. Throughout the process of data annotation, we identified that the majority of transphobic instances included rhetoric aligned with three primary groups: right-wing transphobes, TERF transphobes, and casual transphobes. Truth Social was strongly linked to right-wing transphobes. These are users who are politically aligned with the U.S. Republican Party as it is currently led by Donald Trump. They are transphobes who subscribe to conservative beliefs and

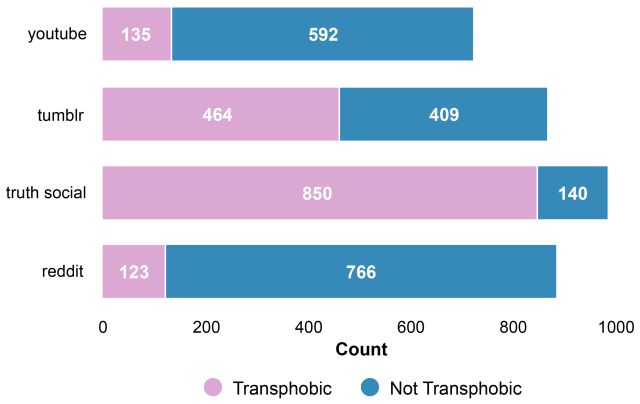


Figure 1: Dataset distribution across source and label

support the numerous legislative attacks on trans people. Similarly, trans-exclusionary radical feminist (TERF) [8] rhetoric was frequently found in the Tumblr data, and in some Reddit instances. TERFs tend to be women and associated with the U.S. "left," self-identifying as feminist and believing that trans people and "gender ideology" pose a threat to women and feminism. They often target trans women specifically. TERFs frequently use coded language such as "TIM"/"TIF" (trans-identified male/trans-identified female) and "TRA" (trans rights activist) to derogatorily refer to trans people. Lastly, we captured transphobia from casual transphobes, by which we mean people who hold some transphobic beliefs and may sometimes express them online, but who are not overtly committed to spreading transphobic rhetoric directly driven by a strong commitment to far-right conservative beliefs or TERF beliefs. Casual transphobic rhetoric was most concentrated in the YouTube data, though it was found across all data sources.

We collected two types of data. First, we collected comments that are directed or in response to other users: YouTube comments left on videos (n = 727) and Reddit comments addressing other users' videos or posts (n = 594). Second, we collected Reddit, Tumblr, and Truth Social posts meant as general statements, opinions, or questions (n = 2158).

4.1.1 Additional labeling. Data annotator consensus meetings highlighted several categories of posts that were difficult to objectively label as transphobic or not. Take for example "anti-man" posts, which express negative views about men generally or a given behavior that a poster may associate with men. Many women and trans and nonbinary people who are not men may post online about their frustrations with men in a patriarchal society and, as such, we chose to label some of these posts as not transphobic. However, this is a noteworthy example because some TERFs may post about their grievances with "men" or "males" in a coded manner that refers to trans women and nonbinary people.

> "The sooner you as a Woman accept that males are biologically incapable of emotional intelligence, love or empathy; the sooner you set yourself free."
> **Label**: *Transphobic*

This post includes no explicit reference to trans people, but because the use of the terms "males" and "biologically" to state that men cannot feel basic human emotions, we agreed that this instance to very likely be transphobic. Even if the text was meant to target cisgender men and not trans women, there is still a strong implication of biological determinism and essentialism that is highly incompatible with trans identity.

In terms of potential applications for TIDEs and what a transphobia classifier should do, we considered that some potential users may definitely consider "anti-man" posts to be transphobic, offensive, or essentialist, while other users may not mind them at all. We descriptively coded these instances during consensus meetings with sub-codes, arriving at the following categories: anti-queer (n = 16), bio-essentialist (n = 15), uninformed (n = 15), radical feminist (n = 11), and sex-based language (n = 11).[6] We chose to add this layer of qualitative coding to reflect some common characteristics of different instances in the dataset. Further descriptive coding of the dataset could allow for more granular labels among and across the three primary groups of transphobes.

Table 1: Performance on recognizing transphobic posts

Model	Acc	Prec	Rec	Mac F1	F1
Perspective API (*toxicity*)	0.63	0.91	0.59	0.60	0.72
Perspective API (*identity attack*)	0.71	0.80	0.68	0.70	0.74
Logistic Regression	0.79	0.87	0.70	0.79	0.77
DeBERTa	0.82	0.89	0.74	0.82	0.81

4.2 Model Performance

Overall, the TIDEs model succeeded in detecting transphobia in the dataset, outperforming our baselines. The DeBERTa model achieved an F1 score of 0.81, with precision of 0.89 and recall of 0.75. The Logistic Regression model attained an F1 score of 0.77, with precision of 0.87, and recall of 0.70. Using the *toxicity* metric on the best-performing threshold (0.1), the Perspective API attained an F1 score of 0.72, with precision of 0.91, and recall of 0.59. Using the *identity attack* metric on the best-performing threshold (0.1), the Perspective API attained an F1 score of 0.74, with precision of 0.80, and recall of 0.68. Model performance metrics are summarized in Table 1.

The DeBERTa model outperformed the Logistic Regression model, indicating that, in come contexts, transphobia is covert enough that the purely lexical cues in a bag-of-words model do not sufficiently describe it. Figure 2 shows the model performance on data from each individual source domain. The model performs best on Tumblr and Truth Social data and worst on Reddit and YouTube data, especially on recall. As a comparative baseline, the Perspective API was able to identify some transphobia when set to a very low threshold, though it performed worse than the TIDEs model.

[6]These numbers are lower bounds, since we only added these additional codes to posts that required extra discussion to reach consensus.

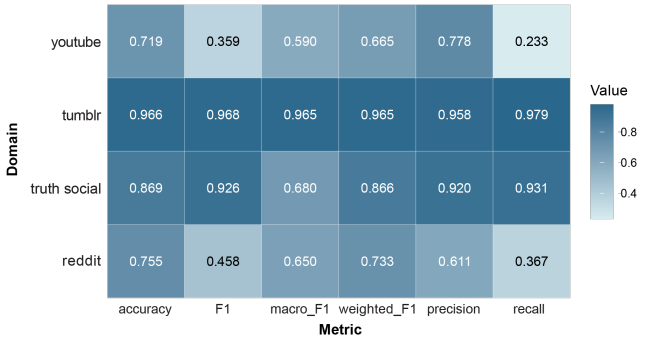


Figure 2: TIDEs model performance on each source domain

Further iteration on the dataset and model would be needed before it could be deployed for content moderation tasks that remove or hide content. Table 5 in the Appendix shows some examples of misclassifications from the testing set. Since our dataset included a very wide range of rhetoric, the model detects many kinds of transphobia, but it still struggles to detect transphobia in some subtle (row 2) or harder to parse (row 1) instances. Conversely, there were some false positives on shorter posts with less context (row 4) and similarly on harder to parse posts. Nevertheless, our metrics demonstrate good performance on a model that combines many different kinds of transphobic data. TIDEs can detect a high percentage of transphobic rhetoric, which indicates its potential for being deployed in limited settings where content is flagged rather than censored. Model performance was best on the two source domains that had the highest proportion of transphobic data (Tumblr and Truth Social).

5 Discussion

We outlined the process of creating TIDEs, a transphobia classifier for social media contexts, through the use of care-centered data practices. Our model performance metrics demonstrate that a single classifier is able to detect transphobic rhetoric from a variety of different sources and actors, highlighting the potential for it to be deployed to address transphobia across social media contexts, with some limitations. As a result of our care-centered approach, we understand transphobic rhetoric online as a multifaceted issue that can manifest in different forms and contexts, and we capture these in the TIDEs dataset.

In this section, we discuss both the development and implication of care-based practices, as well as the contribution of TIDEs as a publicly available system. We then address limitations and possible directions for future work.

5.1 Adopting Care-Centered Data Practices

In constructing and labeling the TIDEs dataset, we intentionally adopted several practices of care [21] that are not commonly used when building datasets for classifiers. We draw from de la Bellacasa's notion of *matters of care*, which views practicing care as a collaborative, political, ethical process, and which has inspired new approaches to data and technology research [10, 20].

Drawing from our process creating TIDEs we introduce the concept of *care-centered data practices*, which span hiring and working with annotators, collecting data, and labeling data. We define care-centered data practices as practices grounded in care for a marginalized group (in this case trans people), and which shape every stage of the construction of a dataset. Care-centered data practices allowed us to revisit commonly-used methods in data work and consider how they could be adapted to better support trans people in the context of transphobia detection. Next, we discuss the motivation behind these practices and the benefits that we observed from their adoption.

5.1.1 Care-centered hiring and working with data annotators. Trans and nonbinary social media users have significant experience with encountering online transphobia and, as such, we chose to hire annotators from these groups. As researchers conducting work that is meant to promote inclusion and safety for marginalized communities, we explicitly chose to ground our work in practices that center not only annotator expertise, but also their well-being. While we believe that our care-centered data practices had a positive impact on the data, resulting in a carefully constructed dataset, our priority was to ensure that annotators would be safe and comfortable.

Prioritizing annotator well-being can help to mitigate the negative impacts of hate speech annotation [3]. Annotators were encouraged to space out labeling across several sessions and take frequent breaks [52]. We explicitly told annotators that they could choose to stop working at any point with no penalty. All of these decisions helped annotators understand the stakes of their work and protect themselves. Annotators were meaningfully involved throughout the entirety of the project; they understood the goal of the dataset and classifier and how their annotations would be used. Checking in frequently and working closely with annotators was crucial for safeguarding their well-being.

During consensus labeling sessions, annotators reflected on their labeling process and how they throughout, following similar practices to those highlighted in [6] and [5]. As Andalibi and Forte [4] have discussed, it is crucial that researchers be aware of the negative impacts that conducting difficult research may have on them. While prior work has emphasized the importance of building peer support networks among researchers [16], we highlight that data annotators are often not considered part of a research team, and as such they are excluded from these networks. Research teams may not consider the impact that hate speech annotation may have on annotators, not only because annotators are not considered part of the research team, but also because there is typically very little interaction between researchers and annotators who are hired on platforms like Amazon Mechanical Turk. We address this by not only thinking deeply about and making intentional decisions to prioritize annotators' well-being, but also by making them part of our research team. We encourage research teams who frequently construct annotated datasets, especially those seeking to serve marginalized communities, to integrate annotators from the communities they study into their research teams, and to consider annotator well-being a central ethical concern.

5.1.2 Care-centered data collection. Detecting transphobia is a complex task because transphobic actors with varying ideologies express their transphobia in considerably different ways. Grounded in *care* for trans communities [21], we captured a range of transphobic ideologies by collecting training data from several platforms with differing post formats and user bases. Through this process, we were able to include data from right-wing transphobes, TERF transphobes, and casual transphobes. Our approach extends prior work that has separated these tasks [15, 47] by joining them into one, thus making it possible to detect complex and varied transphobic speech.

These data collection decisions are practices of care, because we approach this research problem in a way that accounts for the many complex ways transphobia emerges and affects trans people. Due to the variety of transphobic users' posts represented, our dataset contains a wide variety of transphobic arguments with differing motivations. We avoid thinking of transphobia as a form of bigotry that is simply expressed and easily recognizable. Instead, we reckon with the fact that a useful transphobia detection tool must be developed with an awareness of the idiosyncrasies of transphobia, so that non-conventional speech does not go undetected. Without acknowledging the complexity and variety of transphobic rhetoric online, we would risk creating a classifier that may perform well in tests, but that cannot be applied in social media contexts to detect non-extreme forms of transphobia. In this sense, our work is grounded in a situated understanding of what makes transphobia complex and difficult to detect, and we plan to grow the applications of TIDEs beyond the scope of this paper.

5.1.3 Care-centered data labeling. We resolved discrepancies in labeling between annotators by holding consensus meetings, where the two annotators who labeled each portion of the dataset discussed each specific case where their labels differed. By reading each of these instances together and discussing them at length, we reached a carefully constructed final dataset that annotators fully agreed on, with additional comments on particularly difficult instances. Spending additional cost and time to reach consensus on data annotation is not a standard practice in NLP hate speech classification tasks, but it allowed for deeper engagement with the data and the data annotators. Further, dedicated discussions that considered what transphobia is and how it manifests served as processes that demonstrated care for the annotators and for the data.

The careful process of creating TIDEs foregrounded the context- and reader-dependent nature of transphobic rhetoric. We annotated difficult instances with sub-codes to better capture the data's complexity. Instead of prescriptively and definitively deciding whether all "anti-man" or "anti-queer" instances are transphobic or not, we have labeled these instances as "anti-man" or "anti-queer", so that users of the dataset may have access to this additional level of detail. While every instance has a final "transphobic" or "not transphobic" label, the instances that yielded further discussion include notes that users may use as they wish. As such, we contribute a perspective on data annotation that makes use of qualitative coding methods (e.g., [63]) and prioritizes careful consideration of trans people's varying opinions over arriving at a falsely cohesive understanding of what is and is not transphobic.

Additionally, considering the potential real world applications of TIDEs surfaced multiple discussions within our research team about

the practice of singling out a specific form of bigotry such as transphobia and creating a classifier specific to it. While labeling the data, we found that many posts, especially on Truth Social, were racist in addition to being transphobic, or were racist without being transphobic. The majority of these posts were anti-Black and a smaller number were anti-immigrant. These posts, as well as the general prevalence of racist rhetoric in the dataset, foreground the importance of understanding transphobia in conjunction with racism. Blackness and transness are often lumped together by racists and transphobes as markers of a progressive society that they are in ideological disagreement with (e.g. "Why most of terrorist shooters are blacks and "transgenders"?"). This complicates hate speech detection tasks like this one, where a single type of bigotry like transphobia is singled out. As anti-racist researchers, we struggled to conceive of a transphobia classifier that we would use in our daily lives that would flag transphobia and not white supremacy. Additional research on this topic could generate new, more expansive ways of understanding and classifying hate speech in NLP that can account for the relationship between different kinds of hateful content.

In conjunction, these three types of care-centered data practices (while hiring and working with data annotators, while collecting data, and while labeling data) positively shaped how we engaged with the process of transphobia detection, and allowed for a more thoughtful project for the benefit of trans communities. Our work is grounded both in a situated understanding of what makes transphobia complex and difficult to detect and a perspectivist approach to data annotation and hate speech classification [25]. While many kinds of speech are undoubtedly transphobic, care for trans people entails reckoning with the different experiences and perspectives within the trans community, which in turn shape how different trans people understand transphobia.

5.2 Public data and models for transphobia detection

By creating TIDEs, we sought to release a model that can serve as the basis for future work that addresses the issue of online transphobic rhetoric. Our goal is not only to demonstrate that different kinds of transphobia can be detected by the same model, but also to share that model with researchers and people who wish to build upon it. Some of the motivation for the creation of TIDEs came from conversations with other researchers at the intersection of trans studies and technology about how useful a tool like this would be for many different applications. We view TIDEs and systems like it as potential crucial points of intervention for users facing hate speech in day-to-day social media use. It is for this reason that we chose to finetune our own model with the input of trans and nonbinary people, instead of simply using the API of a commercial large language model. This system can be improved upon, iterated on, and used to develop tools to address transphobia detection without relying on companies or services that may become hostile against trans people in the future. We will build on this work in the future by working with trans social media users to co-design a browser extension that makes use of TIDEs.

One of the primary challenges with hate speech detection is the quick pace at which new coded language develops on social media. Our dataset contains many context-dependent posts that we identified as transphobic because of their relevance in current transphobic discourse. New transphobia campaigns will continue to emerge, and with them new language that our current dataset is not trained on. To address this challenge and continue to improve the dataset, we will implement a community labeling feature for TIDEs (similar to the one employed by Shinigami Eyes) in our future trans technology that uses TIDEs to alert users of transphobic content. This will allow users to label transphobic posts that they come across that are not adequately flagged by the model. Moving toward community-centered labeling and moderation practices is crucial for creating systems that can last beyond a few months and be used in real-world contexts. This is a form of trans-centered moderation [69], and a research space that we hope will continue to grow.

Our practices in developing and releasing the TIDEs dataset and model reflect our commitment to conducting research that centers and benefits trans social media users. We sought to create a realistic and multi-site dataset of online transphobia, and we are choosing to share it publicly to encourage future work that builds and improves upon TIDEs. This trans-centered research is, first and foremost, about creating tangible tools that can help trans people experience a less toxic internet.

5.3 Limitations

TIDEs as an intervention is focused on individual responses to transphobia in the face of a lack of protection from platforms. Our work here is not centered around proposed changes to platform policy that would significantly shift how platforms handle transphobia. As such, even if TIDEs is used in the future to create a novel tool that helps protect trans social media users from toxicity by hiding it or flagging it through an interface, the transphobic content in question would still be present on the platform.

Additionally, transphobic discourse online is a manifestation of transphobia as a structural form of oppression. While we believe that there is value to deploying technological tools for transphobia detection, we acknowledge that this is a limited response to a problem that cannot be simply addressed through technology. We strive to achieve balance between addressing the root causes of prevalent issues such as transphobia while also developing technology that works to alleviate or address the immediate effects of transphobic speech online.

6 Conclusion

TIDEs makes a contribution to trans-centered content moderation efforts by serving as a novel tool for transphobia detection which we hope to build on and share with other researchers. Throughout its development, we adopted practices of that allowed us to engage with the data and annotators through a lens of care for trans people. This contributed to a rich dataset that surfaces interesting and useful data without sacrificing annotator well-being. Ultimately, our work is a step toward developing trans technologies that can work for trans people and address transphobic content online.

Acknowledgments

We thank the ACM FAccT reviewers for their thorough engagement with and feedback on our paper. We would also like to thank

the team of annotators who made this project possible; the fellow researchers who shared their expertise and data with us, including Christina Lu, Lydia Bliss, Kylie Boyd, and Dr. Ellen Selkie; and our peers who gave us invaluable feedback throughout the writing process, including Nathan Kim and the members of the CRIT and comp.social labs at the University of Michigan. This work was funded by National Science Foundation grant #2210841, which was recently abruptly terminated in an attack against research about and for trans people. Lastly, we thank all the trans people who know that a better world is possible, and who are willing to fight for it while making cool technology along the way.

References

[1] 2018. TERF Block List (@TERFblocklist). https://x.com/terfblocklist.
[2] 2024. Aegis. https://www.aegis.report/.
[3] Maryam M. AlEmadi and Wajdi Zaghouani. 2024. Emotional Toll and Coping Strategies: Navigating the Effects of Annotating Hate Speech Data. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, Ingo Siegert and Khalid Choukri (Eds.). ELRA and ICCL, Torino, Italia, 66–72.
[4] Nazanin Andalibi and Andrea Forte. 2016. Social Computing Researchers, Vulnerability, and Peer Support.. In *Proceedings of Position Paper at the CHI 2016 Ethical Encounters in HCI: Research in Sensitive and Complex Settings Workshop.*
[5] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. 2018. Social Support, Reciprocity, and Anonymity in Responses to Sexual Abuse Disclosures on Social Media. *ACM Transactions on Computer-Human Interaction* 25, 5 (Oct. 2018), 1–35. doi:10.1145/3234942
[6] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 3906–3918. doi:10.1145/2858036.2858096
[7] Laima Augustaitis, Leland A. Merrill, Kristi E Gamarel, and Oliver L. Haimson. 2021. Online Transgender Health Information Seeking: Facilitators, Barriers, and Future Directions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3411764.3445091
[8] Serena Bassi and Greta LaFleur. 2022. Introduction: TERFs, Gender-Critical Movements, and Postfascist Feminisms. *TSQ: Transgender Studies Quarterly* 9, 3 (Aug. 2022), 311–333. doi:10.1215/23289252-9836008
[9] Ashley Belanger. 2024. X Quietly Revived Anti-Misgendering Policy That Musk Dropped Last Year. https://arstechnica.com/tech-policy/2024/02/x-quietly-revived-anti-misgendering-policy-that-musk-dropped-last-year/.
[10] Cynthia L. Bennett, Daniela K. Rosner, and Alex S. Taylor. 2020. The Care Work of Access. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3313831.3376568
[11] Sam Biddle. 2025. Leaked Meta Rules: Users Are Free to Post "Mexican Immigrants Are Trash!" Or "Trans People Are Immoral". https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/.
[12] Yuliya Cannon, Stacy Speedlin, Joe Avera, Derek Robertson, Mercedes Ingram, and Ashely Prado. 2017. Transition, Connection, Disconnection, and Social Media: Examining the Digital Lived Experiences of Transgender Individuals. *Journal of LGBT Issues in Counseling* 11, 2 (April 2017), 68–87. doi:10.1080/15538605.2017.1310006
[13] Will Carless. 2023. When Libs of TikTok Posts, Threats Increasingly Follow. *USA Today* (Nov. 2023), 01A–01A.
[14] Center for Countering Digital Hate. 2022. *Digital Hate: Social Media's Role In Amplifying Dangerous Lies About LGBTQ+ People.* Technical Report. 46 pages.
[15] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 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*. Association for Computational Linguistics, Dublin, Ireland, 369–377. doi:10.18653/v1/2022.ltedi-1.57
[16] Stevie Chancellor, Nazanin Andalibi, Lindsay Blackwell, David Nemer, and Wendy Moncur. 2019. Sensitive Research, Practice and Design in HCI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland Uk, 2019-05-02). ACM, 1–8. doi:10.1145/3290607.3299003
[17] Alexander Cho. 2018. Default Publicness: Queer Youth of Color, Social Media, and Being Outed by the Machine. *New Media & Society* 20, 9 (Sept. 2018), 3183–3200. doi:10.1177/1461444817744784
[18] Shelley L. Craig, Andrew D. Eaton, Lauren B. McInroy, Sandra A. D'Souza, Sreedevi Krishnan, Gordon A. Wells, Lloyd Twum-Siaw, and Vivian W. Y. Leung. 2020. Navigating Negativity: A Grounded Theory and Integrative Mixed Methods Investigation of How Sexual and Gender Minority Youth Cope with Negative Comments Online. *Psychology & Sexuality* 11, 3 (July 2020), 161–179. doi:10.1080/19419899.2019.1665575
[19] Shelley L. Craig, Andrew D. Eaton, Lauren B. McInroy, Sandra A. D'Souza, Sreedevi Krishnan, Gordon A. Wells, Lloyd Twum-Siaw, and Vivian W. Y. Leung. 2020. Navigating Negativity: A Grounded Theory and Integrative Mixed Methods Investigation of How Sexual and Gender Minority Youth Cope with Negative Comments Online. *Psychology & Sexuality* 11, 3 (July 2020), 161–179. doi:10.1080/19419899.2019.1665575
[20] Amber L. Cushing. 2023. PIM as a Caring: Using Ethics of Care to Explore Personal Information Management as a Caring Process. *Journal of the Association for Information Science and Technology* 74, 11 (Nov. 2023), 1282–1292. doi:10.1002/asi.24824
[21] Maria Puig De La Bellacasa. 2011. Matters of Care in Technoscience: Assembling Neglected Things. *Social Studies of Science* 41, 1 (Feb. 2011), 85–106. doi:10.1177/0306312710380301
[22] Michael Ann DeVito. 2022. How Transfeminine TikTok Creators Navigate the Algorithmic Trap of Visibility Via Folk Theorization. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–31. doi:10.1145/3555105
[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423
[24] Mark Diaz, Razvan Amironesei, Laura Weidinger, and Iason Gabriel. 2022. Accounting for Offensive Speech as a Practice of Resistance. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics, Seattle, Washington (Hybrid), 192–202. doi:10.18653/v1/2022.woah-1.18
[25] Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Mexico City, Mexico, 2279–2292. doi:10.18653/v1/2024.naacl-long.126
[26] R. Stuart Geiger. 2016. Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space. *Information, Communication & Society* 19, 6 (June 2016), 787–803. doi:10.1080/1369118X.2016.1153700
[27] Tarleton Gillespie. 2018. *Regulation of and by Platforms*. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road London EC1Y 1SP, 254–278. doi:10.4135/9781473984066.n15
[28] Tarleton Gillespie. 2020. Content Moderation, AI, and the Question of Scale. *Big Data & Society* 7, 2 (July 2020), 2053951720943234. doi:10.1177/2053951720943234
[29] GLAAD. 2023. Transgenderism: Definition, Meaning, and Origin in Anti-LGBTQ Hate. https://glaad.org/transgenderism-definition-meaning-anti-lgbt-online-hate/.
[30] GLAAD. 2024. Report – Unsafe: Meta Fails to Moderate Extreme Anti-trans Hate Across Facebook, Instagram, and Threads | GLAAD. https://glaad.org/smsi/report-meta-fails-to-moderate-extreme-anti-trans-hate-across-facebook-instagram-and-threads/.
[31] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data & Society* 7, 1 (Jan. 2020), 205395171989794. doi:10.1177/2053951719897945
[32] Oliver Haimson. 2018. Social Media as Social Transition Machinery. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–21. doi:10.1145/3274332
[33] Oliver L. Haimson. 2025. *Trans Technologies*. MIT Press.
[34] Oliver L. Haimson, Avery Dame-Griff, Elias Capello, and Zahari Richter. 2019. Tumblr Was a Trans Technology: The Meaning, Importance, History, and Future of Trans Technologies. *Feminist Media Studies* 21, 3 (2019), 345–361. doi:10.1080/14680777.2019.1678505
[35] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 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 (Oct. 2021), 1–35. doi:10.1145/3479610
[36] Oliver L. Haimson, Dykee Gorrell, Denny L. Starks, and Zu Weinger. 2020. Designing Trans Technology: Defining Challenges and Envisioning Community-Centered Solutions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. doi:10.1145/3313831.3376669
[37] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. doi:10.48550/arXiv.2006.03654 arXiv:2006.03654 [cs]

[38] Landon D. Hughes, Kacie M. Kidd, Kristi E. Gamarel, Don Operario, and Nadia Dowshen. 2021. "These Laws Will Be Devastating": Provider Perspectives on Legislation Banning Gender-Affirming Care for Transgender Adolescents. *Journal of Adolescent Health* 69, 6 (Dec. 2021), 976–982. doi:10.1016/j.jadohealth.2021.08.020

[39] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (Oct. 2019), 1–35. doi:10.1145/3338243

[40] Joel Kaplan. 2025. More Speech and Fewer Mistakes.

[41] Kiran. 2022. Shinigami Eyes. https://shinigami-eyes.github.io/.

[42] Kate Knibbs. 2025. Meta Now Lets Users Say Gay and Trans People Have 'Mental Illness'. https://www.wired.com/story/meta-immigration-gender-policies-change/.

[43] Prasanna Kumar Kumaresan, Rahul Ponnusamy, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2023. Homophobia and Transphobia Detection for Low-Resourced Languages in Social Media Comments. *Natural Language Processing Journal* 5 (Dec. 2023), 100041. doi:10.1016/j.nlp.2023.100041

[44] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs]

[45] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs]

[46] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. doi:10.48550/arXiv.1711.05101 arXiv:1711.05101 [cs]

[47] 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)*. Association for Computational Linguistics, Seattle, Washington (Hybrid), 79–91. doi:10.18653/v1/2022.woah-1.8

[48] Abulimiti Maimaitituoheti. 2022. ABLIMET @LT-EDI-ACL2022: A Roberta Based Approach for Homophobia/Transphobia Detection in Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, Dublin, Ireland, 155–160. doi:10.18653/v1/2022.ltedi-1.19

[49] Samuel Mayworm, Michael Ann DeVito, Daniel Delmonaco, Hibby Thach, and Oliver L. Haimson. 2024. Content Moderation Folk Theories and Perceptions of Platform Spirit among Marginalized Social Media Users. *ACM Transactions on Social Computing* 7, 1 (March 2024), 1–27. doi:10.1145/3632741

[50] Miles McCain and David Thiel. 2022. *Truthbrush*. https://github.com/stanfordio/truthbrush

[51] Meta. 2024. Facebook Community Standards. https://transparency.meta.com/policies/community-standards.

[52] Wendy Moncur. 2013. The Emotional Wellbeing of Researchers: Considerations for Practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 1883–1890. doi:10.1145/2470654.2466248

[53] Zein Murib. 2020. Administering Biology: How "Bathroom Bills" Criminalize and Stigmatize Trans and Gender Nonconforming People in Public Space. *Administrative Theory & Praxis* 42, 2 (April 2020), 153–171. doi:10.1080/10841806.2019.1659048

[54] Arianna Muti, Marta Marchiori Manerba, Katerina Korre, and Alberto Barrón-Cedeño. 2022. LeaningTower@LT-EDI-ACL2022: When Hope and Hate Collide. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, Dublin, Ireland, 306–311. doi:10.18653/v1/2022.ltedi-1.46

[55] Casey Newton. 2025. Inside Meta's Dehumanizing New Speech Policies for Trans People. https://www.platformer.news/meta-new-trans-guidelines-hate-speech/.

[56] Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble Modeling for Homophobia and Transphobia Detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, Dublin, Ireland, 258–264. doi:10.18653/v1/2022.ltedi-1.37

[57] Barbara Ortutay. 2023. Twitter Removes Policy against Deadnaming Transgender People. https://apnews.com/article/twitter-elon-musk-transgender-deadnaming-hateful-conduct-ae1b7285bb906e04b26ff9751ec0c2ce.

[58] Anastasia Powell, Adrian J Scott, and Nicola Henry. 2020. Digital Harassment and Abuse: Experiences of Sexuality and Gender Minority Adults. *European Journal of Criminology* 17, 2 (March 2020), 199–223. doi:10.1177/1477370818788006

[59] Siladitya Ray. 2024. Musk Says 'Cisgender' And 'Cis' Are Now 'Slurs' On Twitter. https://www.forbes.com/sites/sidityaray/2023/06/21/musk-says-cisgender-and-cis-are-now-slurs-on-twitter/.

[60] Erin Reed. 2022. "My Twitter Is Made Significantly Better by Three Things... 1) A TERF Blocklist That Is 200k+ Accounts Large... 2) Redblock Extension for Chrome... 3) Shinigami Eyes There Are Many Accounts out There Centered on Attacking Trans People, Often Posing as Members of the Community." / X. https://x.com/ErinInTheMorn/status/1566988180678758400.

[61] Erin Reed. 2024. Post-Election 2024 Anti-Trans Risk Assessment Map. https://www.erininthemorning.com/p/post-election-2024-anti-trans-risk.

[62] Henrik Skaug Sætra and Jo Ese. 2023. Shinigami Eyes and Social Media Labeling as a Technology for Self-care. In *Technology and Sustainable Development: The Promise and Pitfalls of Techno-Solutionism* (1 ed.), Henrik Skaug Sætra (Ed.). Routledge, New York.

[63] Johnny Saldaña. 2016. *The Coding Manual for Qualitative Researchers* (3. edition ed.). SAGE, Los Angeles, Calif. London New Delhi Singapore Washington DC.

[64] 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*. Association for Computational Linguistics, Florence, Italy, 1668–1678. doi:10.18653/v1/P19-1163

[65] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–27. doi:10.1145/3274424

[66] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society* 21, 7 (July 2019), 1417–1443. doi:10.1177/1461444818821316

[67] Ellen Selkie, Victoria Adkins, Ellie Masters, Anita Bajpai, and Daniel Shumer. 2020. Transgender Adolescents' Uses of Social Media for Social Support. *Journal of Adolescent Health* 66, 3 (March 2020), 275–280. doi:10.1016/j.jadohealth.2019.08.011

[68] @terf-blocklist. [n. d.]. TERF Blocklist!! https://www.tumblr.com/terf-blocklist/189407806815/terf-blocklist.

[69] Hibby Thach, Samuel Mayworm, Michaelanne Thomas, and Oliver L. Haimson. 2024. Trans-Centered Moderation: Trans Technology Creators and Centering Transness in Platform and Community Governance: Trans-centered Moderation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 326–336. doi:10.1145/3630106.3658909

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems*.

[71] Jess Weatherbed. 2023. Block Party's Anti-Harassment Service Is Leaving Twitter. https://www.theverge.com/2023/5/31/23743538/block-party-hiatus-twitter-app-anti-harassment-service-api.

A Dataset Characteristics

Table 2 provides the distribution of annotations across source domains. Table 3 lists the broad categorizations that were used in identifying subreddits (see §3.1.2).

Table 2: Dataset distribution across source and label

Source	Transphobic	Not transphobic	Total
Youtube	135	592	727
Tumblr	464	409	873
Truth Social	850	140	990
Reddit	123	766	889
Total	1572	1907	3479

Table 3: Subreddit categories

Subreddit	Community Category
actual_detrans	Trans-related (not transphobic)
ask_transgender	Trans-related (not transphobic)
gender	Trans-related (not transphobic)
GenderCynical	Trans-related (not transphobic)
genderfluid	Trans-related (not transphobic)
trans	Trans-related (not transphobic)
transgender	Trans-related (not transphobic)
Transgender_Surgeries	Trans-related (not transphobic)
TransSpace	Trans-related (not transphobic)
TransSupport	Trans-related (not transphobic)
assignedmale	Trans-related (transphobic)
Gender_Critical	Trans-related (transphobic)
GenderCritical	Trans-related (transphobic)
GenderCriticalGuys	Trans-related (transphobic)
GenderCynicalCritical	Trans-related (transphobic)
TumblrAtRest	Anti-SJW
TumblrCirclejerk	Anti-SJW
TumblrInAction	Anti-SJW
actuallesbians	LGBT-related (not transphobic)
ActualLesbiansOver25	LGBT-related (not transphobic)
Actuallylesbian	LGBT-related (transphobic)
Ask_Politics	Political discussion
Askpolitics	Political discussion
politics	Political discussion

B Keywords used for political subreddits

"gender identity" , "gender ideology" , "nonbinary" , "non-binary" , "gender dysphoria" , "gender-affirming care" , "transphobia" , "transphobic" , "pronouns" , "bathroom bills" , "sports bans" , "trans healthcare bans" , "puberty blockers" , "hormone replacement therapy" , "HRT" , "transition-related care" , "detransition" , "detransitioners" , "drag queens" , "trans kids" , "parental rights" , "grooming" , "woke ideology" , "biological sex" , "indoctrination" , "radical gender theory" , "protect our children" , "parental consent" , "J.K. Rowling" , "Matt Walsh" , "the Trevor project" , "terf" , "drag queen story"

C Codebook

Instructions given to annotators: Here are some broad categories for transphobic speech. Please keep in mind that this list is incomplete. You should use your own judgment and expertise when labeling, and you should add categories and examples here if you encounter them.

Pronouns/Names

- Misgendering
- Deadnaming
- Making light of pronouns (e.g., complaining about or mocking pronouns)
- Improper language: "transgenders," "transgendered," "transsexual," "transvestite," "cross-dresser"
- Implying that trans people are deceptive, hiding their identities, or "in disguise"

Medical

- Hateful comments about medical transition
- Referring to transition as mutilation
- Invasive questions about a person's transition (e.g., asking or speculating on whether someone has undergone specific procedures)
- Concerns about medical interventions for trans children (e.g., "save the children" rhetoric) or claims that those performing gender-affirming healthcare are harming children

Religious

- Associating trans identity with sin, hell, demons, etc.
- Religious shaming (e.g., telling trans people to find God, or that they will not go to Heaven)

TERF Rhetoric

- Terms like "TIMs," "TIFs," "TRAs," or "gyns"
- "Concerns" about transition
- "Sex is real" as code for "only sex matters," and therefore trans identity is not real or valid
- Calls to abolish gender when the intent is to erase trans people

Dogwhistles

- "What is a woman?"
- "41%," a transphobic joke referencing a statistic on the trans suicide rate

General Hateful Conduct

- Wishing harm upon trans people
- Associating trans people with violence, crime, or deviance
- Statements like "There is no such thing as transgender"

Political

- Statements like "Transgender isn't in the Bill of Right" used to justify that trans people "don't exist"
- Concerns about a liberal "transgender agenda," "transgenderism," or "gender ideology"

- Positive reactions to anti-trans legislation or news
- Making assumptions about the gender identities of politicians or engaging in "transvestigation"
- Conspiracy theories about trans people, such as:
  - Claims that "the FBI threatened a nurse who accused the nation's largest children's hospital of illegally billing taxpayers for transgender medicine"
  - Mention of a so-called "trans program" (e.g., unfounded theories like imagining a "Charles Xavier school for trans people")

News Media

- If there is a headline with no context, evaluate the headline itself:
  - Is the headline transphobic or perpetuating a transphobic narrative? For example, if the headline is about a crime committed by a trans person, does it explicitly draw attention to the trans person's identity in a way that implies that they committed that crime *because* of their identity? If so, label it as "transphobic."
  - Otherwise, (for example, if the headline is about a trans athlete going to the Olympics and it does not use transphobic terms) mark it as "not transphobic."

D Logistic regression features

The highest weighted features (positive and negative) from the logistic regression model are given, along with model coefficients, in Table 4.

Table 4: Top 10 strongest tokens associated with the transphobic (left) and not transphobic (right) classes.

Feature	Coefficient	Feature	Coefficient
transgender	2.91	good	-1.08
attracted	1.22	choice	-1.18
cult	1.21	stupid	-1.17
gb	1.15	ban	-1.14
follow	1.12	materialism	-1.06
rad	1.08	desire	-0.98
fem	1.06	respect	-0.96
month	1.05	surgeries	-0.96
gy	1.04	rules	-0.95
disgusting	1.03	let	-0.94

E Model performance

E.1 Misclassifications

Table 5 provides four representative examples of model misclassifications.

Table 5: Examples of model misclassifications

Label	Pred.	Post content
Transphobic	Not	"she is very mature, whether she has a male voice or not i see her as a woman because she was assigned that gender at birth, hope shes doing better now"
Transphobic	Not	"To just swap at will would be pretty cool, you've got to admit. Anyway I've completely lost track of everything, does transgender mean anything anymore? If they are going to receive adequate help then there needs to be some agreement on what is dysphoria/ transgender and what severity is worth any treatment"
Not	Transphobic	"Fun fact, the Nazis targeted transgender folks as well. So MAGAs hateful crusade against them is 100% on point for being the fascists they are."
Not	Transphobic	"You are transgender, right?"

E.2 Cross-source testing

We also experimented with iteratively holding out one source domain during training, given the importance of the different forms of transphobia expressed on each platform. Hyperparameter tuning in this case involved 10 trials for each model, rather than 25. Results are given below in Table 6.

Table 6: Performance of the best model trained on only the other three sources for each source.

Source	Accuracy	Macro F1	Weighted F1	F1
YouTube	0.71	0.63	0.69	0.46
Tumblr	0.53	0.53	0.52	0.48
Truth Social	0.43	0.40	0.50	0.54
Reddit	0.27	0.21	0.12	0.43