Beyond the Binary: Analysing Transphobic Hate and Harassment Online

Anna Talas
University of Cambridge
at2008@cam.ac.uk

Summer Leigh
University of Cambridge
sdrfhl2@cantab.ac.uk

Alice Hutchings
University of Cambridge
ah793@cam.ac.uk

Abstract

Online communities provide support and help to individuals transitioning gender. However, this point of transition also increases vulnerability, coupled with increased exposure to online harms. In this research, we analyse a popular hate and harassment site known for targeting minority groups, including transgender people. We analyse 17 million posts dating back to 2012 to gain insights into the types of information collected about targets. We find users commonly link to social media sites such as Twitter/X and meticulously archive links related to their targets. We scrape over 150,000 relevant links posted to Twitter/X and their archived versions and analyse the profiles and posts. We find targets often tweet about harassment, popculture, and queer and gender-related discussions. We develop and evaluate classifiers to detect calls for harassment, doxxing, mention of transgender individuals, and toxic/abusive speech within the forum posts. The results of our classifiers show that forum posts about transgender individuals are significantly more likely to contain other harmful content.

1 Introduction

Life transitions can be tricky, and many people turn to online resources to help navigate this change (Haimson et al., 2019; Zhang et al., 2022). While online communities may provide help and support (Geeng et al., 2022), there is also the potential for users to experience unintended consequences. Engaging online also exposes people to online harms, making them even more vulnerable during difficult times. One major life change that some people face is gender transition (Haimson, 2017; Thomas et al., 2021). To design better tools to help users navigate life transitions, we need to first understand the risks being faced.

In this research, we explore online hate and harassment directed towards those who have undergone gender transition. We use data from a hate and harassment site included in the ExtremeBB corpus (Vu et al., 2023b) of posts scraped from extremist forums. The forum was chosen due to the coordinated harassment of minorities, including transgender people. Users on this site commonly link (include a URL in their post) to the social media profiles (usually Twitter/X[1]) of targets, often using an archive service. We find instances of cross-posted links to social media and archive sites and scrape additional data from the URLs. To minimise the quantity of hateful content we view and make it feasible to analyse such a large amount of data, we develop classifiers to predict which posts relate to calls for harassment, transgender people, contain doxxing,[2] and contain abusive/toxic content.

Archiving services take snapshots of a website, preserving it exactly as it was when the snapshot was taken. This can preserve posts that are later deleted, or accounts that are made private. We collect the archived versions of Twitter/X and compare these with more recent data. By evaluating the changes between archived and current versions, we can identify what changes have taken place, such as the profile being made private, profiles being suspended, tweets being deleted, or profiles being deleted. We use these three main data sources to address the following research questions:

RQ1 Is there evidence of displacement away from Twitter/X towards other sites, such as Mastodon, Blue Sky, or Threads?

RQ2 How do victims respond to harassment? Do they make their profiles private or delete posts? Do they change handles?

RQ3 Is harassment moderated?

[1]Twitter was rebranded as X in July 2023. We use Twitter to refer to the platform before this change, X to refer to it after, and Twitter/X to refer to both before and after.

[2]Doxxing refers to compiling and publishing personal information about a person.

RQ4 How do attackers select targets?

RQ5 Are transgender individuals more likely to be doxxed or harassed compared to non-transgender individuals?

We observe recent changes to online platforms, to understand how moderation activities and the spaces where abuse is occurring change over time. We analyse the distribution of links on the forum and archive sites to assess displacement effects as users leave websites, or moderation efforts change. This provides insight into how targets are doxxed, what information is being spread, and where.

We make several contributions. We train four binary classifiers to detect calls for harassment, doxxing, transgender targets, and abusive or toxic speech in the forum posts. Overall, we find an upward trend in the presence of harmful content over time. We scrape over a decade worth of archived links from archival sites and extra content from those linking to Twitter/X to understand the user behaviour and the types of links shared. We analyse the contents of the archived tweets. We analyse the changes in the profiles over time. We also find that most profiles belong to relatively small creators.

We provide an overview of related work in §2. §3 describes our dataset and methods. We evaluate classifiers in §4, and in §5 provide our findings. We discuss our findings in §6, including potential future research directions. Our conclusion are found in §7, followed by the limitations in §8. Tables and Figures are provided in the Appendix.

2 Related Work

2.1 Online risks faced by the LGBTQ+ community

We build on research into risks faced by the transgender and wider LGBTQ+ community online. Thomas et al. (2021) taxonomise hate and harassment directed towards the LGBTQ+ community, arguing the problem is increasing over time. While they do not focus specifically on transgender individuals, they do acknowledge attackers may 'deadname' (use the former name of a transgender individual) targets. A related harm is intentional 'misgendering' (labelling a transgender person with a gender that does not match their gender identity).

According to the typology, the attacks we see in this research would be classified as toxic content, content leakage, and overloading. Toxic content includes bullying, trolling, and intentional provocation. Content leakage, which includes doxxing, refers to the spread of sensitive private information with the intention of embarrassing, threatening, or intimidating the target. We see overloading through the calls for harassment, where social media accounts of targets are posted. This can be considered coordinated trolling activity.

Similar to our research, research by Haimson (2017) focuses on the online experiences of the transgender community. Haimson et al. (2016) use survey methods to explore the social complexities involved in managing online information disclosure and identities when transitioning gender. Haimson et al. (2015) explore how online communities can provide those going through major life transitions with support and friendship, and help mitigate stress. Likewise, Geeng et al. (2022) find that queer participants value online community support, but must navigate the online risks that come with this.

Locatelli et al. (2023) analyse online homophobia across seven different languages on Twitter. They find homotransphobia is a global problem, while its expression is highly dependent on cultural context. Tanni et al. (2024) analyse direct messages on Instagram and concluded that LGBTQ+ youth experience significantly more high-risk online interactions and report worse mental health. They also highlight the importance of creating supportive online environments that tailor to LGBTQ+ youth.

Vu et al. (2023a) show that suppressing harmful online forums is hard even when efforts are combined. They find collective industry attempts to take down a hate and harassment forum were ultimately unsuccessful. Furthermore, while loosely connected users left the platform, many others joined who were much more toxic and active.

2.2 Detecting and classifying online harassment and doxxing

We also focus on the detection and classification of online harassment and doxxing. Franz and Thatcher (2023) analyse the victim perspective after being doxxed and how it influences behaviour. Aliapoulios et al. (2021) classify calls to harassment and doxxing, both of which can lead to harassment of targets online and in physical spaces. Snyder et al. (2017) analyse the frequency of doxxing on sites like 4chan.org and 8ch.net. They find anti-abuse efforts by social networks help reduce the frequency of targets responding to harassment by restricting or closing accounts. Arora et al. (2020)

develop classifiers to automatically detect harassment aimed at women journalists on Twitter. To our knowledge, there is no previous research specifically analysing transphobic hate and harassment in extremist forums on a large scale.

Inspired by Dias Oliva et al. (2021) and Talas and Hutchings (2023), we evaluate Jigsaw and Google's Perspective API, used for online moderation on platforms such as Reddit and major media outlets. The Perspective API defines toxicity as a "rude, disrespectful, or unreasonable comment that is likely to make someone leave the conversation". Talas and Hutchings (2023) discover the classifier is not reliable in detecting toxicity in music lyrics shared in underground forums. Their exploration reveals the classifier only takes the first 501 characters as input. Therefore, for longer posts, performance drops considerably. Dias Oliva et al. (2021) find the Perspective API is biased against LGBTQ+ content creators, classifying Twitter posts made by drag queens as more toxic than those of white nationalists and labelled tweets using words like "gay" and "lesbian" as highly toxic even if they contained positive content. We explore how well the classifier works on our forum dataset, which contains very long posts (averaging over 1,000 characters), many targeting the LGBTQ+ community.

3 Research Methods

In this section we explain our research methods. We start with an overview of the ethical concerns we considered when designing the project. For our data collection phase we collected the links posted in the extremist forum and scraped further data from X and archival sites. An overview of the data processing pipeline can be seen in Figure 2.

3.1 Ethical considerations

We obtained ethics approval from our department's ethics committee. The ExtremeBB dataset and the data from Twitter/X and archive websites are collected from publicly available websites using web scrapers, and informed consent is not requested from users. Under the Ethics Statement of the British Society of Criminology (2015), informed consent is not required for research into online communities where the data is publicly available and the research outputs focus on collective rather than individual behaviour. Where example posts are provided, they are paraphrased to reduce the likelihood the author is identified or attributed. We minimise providing examples of toxic/transphobic content to limit the exposure of these views.

Recent changes to X have made API access prohibitively expensive for academic research. We discussed this with other academics affected by these changes, who advised requesting data or permission from X is unlikely to be successful, and the community is moving towards scraping relevant content for research purposes, where there is a clear social benefit that outweighs potential risks.

We considered the possibility of a legal case to be made against us. We found this was unlikely to be successful for several reasons. First, relevant case law from the US has ruled that web scraping from public sites does not violate the Computer Fraud and Abuse Act. Second, there is an exception to UK copyright law which allows researchers to make copies of any copyrighted material for the purpose of computational analysis if they already have lawful access to the material. The UK Government has asserted that 'Publishers and content providers will be able to apply reasonable measures to maintain their network security or stability, but these measures should not prevent or unreasonably restrict researcher's ability to text and data mine. Contract terms that stop researchers making copies to carry out text and data mining will be unenforceable' (Intellectual Property Office, 2021). We also reduce harm to the platforms by throttling our scraping to avoid overloading their services.

To further reduce the likelihood of harm to users of these platforms, we do not publish identifying information. We do some analysis of online content linked to on the hate and harassment platform. Some of these tweets have since been deleted or made private by the user, but we collect older archived versions. To respect the content creators' wishes for these tweets to no longer be available, we discarded deleted or privated content before running our analyses.

Another consideration is that working with hate and harassment data poses risks for researchers. To mitigate these risks, researchers participating in this project met regularly to discuss and offer support. Team members were aware of the counselling services (provided at no cost) they could turn to if required. An additional risk to researchers is the possibility of reprisals (Doerfler et al., 2021). Therefore, we do not name the hate and harassment platform to reduce the likelihood that its members will target us. We followed best practice in conducting risky research (Marwick et al., 2016) and com-

municated with the department and university's communications teams about the research and potential for harassment to be directed to us.

3.2 ExtremeBB

The original dataset we use is a subset of ExtremeBB (Vu et al., 2023b). This dataset is available for academic research through datasharing agreements.[3] The dataset consists of posts scraped from various extremist forums going as far back as far as 2001. We analyse one English-language forum, but avoid providing details which may identify which one (see §3.1). The forum largely focuses on targeting different individuals and minority groups, i.e., the LGBTQ+ community and neurodivergent people, with boards and threads dedicated to specific people or groups of individuals. The forum also contains boards dedicated to everyday topics such as music, gaming, and other hobbies. This forum has been associated in the media with harassment of members of the LGBTQ+ community.

The forum contains more than 17M posts dating back over a decade and has become increasingly more popular over time (see Figure 3). The decline during 2022 and 2023 is due to forum disruptions, which made it inaccessible for some periods. The data relating to links posted to the forums is from 2013 to September 2023, as scraping and parsing this volume of links with a custom scraper took a significant amount of time. The classifier results include all posts in the forum up to April 2025.

To understand how targets are doxxed, we focus on the types of links shared in the posts. First, we filter the dataset to extract links contained in the content of the posts. We analyse the links, finding users commonly link to social media sites such as YouTube, Twitter, Reddit and Facebook (see Table 1). A significant number of links direct to different 'archive' domains, such as archive.md, arhive.ph etc., all of which (except for web.archive.org) lead to the same archival site.

3.3 Archive sites

One of the most commonly linked sites from the extremist forum is an archive site similar to the Wayback Machine.[4] This site is mostly used to archive content (often targets' personal accounts) and ensure accessibility even if the original content is taken down or the shared social media post is deleted. This provides the unique opportunity to analyse the actions taken by victims and online platforms over time by comparing the archived version with the current state. The archive has used various domains and mirror sites over time. Despite showing up as separate domains they all redirect to the same content. There are over 1.17 million links to this archive site posted on the forum. Many are duplicated, leaving 382,114 unique links.

We visit each of the links and use a customised scraping tool to recover the original archived URL. A total of 14,854 unique domains are archived, with the majority only appearing a few times. Only the top 1,200 domains have more than 10 links. Table 2 shows the most commonly linked domains.

Table 1: Most common domains linked to outside the forum

Domain	No. of occurences
youtube[.]com	999,301
twitter[.]com	603,901
archive[.]md	583,227
archive[.]ph	265,323
imgur[.]com	189,178
wikipedia[.]org	156,619
archive[.]fo	145,339
archive[.]vn	140,002
reddit[.]com	95,657
youtu[.]be	93,783
facebook[.]com	54,998
web[.]archive[.]org	54,551
mobile[.]twitter[.]com	33,212
instagram[.]com	30,637

Table 2: Most common original domains in the archived sites (counting unique links only)

Domain	No. of occurences
twitter[.]com	160,629
reddit[.]com	15,807
tumblr[.]com	15,767
facebook[.]com	8,014
youtube[.]com	6,767
deviantart[.]com	4,254

3.4 Scraping Twitter/X and archive data

The most commonly archived domain was Twitter (Table 2). Therefore, we analyse these links further, scraping the content from both the archived and current versions of the site. There are three types of Twitter, those that point at specific tweets (92.9%),

[3] https://www.cambridgecybercrime.uk/process.html
[4] https://web.archive.org/

Twitter profiles (5.8%), and miscellaneous links that do not fit either category (such as to the front page or specific search queries, 1.3%). We disregard the third category as they are a small proportion of the total, and there is no clear way to compare their content to the current state.

To scrape data from these links (both the archive and the current version) a custom scraper was built. For URL pointing to a tweet or retweet, the scraper collects the unique tweet ID. This allows tweets to be associated with a profile through username changes, so we can capture when usernames are changed. Due to the nature of the posts that contain these links, many point to tweets that are no longer accessible or to profiles that have been deleted, suspended, or made private. For all accessible tweets, the scraper collected the following information: current username and display name of the poster, text contents of the tweet (disregarding any media content), date the tweet was created, number of retweets, likes, replies, views and bookmarks. Not all information was available for every tweet. If the tweet was a retweet, the same information was gathered about the original tweet.

If the URL points directly to a profile, the scraper records whether a profile is accessible, private, suspended, or does not exist. Where possible, the following information was also collected: current username and display name of the poster, if the profile is verified, profile creation date, number of tweets posted, number of followers and followings, and bio(graphy). We originally intended to collect follower lists from each profile to create a social graph for identifying if targets are part of similar social circles. However, due to recent changes in X's profile display, only a small subset of followers and following are displayed on individual profiles and we are unable to collect the full follower lists.

We also intended to collect the same data for each profile and tweet from the archive sites. This was complicated by changes in Twitter/X over the years. The archive capture the site in its original state, going back as far as 2012. This means that we were unable to obtain data about views and bookmarks for older archived tweets, as they were not metrics offered at the time of archival. We also discovered some archive links (<1%) had not correctly archived the posts, which also hindered our ability to get all the information initially planned.

To analyse the content of the tweets archived and shared on the platform we perform topic analysis on the tweets available on X at the time of scraping (see §5.2). While we count how many tweets have been deleted or made private, for ethical reasons we remove the content of these tweets from our subsequent analyses (§3.1). For the remaining tweets, we use standard NLP pre-processing steps, including removing non-alphabetical characters such as punctuation, and stop-words. We use lemmatisation to get the root forms of the words. Finally, we use BERTopic (Grootendorst, 2022) for unsupervised topic modelling to identify groups of words that commonly appear in the same context.

4 Classifier development

4.1 Manual annotations

To evaluate and develop classifiers for large-scale analysis of the data, three annotators manually label a subset of the forum posts. We label all posts as positive or negative for the four categories (calls for harassment, doxxing, transgender target and toxic or abusive content) outlined in the annotation guidelines provided in Table 4. The annotation categories were inspired by previous research which also aimed to detect doxxing and calls for harassment (Aliapoulios et al., 2021), including an annotation for whether the posts reference transgender individuals. The four categories were all viewed separately and were not mutually exclusive, the annotations could contain any possible combinations of the four categories. Annotators included the authors and an additional team member with domain expertise. The annotators received the posts in text form, including metadata to show embedded links, but were asked not to open any of them. Annotators were aware of the potentially distressing content of the posts beforehand and the option to discontinue annotating if they wished to.

The initial set of 300 threads was randomly selected from all threads on the forum. We annotated the first five posts in each thread (or fewer if there were less than five posts), totalling 1,491 posts. We annotate the first posts from a thread as those tend to provide the most context, and annotating multiple posts together helps to understand the context. We find the first posts tend to contain the initial call for harassment and information about the target. We met to discuss disagreements, which were mainly decided by majority vote. Two annotators also completed a second round of annotations to improve classifier performance by annotating another 100 posts with the lowest confidence in the classifier results out of the forum dataset.

Table 5 shows a breakdown of annotation results. We evaluate the agreement between annotators by using Fleiss' κ coefficient for annotations with three annotators and Cohen's κ coefficient for the second batch of annotations with two annotators. Landis and Koch (1977) propose that κ greater than 0.2 indicates fair agreement, 0.4 moderate agreement, 0.6 substantial agreement, and above 0.8 almost perfect agreement.

4.2 Evaluation of existing classifiers

We identify existing classifiers created to detect doxxing and calls for harassment by Aliapoulios et al. (2021). We evaluate these classifiers and find poor performance on data from the hate and harassment forum. This may be because their classifiers are trained on much shorter texts and different lexicons than those found in the forum dataset.

We also evaluate the accuracy of the Perspective API (Google Jigsaw, 2017). The API measures the toxicity score of text posts. The API is free to use, and given a text in any of the 18 languages currently supported, returns a score between 0 and 1 that represents the likelihood the comment is considered toxic by the reader. While the API offers classification for more specific categories, we only outline our evaluation of the "toxicity" classifier, as we achieved the best performance with this category.

We compare our manual annotations for toxicity or abusive content ($N = 1,591$) with the Perspective API. Some results returned an error, leaving us with $N = 1,565$ observations. As the Perspective API toxicity scores are not normally distributed (Figure 6), a Mann-Whitney U test is used to examine the relationship between the two. We find a significant difference, with those annotated as toxic more likely to have higher Perspective API toxicity scores ($U = 481615.5, p < .001$). However, we note that the Mann-Whitney U test is often used when sample sizes are small. When sample sizes increase, it becomes particularly sensitive to small changes. Therefore, we look at Figures 6 and 5. Although the median Perspective API toxicity score is higher for those we annotated as toxic, there is still a considerable proportion of posts we consider toxic or abusive that had a very low toxicity score.

4.3 Classifier development

We apply common pre-processing steps to convert the natural language inputs into a suitable format, such as word embedding, TF-IDF, and TF-IDF vectorisation. Machine learning risks ignoring minority classes in imbalanced datasets. As our dataset is imbalanced for most categories, we apply SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002), which synthesises new examples of minority classes using already vectorised inputs. We also experiment with adjusting the loss function to account for the imbalance.

We test and evaluate suitable models, narrowing our focus down to three models used for similar research working with NLP data from underground forums, e.g. Zhou et al. (2023); Man et al. (2023). These include XGBoost (eXtreme Gradient Boost) and BERT (Bidirectional Encoder Representations from Transformers). We also include ModernBERT, a newer version of BERT with a larger context window. To fine-tune the classifiers we also change the confidence interval thresholds, as we find most false positives and false negatives receive low confidence scores. For classifiers with limited context windows we also attempt chunking (breaking the input down into smaller parts).

We attempt using local LLMs of various sizes and one-shot learning to classify the posts. The LLMs tested (Mistral and Gemma) do not provide a significant improvement in classifier accuracy, and are significantly slower (classifying around 15.5 posts per minute for one category). Classifying 17 million posts at this rate is resource intensive, taking around 761 days. There are ways to speed up this classifying process, but they are prohibitively expensive and would go against our ethics agreement to not upload the data, which includes personal information, to cloud providers.

4.4 Evaluation of our classifiers

Due to the imbalanced dataset, accuracy scores for the classifiers are high even when they had high false positive and false negative rate. We did a second round of annotations on a smaller subset of posts that received low confidence scores to improve precision and recall. The second round of annotations, despite being a much smaller amount, reliably improved the metrics of the classifiers.

The evaluation metrics are in Table 6. The XGBoost models outperform BERT for all four categories, while ModernBERT outperforms XGBoost in "calls to harassment" and "toxic/abusive content". This may be because BERT has an input limit of 512 tokens, which is much smaller than most posts we classify. ModernBERT has a much higher limit of 8,192 tokens. Chunking inputs did not significantly improve performance. XGBoost

may outperform both BERT versions as pre-trained language models can struggle to adapt to slang and specialised vocabulary, which are frequent in the forum posts, especially when referring to minorities, including transgender people. We subsequently use the best performing models for each category to classify the entire forum dataset (see §5.3).

5 Analysis and Results

5.1 Analysis of archive and Twitter/X data

Figure 4 shows the distribution of archived links over time, including archived Twitter/X links. Links to Twitter/X are posted frequently for most of the forum's lifespan, peaking at the beginning of 2020 before slowly declining. In **RQ1** we ask if there is displacement towards other competing platforms due to recent changes in X. We do not see any significant displacement in links being shared on the forum. We find 667 links point to Mastodon, 49 links point to Threads, and nine to Blue Sky. This includes archive links where the original site was the social media platform. Recent declines in archived Twitter/X links may be due to changes to X which prevent archiving, as the site now requires users to log in to view posts. The forum was also inaccessible for some parts of 2022 and 2023.

By analysing the archival links posted we find links for 30,231 Twitter/X accounts. The average number of links archived per account is five. The most links posted for one account is 2,871. We analyse the 149,189 links posted for tweets and 9,350 links directing to profiles. In the archived version of Twitter we are able to scrape the content for most tweets, with <1% being unavailable. While scraping X we find almost half of the posts (67,284, making up 45.1%) are no longer available due to the post being deleted or the profile being private. While we cannot make claims about causality, this provides some indication that victims make their profiles private or delete posts at some point after being linked to on the forum (**RQ2**).

Only in a small number of links (205) had a username change when comparing the archive and current versions. As Twitter/X redirects to the correct tweet using the post ID when the user handle is changed (allowing us to identify changes), it is unlikely to be an effective way to mitigate potential harassment once the post has been linked to. We find some indications of moderation (**RQ3**), with 14,758 tweets no longer available due to account suspension. A further 4,060 tweets are unavailable as the account was made private. For linked profiles, 745 are suspended and 856 are made private.

5.2 Target selection

While it is difficult to accurately assess how targets are selected (**RQ4**), analysing common characteristics of linked Twitter profiles and contents of tweets provides some insights. We analyse the follower counts of the archived Twitter profiles. Follower counts range from 0 to over 140 million. The majority of archived profiles have relatively few followers: half have fewer than 328 followers, and 75% have less than 1,580. Therefore, these accounts likely belong to small creators or regular users rather than celebrities and popular figures.

We analyse the keywords found on the biographies of the profiles linked to from the forum (see Table 11). As well as terms relating to occupations and hobbies, these frequently include queer/trans and activism-related language and preferred pronouns. We also find several emojis, including the "No one under eighteen" sign (often signalling sexual content, 1,366 occurrences), the rainbow pride flag (982), and the transgender pride flag (684).

We use BERTopic to analyse the topics of linked tweets that had not been subsequently deleted. We exclude tweets only containing images or videos. We find most tweets are covered by eight topics (Table 12). The most frequent topic relates to harassment, with associated keywords including 'assault', 'abuse', 'evidence' and 'doxxed'. The tweets in this topic often reference the forum we analyse, referring to it as a doxxing or stalking site. Personal topics include Storytelling, Queer/Transgender, Body-image and Mental Health. Other common topics are Politics/News and Popculture. Almost 5% of tweets were about fundraising and donations. Only 17% of tweets do not fit into the above topics.

5.3 Analysis of posts using automatically classified data

Table 3 shows the results of our classifications across the entire forum. We find 0.9% of posts are classified as a call for harassment, 0.4% are predicted to contain doxxing, and 4.0% mention a transgender target. Over one-third of posts are predicted to have toxic or abusive content. This overall low proportion is because most calls for harassment and doxxing are contained in the threads' first post, which attract many replies. Table 10 shows that in the first post of threads, the proportion of positive results is much higher. Of the first posts, 8.0%

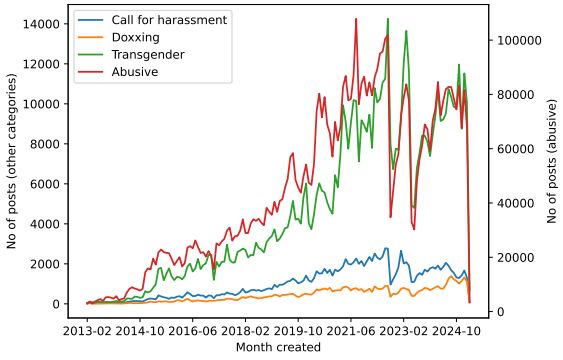

Figure 1: All positive results of classifiers over time

contain a call for harassment, almost nine times higher than the proportion of all posts. Similarly, 5.2% contain doxxing, and 5.4% mention a transgender individual. The percentage of toxic content is slightly lower (34.9% of first posts).

Figure 1 shows the distribution of positively classified calls for harassment, doxxing, transgender target and toxic/abusive content over time, using a different scale for the 'abusive' category, given the high volume of posts. This indicates an increase in these types of posts, except for drops from forum disruption in 2022/2023.

The comparison of calls for harassment, doxxing, and toxic/abusive content against posts mentioning transgender targets shows that transgender targets are significantly more likely to be negatively affected (**RQ5**). The difference is greatest for posts containing toxic/abusive speech ($\chi^2(1, N = 17,090,519) = 565,677.02, p < .001$, Table 9). This may be because forum users often use slurs and specific offensive slang to refer to transgender individuals. The difference for calls for harassment ($\chi^2(1, N = 17,090,519) = 76,555.21, p < .001$, Table 8), is greater than for doxxing ($\chi^2(1, N = 17,090,519) = 19,299.81, p < .001$, Table 7).

5.4 Additional observations

During our analysis, we note several additional observations. First, some users' activities border on cybercrime. While cybercrime forums are more focused on profit-motivated activities (Hughes and Hutchings, 2023), this forum uses malicious ways to gather information about specific targets. For example, forum users use OSINT tools, and there is some evidence that they explore exploits aimed at gaining access to targets' accounts. Forum members frequently seek advice on which tools are best for acquiring information about targets.

The personal information posted commonly includes the person's full name, address, date of birth, phone number, and email address. Similar personal information about targets' family members is often posted. Transgender individuals are frequently referred to by their deadname (the name used before transitioning). Deliberate misgendering (not using the person's preferred pronouns) is encouraged.

Users teach others how best to archive and share their findings, e.g.: *"The process: 1) Use a converter to download a video/audio submission on YouTube and the like. I like [app]. There are various editing tools for audio and video. 2) Upload the files to one or more publicly accessible repositories- the more the better. Examples include [archive sites]. 3) Link to the original source and other backups in the file descriptions. Use proper tagging and naming. 4) Promote all copies in appropriate places such as [forum name], drama groups, etc."*

Archiving information about targets, so content survives if hidden/deleted, is encouraged. We discover the reason multiple domains point to the one archive site is because, unlike the Wayback Machine, this site does not comply with takedown requests: *"Good work on the archive. Don't use Wayback Machine for tweets because they might counteract archiving certain twitter accounts, as demonstrated by [USERNAME]"*

6 Discussion

We explore the harassment and doxxing posted on hate forums with a focus on how it affects transgender people. The most popular social media site linked to is Twitter/X. This may be because on Twitter/X users often post using accounts linked to their real world identity (unlike Reddit, where users tend to be anonymous), and posts are by default public (unlike Facebook, where you often need to friend someone before seeing their whole profile and post history). However, personal information shared about targets, such as full names, birthdays, and names of family members, is likely collected from these sites using publicly available information.

Targets on Twitter/X mostly respond by self-censorship (deleting tweets), but do not tend to private or delete their accounts. We find little evidence of users changing their username on the platform, perhaps because Twitter/X still redirect links to posts to the new handle. While we find evidence of Twitter/X moderating content posted on the platform, it is difficult to evaluate the type of

Table 3: Results of the classifier on all posts in the forum

Category	Yes	No
Calls for harassment	147,509 (0.9%)	16,943,010 (99.1%)
Doxxing	62,689 (0.4%)	17,027,922 (99.6%)
Transgender target	672,311 (4.0%)	16,418,200 (96.0%)
Toxic/abusive content	6,390,892 (37.4%)	10,699,619 (62.6%)

content they moderate, as we do not know which post(s) caused the suspension, and if they belong to the subset of archived posts. Moderation may also incorrectly flag accounts, as Haimson et al. (2021) find minority groups experience more frequent content removal despite following site guidelines.

Existing classifiers trained to detect doxxing and calls for harassment do not perform well across domains, hence we train our own classifiers. Existing classifiers are trained on relatively short posts, while many forum posts are essay-length. The hateful language also contains a lot of slang, particularly words meant to be insulting or derogatory to minorities. This makes building generalisable classifiers for use on many different platforms difficult.

Compared to non-transgender related content, forum posts relating to transgender people are significantly more likely to including doxxing and toxic content. The proportion of both is much higher in the first posts of the threads, which tend to be longer-form dossiers on the targets. While transgender individuals are more likely to be targeted, there is little evidence of users' displacement from Twitter/X to other platforms. This may point to the importance of online spaces to minority groups. Lucero (2017) suggest for minority groups such as LGBTQ+ youth, offline contexts are often hostile and unsafe, therefore online communities are an important place for self-expression and exploration.

However, increased exposure to online spaces can also increase the likelihood of experiencing online harms. Therefore, there is a need to help the members of queer support communities participate in a safe way. The solution should not be to avoid online spaces altogether. Rather than discouraging the LGBTQ+ community from being open on social media, this should be encouraged and normalised, and moderation efforts should recognise the types of harassment they are subjected to.

7 Conclusion

We analyse a large hate and harassment forum. Twitter/X is by far the most commonly archived site, accounting for almost half of archived links. We do not find evidence of displacement to alternative social media sites such as Mastodon, Threads, or BlueSky. Comparing the archived and current states of Twitter/X links allows us to analyse changes over time. Almost half of posts are no longer available as they are deleted or the accounts are private or suspended. Tweets contain many different topics, including Harrassment, Storytelling, Popculture and Queer/Transgender discussions.

Profile biographies include common themes, such as hobbies or occupations, gender and sexuality-related terminology, preferred pronouns, and pride flags. We manually annotate posts to train four classifiers to detect calls for harassment, doxxing, transgender targets, and toxic/abusive content in the forum posts. The Perspective API performs poorly on toxic forum data. We use our classifiers to automatically label all 17+ million forum posts. Posts mentioning transgender people are significantly more likely to contain calls for harassment, doxxing, and toxic or abusive speech, and these categories are most likely to be found in the first post of the thread.

8 Limitations

This research has attempted to overcome the significant difficulties associated with this challenging area of research. However, a number of limitations remain. First, we only analyse data from one hate and harassment forum, one social media website, and one archive site. Future research could include analysing how the trends might differ on other platforms.

There is an issue with detecting doxxing through the methods used, as it may be obfuscated in multiple ways. Doxxing can appear in the form of images or links to other websites containing personal information that cannot be detected by analysing the posts' textual data. Users occasionally add self-authored PDF or text file attachments, which we did not collect, but potentially contain doxxing.

The NLP methods we use to automate the clas-

sification of posts are not without their limitations, and our classifiers are not 100% accurate. This is further complicated by the subjective difficulty of defining what falls under categories like toxic speech or calls for harassment. Some posts are, therefore, likely to be misclassified. Similarly, the classification of the topics covered by the tweets is complicated by some of them containing mostly (or only) images or videos, which we didn't scrape, and our classification is based purely on the text.

The classification of specific types of posts or motivations in these kinds of forums is made more difficult by the use of specialised slang and argot, which pre-trained language models are very unlikely to have in their training dataset. Moreover, many of the posts are much longer than context windows offered by models such as BERT, increasing the difficulty even more.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/T517847/1 (for AT, SL, and AH) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 949127) (for AH).

References

Max Aliapoulios, Kejsi Take, Prashanth Ramakrishna, Daniel Borkan, Beth Goldberg, Jeffrey Sorensen, Anna Turner, Rachel Greenstadt, Tobias Lauinger, and Damon McCoy. 2021. A large-scale characterization of online incitements to harassment across platforms. In *Proceedings of the 21st ACM Internet Measurement Conference*, IMC '21, page 621–638, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3487552.3487852.

Ishaan Arora, Julia Guo, Sarah Ita Levitan, Susan McGregor, and Julia Hirschberg. 2020. A novel methodology for developing automatic harassment classifiers for Twitter. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 7–15, Online. Association for Computational Linguistics. Https://aclanthology.org/2020.alw-1.2.

British Society of Criminology. 2015. Statement of ethics. https://www.britsoccrim.org/documents/BSCEthics2015.pdf.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357. https://doi.org/10.1613/jair.953.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, 25:700–732. https://link.springer.com/article/10.1007/s12119-020-09790-w.

Periwinkle Doerfler, Andrea Forte, Emiliano De Cristofaro, Gianluca Stringhini, Jeremy Blackburn, and Damon McCoy. 2021. I'm a professor, which isn't usually a dangerous job: Internet-facilitated harassment and its impact on researchers. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–32. https://doi.org/10.1145/3476082.

Anjuli Franz and Jason Bennett Thatcher. 2023. Doxing and doxees: A qualitative analysis of victim experiences and responses. In *European Conference on Information Systems (ECIS)*. https://aisel.aisnet.org/ecis2023_rp/397/.

Christine Geeng, Mike Harris, Elissa Redmiles, and Franziska Roesner. 2022. "Like lesbians walking the perimeter": Experiences of US LGBTQ+ folks with online security, safety, and privacy advice. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 305–322. https://www.usenix.org/system/files/sec22-geeng.pdf.

Google Jigsaw. 2017. Perspective API. https://www.perspectiveapi.com/.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Oliver L Haimson. 2017. The social complexities of transgender identity disclosure on social network sites. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 280–285. https://dl.acm.org/doi/abs/10.1145/3027063.3027136.

Oliver L Haimson, Jed R Brubaker, Lynn Dombrowski, and Gillian R Hayes. 2015. Disclosure, stress, and support during gender transition on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1176–1190. https://dl.acm.org/doi/abs/10.1145/2675133.2675152.

Oliver L Haimson, Jed R Brubaker, Lynn Dombrowski, and Gillian R Hayes. 2016. Digital footprints and changing networks during online identity transitions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2895–2907. https://dl.acm.org/doi/abs/10.1145/2858036.2858136.

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. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2). https://doi.org/10.1145/3479610.

Oliver L Haimson, Bryan Semaan, Brianna Dym, Joey Chiao-Yin Hsiao, Daniel Herron, and Wendy Moncur. 2019. Life transitions and social technologies: Research and design for times of life change. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, pages 480–486. https://dl.acm.org/doi/abs/10.1145/3311957.3359431.

Jack Hughes and Alice Hutchings. 2023. Digital drift and the evolution of a large cybercrime forum. In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 183–193. IEEE. https://ieeexplore.ieee.org/abstract/document/10190639.

Intellectual Property Office. 2021. Guidance: Exceptions to copyright. https://www.gov.uk/guidance/exceptions-to-copyright.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174. https://www.jstor.org/stable/2529310.

Davide Locatelli, Greta Damo, and Debora Nozza. 2023. A cross-lingual study of homotransphobia on Twitter. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 16–24. Association for Computational Linguistics. https://aclanthology.org/2023.c3nlp-1.3.

Leanna Lucero. 2017. Safe spaces in online places: Social media and LGBTQ youth. *Multicultural Education Review*, 9(2):117–128. https://doi.org/10.1080/2005615X.2017.1313482.

Jessica Man, Gilberto Atondo Siu, and Alice Hutchings. 2023. Autism disclosures and cybercrime discourse on a large underground forum. In *2023 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–14. IEEE. https://ieeexplore.ieee.org/abstract/document/10485504/.

Alice E. Marwick, Lindsay Blackwell, and Katherine Lo. 2016. Best practices for conducting risky research and protecting yourself from online harassment. https://datasociety.net/wp-content/uploads/2016/10/Best_Practices_for_Conducting_Risky_Research-Oct-2016.pdf.

Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. 2017. Fifteen minutes of unwanted fame: detecting and characterizing doxing. In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, page 432–444, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3131365.3131385.

Anna Talas and Alice Hutchings. 2023. Hacker's paradise: Analysing music in a cybercrime forum. In *2023 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–14. IEEE. https://ieeexplore.ieee.org/abstract/document/10485503/.

Tangila Islam Tanni, Mamtaj Akter, Joshua Anderson, Mary Jean Amon, and Pamela J. Wisniewski. 2024. Examining the unique online risk experiences and mental health outcomes of LGBTQ+ versus heterosexual youth. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642509.

Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. SoK: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267. IEEE. https://ieeexplore.ieee.org/abstract/document/9519435/.

Anh V Vu, Alice Hutchings, and Ross Anderson. 2023a. No easy way out: the effectiveness of deplatforming an extremist forum to suppress hate and harassment. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 7–7. IEEE Computer Society. https://www.computer.org/csdl/proceedings-article/sp/2024/313000a007/1RjE9LYWfTy.

Anh V Vu, Lydia Wilson, Yi Ting Chua, Ilia Shumailov, and Ross Anderson. 2023b. ExtremeBB: A database for large-scale research into online hate, harassment, the manosphere and extremism. In *The 61st Annual Meeting Of The Association For Computational Linguistics*. https://virtual2023.aclweb.org/paper_ACL_33.html.

Ben Zefeng Zhang, Tianxiao Liu, Shanley Corvite, Nazanin Andalibi, and Oliver L Haimson. 2022. Separate online networks during life transitions: Support, identity, and challenges in social media and online communities. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–30. https://dl.acm.org/doi/abs/10.1145/3555559.

Linda Zhou, Andrew Caines, Ildiko Pete, and Alice Hutchings. 2023. Automated hate speech detection and span extraction in underground hacking and extremist forums. *Natural Language Engineering*, 29(5):1247–1274. https://doi.org/10.1017/S1351324922000262.

A Appendix

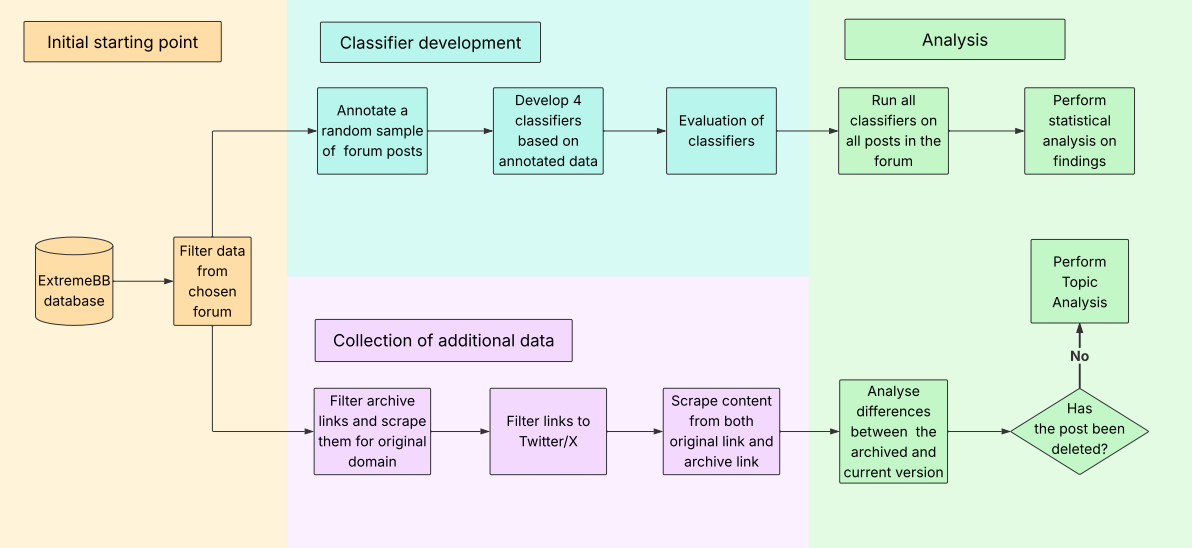

Figure 2: Data processing pipeline

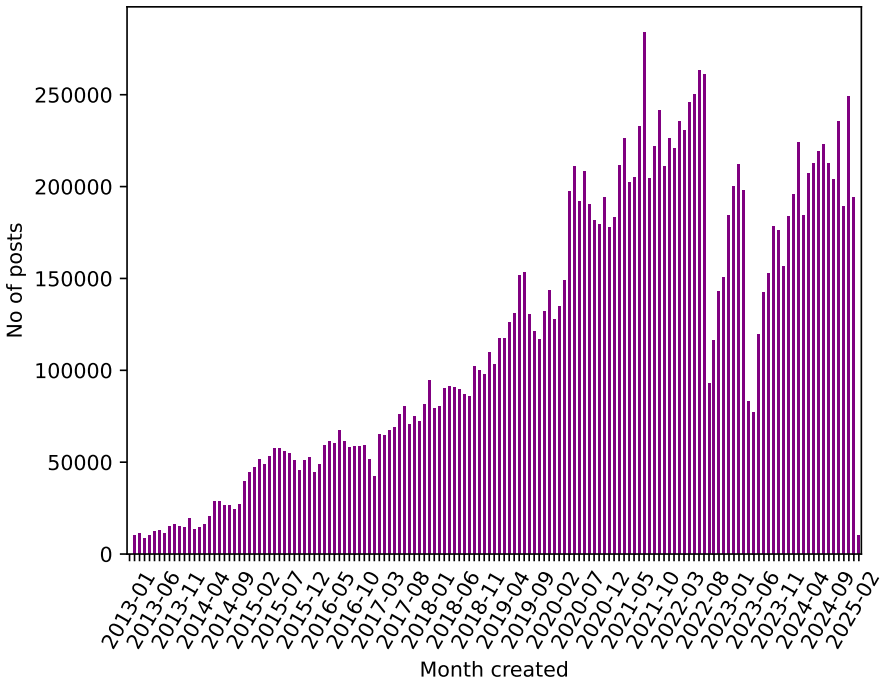

Figure 3: Number of posts scraped from forum by month

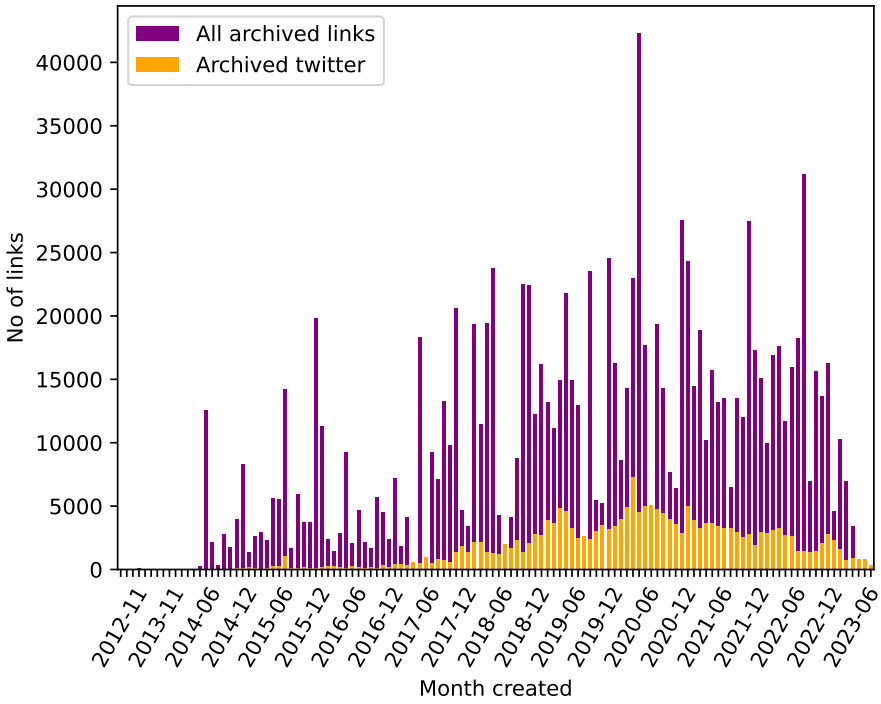

Figure 4: Number of archived links posted over time

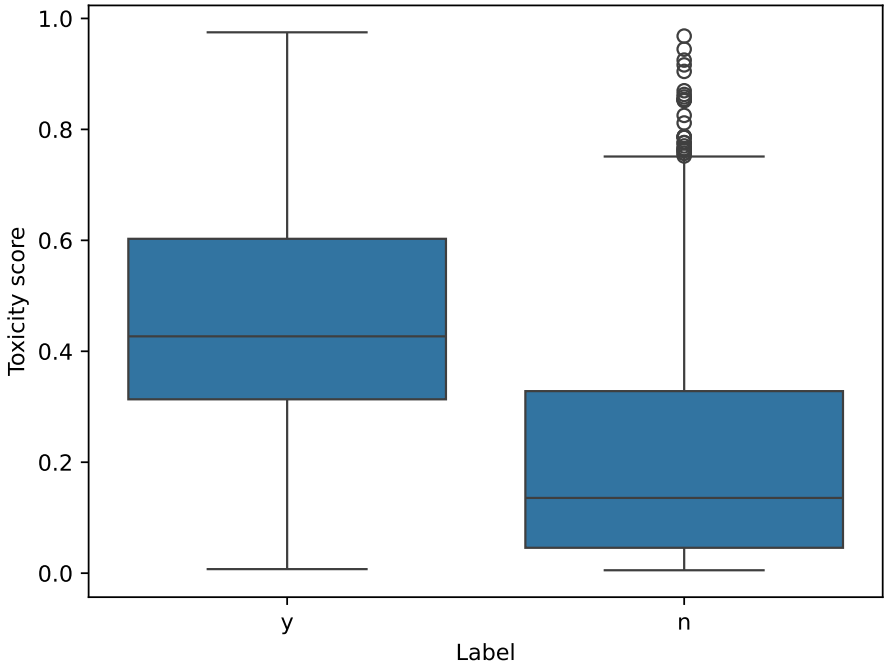

Figure 5: Boxplots comparing Perspective API toxicity classifier results (0.0-1.0) with manual annotations (y/n)

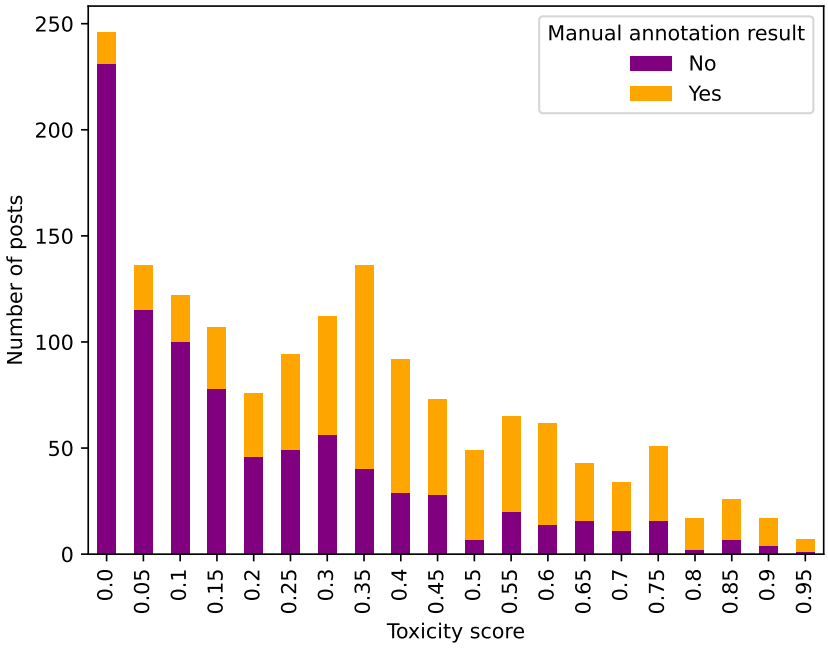

Figure 6: Comparison of Perspective API toxicity classifier results (0.0-1.0) and manual annotations (y/n)

Table 4: Annotation guidelines, each category is annotated separately with true or false

Category	**Description**	**Anonymised example**
Call for harassment	Does the post attempt to mobilise others to collaborate in harassing the target?	*"[WEBSITE] has so much that needs to be archived, I am not even done looking into this idiot, this is just to get started."*
Doxxing	Does the post contain doxxing? (personal information about the target i.e. address or phone number)	*"Dox courtesy of [USERNAME]: [ADDRESS] [LEGAL NAME]"*
Transgender target	Is the post about a transgender individual?	*"[NAME] is another autistic and dumb trans freeloading off of taxes"*
Toxic or abusive content	Would you describe the post as abusive or toxic?	*"Can you imagine losing a beauty contest to this fat and ugly freak?"*

Table 5: Results of manual annotation of posts, including annotation agreements using κ- coefficient

	1st round				**2nd round**				**Total**	
	Yes	**No**	κ	**Agreement**	**Yes**	**No**	κ	**Agreement**	**Yes**	**No**
Call for harassment	90	1,401	0.668	Substantial	99	1	1.000	Almost perfect	189	1,402
Doxxing	45	1,446	0.807	Almost perfect	35	65	0.889	Almost perfect	80	1,511
Transgender target	152	1,339	0.801	Almost perfect	38	62	0.918	Almost perfect	190	1,401
Abusive or toxic speech	609	882	0.670	Substantial	99	1	0.492	Moderate	708	883

Table 6: Comparison of XGBoost and BERT

Calls to harassment	Accuracy	Precision	Recall	F-score
BERT	0.884	0.809	0.421	0.554
XGBoost	0.956	0.795	0.838	0.815
ModernBERT	0.992	0.833	0.882	0.857
Doxxing	Accuracy	Precision	Recall	F-score
BERT	0.988	0.500	0.429	0.461
XGBoost	0.975	0.600	0.818	0.692
ModernBERT	0.989	0.571	0.571	0.571
Transgender target	Accuracy	Precision	Recall	F-score
BERT	0.936	0.882	0.468	0.612
XGBoost	0.934	0.719	0.586	0.645
ModernBERT	0.920	0.546	0.660	0.598
Toxic/Abusive content	Accuracy	Precision	Recall	F-score
BERT	0.695	0.985	0.266	0.419
XGBoost	0.685	0.657	0.591	0.622
ModernBERT	0.782	0.723	0.739	0.734

Table 7: Contingency table for transgender target and calls for harassment (expected frequencies in parentheses)

Trans target	Call for harassment No	Call for harassment Yes	Total
No	16,297,069	121,137	16,418,206
	(16,276,499.8)	(141,706.2)	
	Std. Res 10.49	Std. Res -10.48	
Yes	645,941	26,372	672,313
	(666,510.2)	(5,802.7)	
	Std. Res -10.48	Std. Res 10.48	
	16,943,010	147,509	17,090,519

$\chi^2(1, N=17{,}090{,}519)=76{,}555.21$
$p<.001$

Table 8: Contingency table for transgender target and doxxing (expected frequencies in parentheses)

Trans target	Doxxing No	Doxxing Yes	Total
No	16,364,821	53,385	16,418,206
	(16,358,076.3)	(60,129.7)	
	Std. Res 5.27	Std. Res -5.27	
Yes	663,106	9,207	672,313
	(669,850.7)	(2,462.2)	
	Std. Res -5.27	Std. Res 5.27	
	17,027,927	62,592	17,090,519

$\chi^2(1, N=17{,}090{,}519)=19{,}299.81$
$p<.001$

Table 9: Contingency table for transgender target and toxic/abusive content (expected frequencies in parentheses)

Trans target	Abusive/toxic content		Total
	No	Yes	
No	10,571,176	5,847,030	16,418,206
	(10,278,716.4)	(6,139,489.6)	
	Std. Res 28.50	Std. Res -28.50	
Yes	128,446	543,867	672,313
	(420,905.6)	(251,407.4)	
	Std. Res -28.50	Std. Res 28.50	
	10,699,622	6,390,897	17,090,519

$\chi^2(1, N=17{,}090{,}519)=565{,}677.02$
$p<.001$

Table 10: Results of the classifier on the first post of each thread

Category	**Yes**	**No**
Calls for harassment	6,057 (8.0%)	69,457 (92.0%)
Doxxing	3,904 (5.2%)	71,610 (94.8%)
Transgender target	4,090 (5.4%)	71,424 (94.58%)
Toxic/abusive content	26,324 (34.9%)	49,190 (65.1%)

Table 11: Most commonly used keywords from profile biographies

Keyword	**Frequency**
'she/her'	2819
'artist'	2112
'he/him'	2093
'game'	1868
'writer'	1809
'love'	1783
'account'	1782
'make'	1574
'trans'	1566
'like'	1515
'art'	1506

Table 12: Topics present in tweets along with most commonly associated keywords (percentage of posts in brackets)

Topic	**Most commonly associated keywords**	**Example tweet**
Storytelling (15.6%)	stop, friend, read, believe, feel, support, talk	*"Thanks for letting me talk about this stuff online. I know it hurts my numbers, but it's one of the few ways I can process things. Really appreciate you being my support structure."*
Queer/Transgender (10.6%)	transphobic, transgender, lgbt, lgbtq, trans, homophobic, queer, misogyny, surgery, hrt, dysphoria	*'If you think kids shouldn't be taught that LGBTQ people exist, you're fueling hate and ignorance. That's inexcusable."*
Harrassment (25.7%)	doxxing, doxxed, harassment, banned, blocked, lawsuit, evidence, accused, investigation, assault, abuse, story, stalking	*"If you believe doxxing and harassment are ever 'justified,' you're no better than the people committing real harm. I don't care if it's 'just against people you find gross' — you're still enabling abuse, and that's messed up."*
Politics/News (9.9%)	coronavirus, deaths, pandemic, biden, trump, voters, putin, ukraine, russia, antifa, fascism, leftism	*"Each passing day sees more young Ukrainians and Russians dying over minor territorial shifts, with borders that remain largely unchanged. Their lives are worth far more than this."*
Bodyimage (1.9%)	obesity, obese, fatphobia, overweight, fat, weightneutral, diet, weigh, stigma	*"Stop assuming that when fat people face certain health issues more often than thin people, their bodies are inherently to blame. Fat individuals endure constant stigma, repeated weight fluctuations, and discrimination throughout the healthcare system — all of which take a serious toll on their health."*
Popculture (10.8%)	comicsgate, comics, manga, comic, anime, streaming, twitch, furry, animeconvention, gamer, gaming, youtube	*"I'm putting together my 'most disappointing' games of 2019 list and wow, there's no shortage lol. But I'm curious: what was your biggest gaming letdown? Might end up on my list!"*
Mental Health (3.9%)	therapy, meds, pills, antidepressants, therapist, autism, ssris, medication, therapists, aspergers	*"My doctor visit for anxiety went well! We're giving Wellbutrin a shot. I've tried other antidepressants before, but I've heard this one might come with fewer annoying side effects. Fingers crossed!"*
Fundraising (4.6%)	bank, fund, fundraiser, donations, money, deposits, donation, gofundme, paypal	*"Shoutout to the right-wing chuds for getting mad about my donation links, your outrage brought in $16! I don't usually get much from that tweet, so honestly, thanks for the boost!"*