A Large-Scale Characterization of Online Incitements to Harassment Across Platforms

Max Aliapoulios
New York University
New York, NY, USA

Kejsi Take
New York University
New York, NY, USA

Prashanth Ramakrishna
New York University
New York, NY, USA

Daniel Borkan
Jigsaw
New York, NY, USA

Beth Goldberg
Jigsaw
New York, NY, USA

Jeffrey Sorensen
Jigsaw
New York, NY, USA

Anna Turner
Jigsaw
New York, NY, USA

Rachel Greenstadt
New York University
New York, NY, USA

Tobias Lauinger
New York University
New York, NY, USA

Damon McCoy
New York University
New York, NY, USA

ABSTRACT

Attack strategies used by online harassers have evolved over time to inflict increasing harm to their targets. In addition to scaling harassment through incitement and coordination, online communities that commonly engage in harassment are likely a source of “innovation” for harassment attack strategies. We use the incitements or calls to harassment posted by members of these communities as a lens through which to holistically measure and understand this ecosystem. We create a filtering pipeline to discover 14,679 incitements to harassment within four large-scale data sets of messages and posts that span multiple platforms.

Our approach studies the coordination itself, detecting inciting language, rather than individual attack types, to understand a broad range of harassment strategies. In particular, this approach allows us to create a taxonomy of attack strategies. We use this taxonomy to categorize the preferred approaches of coordinated attackers and the proportion of incitements for various types of harassment on different platforms. We find that over 50% of the incitements to harassment included calls to report the target to authorities or their respective platforms. Finally, we provide suggestions for actions and future research that could be performed by researchers, platforms, authorities, and anti-harassment groups.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Computing methodologies** → *Natural language processing*; • **Networks** → **Social media networks**.

KEYWORDS

Online social harm, online coordinated harassment, cyberbullying, doxing, empirical measurement.

ACM Reference Format:
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 *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3487552.3487852

1 INTRODUCTION

Online harassment can have severe consequences including economic loss [16], reputation damage, harms to mental health [17], and in extreme cases has resulted in death [22]. The scale and intensity of online harassment can be amplified when the aggressor incites others to also inflict harm against the target, what we term *calls to harassment*. One prominent coordinated harassment incitement strategy is doxing (or doxxing), which has been defined as the intentional release of personal information seeking to punish, intimidate, threaten or humiliate individuals [15]. In 2021, a survey by the Economist Intelligence Unit found that 55% of women globally have experienced doxing, either firsthand or as witnesses [45].

Attack strategies used by online harassers have evolved over time to inflict increasing harm to their targets. In addition to scaling harassment through incitement and coordination, online communities that commonly engage in harassment are likely a source of “innovation” for harassment attack strategies. We use the calls to harassment posted by members of these communities as a lens through which to measure and understand this ecosystem.

Other work has detected specialized calls for harassment in the form of detecting doxes, which are both a form of harassment and an implicit call to harassment [37]. However, this specialized lens omits a large part of the incitement to harassment landscape. This paper contributes the first measurement study aimed at illuminating the broader ecosystem of these calls to harassment.

IMC '21, November 2–4, 2021, Virtual Event

Our approach studies the coordination itself, detecting inciting language, rather than individual attack types, to understand a broad range of harassment strategies. In particular, this approach allows us to categorize the preferred approaches of coordinated attackers and the proportion of calls for various types of harassment on different platforms.

This research makes several contributions. First, we develop a filtering pipeline to detect coordinated harassment incitements and doxing instances across five platforms (Discord, boards, Gab, pastes, Telegram). We make the models used in the pipeline publicly available, along with our automated analysis tools. In addition to boards, Chat, Gab, and pastes, prior research has indicated that certain ideological blogs are often sources of doxes and calls to harassment [32]. We create a taxonomy of these attacks, and discuss how different communities gravitate to different attack strategies.

As an additional contribution, we use this holistic, empirical data set to provide an analysis of sources of calls to harassment, which we synthesize into a taxonomy of attack types, and provide suggestions for actions and future research that could be performed by researchers, platforms, authorities, and anti-harassment groups. For example, while reporting systems are one of the key approaches used to counter harassment and abuse, our analysis demonstrates that their manipulation is a key goal of coordinated harassers. Over 50% of the calls to harassment we annotated (3,206 instances) included calls to report the targets to the platforms where they hold accounts (24% or 1,496 instances) and other public or private entities (such as law enforcement or employers). To better understand how calls to harassment develop, we conducted an analysis of threads including calls to harassment on image boards. We find that calls to harassment rarely appear in the first post of the thread (only 3.7% of the time) and are fairly evenly distributed over the length of the thread. This analysis demonstrates that analyses that only focus on the first post in threads will miss the majority of coordinated harassment. We also studied the co-occurrence of calls to harassment and doxes within threads. In our data sets, we show that only 95 posts are detected as both doxes and calls to harassment, out of a total of 14,679 posts detected by both filtering pipelines and validated by manual analysis. Additionally, we find that 8.53% of calls to harassment contain a dox and 17.85% of doxing threads contain a call to harassment. We use our analysis of calls to harassment to categorize and contextualize how the PII and other information included in a dox might increase the risk of harm to targets.

Our research illustrates that developing tools that can detect and analyze calls to harassment in addition to actual harassment is important not just for social media platforms, but also to empower individuals so that they may better manage these incidents.

2 BACKGROUND & RELATED WORK

We define *calls to harassment* as "when an individual attempts to mobilize others online to collaborate to conduct online harassment" and define online harassment as "when an aggressor specifically targets another person or group online to inflict emotional harm." The key portion of this definition is that our study is not looking for online harassment on its own, but rather for an effort by an individual or group to rally support from others to conduct online harassment. We define *doxing* as when "a third party posts, broadcasts or publishes personal information about an individual without their consent and with the intention to do harm." We note that while doxes could be seen as an implicit call to harassment (i.e., publishing the target's personal information so that others can use it to contact and harass the target), in this study, we do not treat a dox as a call to harassment unless the dox explicitly contains mobilizing language. We develop two separate classifiers to detect doxes and calls to harassment, respectively.

Interview-based research has studied the nature, experiences and consequences of online harassment, particularly on students [8, 9, 26] and female journalists [2, 7]. Chen *et al.* found that in a sample of 2,120 Hong Kong secondary school students, girls were harassment targets more often than boys, and that there were significant associations between disclosure of Personally Identifiable Information (PII) or personal audio-video materials, and emotional and psychological distress [9]. In a study by the Economist, 35% of women surveyed reported mental health issues as a result of online threats [45]. Researchers have also studied public shaming as an emerging type of doxing [19, 24].

There have been several quantitative studies that examined specific call to harassment attack strategies posted to 4chan, such as raiding (i.e., inciting spamming of) YouTube channels [21, 29] and Zoombombing (i.e., disrupting conference calls) [27]. Snyder *et al.* trained a classifier for detecting doxes and performed a quantitative study of doxes posted to pastebin.com, 4chan, and 8ch.net [37]. We expand on this prior work by creating a filtering pipeline that can detect a broader set of calls to harassment and doxes across platforms. These classifiers enable us to perform a more holistic analysis of the call to harassment ecosystem.

Thomas *et al.* created a taxonomy of hate and harassment attack strategies based on a review of prior studies [42]. The categories of online harassment include toxic content, content leakage, overloading, false reporting, impersonation, surveillance, and lockout and control. We use this taxonomy as a starting point to create a taxonomy of call to harassment strategies. We refine and improve this taxonomy using our empirical data on calls to harassment.

There has also been a large effort to design abusive content detection systems [5, 6, 12, 29, 37]. Online social networks have included anti-harassment warnings in their community guidelines [43, 44] and have also implemented automated detection mechanisms [14]. We focus on studying calls to harassment rather than abusive content to holistically understand how those seeking to incite harassment operate.

3 ETHICAL CONSIDERATIONS

A better understanding of calls to harassment provides clear benefits in prevention and mitigation of the harms caused by harassment. Since the data at hand is sensitive, we took particular privacy precautions, and our protocol was approved by New York University's Institutional Review Board (IRB). We limited our study to secondary analysis of published content, without contacting any doxers or targets, and we release only aggregate trends.

Annotations of our data sets were carried out by a mix of internal and external annotators to address issues of annotation quality and scale. We took steps to minimize unnecessary exposure of annotators to personal data of targets of harassment, or to potentially

Data set	Posts/Messages	Min Date	Max Date
Boards	405,943,342	2001-06-14	2020-08-01
Blogs	115,052	1999-04-23	2020-08-14
Chat	70,273,973	2015-09-21	2020-08-01
Gab	50,165,961	2016-08-10	2020-08-01
Pastes	32,555,682	2008-03-22	2020-08-01

Table 1: Raw data sets that we use to detect instances of doxing and calls to harassment.

distressing or unlawful content. All annotators were given only the text contained in posts (no images or metadata), and were instructed not to open URLs or use the data from the posts in any way (e.g., no web searches). Annotators were briefed on the necessity to maintain confidentiality. They were also briefed beforehand on the type of content to be annotated, its potentially distressing nature, and the option to pause or discontinue the annotation task at any time. Internal annotations were conducted by student coauthors, who participated on their own initiative. They received recommendations to take frequent breaks, and their advisor inquired regularly about their well-being. External annotations were contracted from a professional labelling service for machine learning training data. The service is contractually obligated to confidentiality, and workers were required to sign a non-disclosure agreement during their on-boarding process. Workers' participation in the task was voluntary and subject to successful completion of training and test questions with synthetic examples. The service stated that their workers earn at least minimum wage.

We have contacted several interested platforms to aid with the removal of calls to harassment and doxes, and we will continue this work going forward. Lastly, we will open-source the classifiers discussed in this analysis to help online platforms better detect calls to harassment and doxing. We will not provide PII or actual training data for these classifiers. We acknowledge that determined doxers could use these open-sourced classifiers to reverse-engineer better doxing strategies to evade dox detectors on platforms, and learn which sites are most amenable to calls to harassment or doxing. However, we believe that the content moderation opportunities that these classifiers provide to platforms and websites are more significant than this risk, as is the free availability of baseline classifiers for academics interested in conducting future research.

4 DATA

A third-party threat intelligence company collected and provided to us the data sets that we analyzed for this study. The raw data set that was shared with us only contained text (no images), and was stored on secure servers in accordance with our IRB's protocols for handling sensitive data that might contain PII. The company's web crawlers collected data from a range of different platform types where coordinated harassment activity takes place, including blogs, boards, chat services (i.e., Discord and Telegram), paste sites, and micro blogs (i.e., Gab). The third-party web crawlers collect raw HTML from websites (i.e., pastes) and API responses from applications such as chats. Depending on the data source, the crawlers utilized accounts necessary for data access. Table 1 provides a summary of all the data sets that we analyzed. We deliberately avoid publishing complete lists of individual channels, servers, domains or sub-forums in order not to advertise the online locations where doxing and calls to harassment occur; we will provide them privately to other researchers.

Blogs. The "blogs" data includes posts from ideologically motivated websites that have been involved in high-profile harassment incidents. The blogs studied represent a variety of ideologies from fascist to anti-fascist, and contain long-form posts that did not require an account to access. Sometimes sites are associated with a specific organization, and other times with a specific individual or group of individuals. One example site is "The Daily Stormer," a Neo-Nazi forum. Table 1 includes summary statistics from the three different blog sites we studied.

Boards. The "boards" data spans 43 different domains and includes sites known for hate speech and coordinated harassment such as "4chan" and "8kun" [18]. Each of these domains is considered an "imageboard," or "board" for short. Imageboards are forums that focus on image posting. Users reply to one another in groups of posts, considered a thread, which typically begin with a single image. Most board sites are somewhat ephemeral because they archive old threads in a way that makes it difficult to browse historical data. In addition, these sites are pseudo-anonymous and users are made unidentifiable across posts if anonymity is desired.

Chat. The "chat" data includes selected messages from Discord and Telegram. We utilized channels/servers from these data sources that subject matter experts had hand-labeled as being related to online troll groups, white supremacist organizations, or general online hate and harassment communities. This curated set of channels was provided by our threat intelligence partners. The data was only collected from channels that did not require an invitation, and not from direct or otherwise private messages.

Discord is an instant messaging and voice-over-IP application originally designed for online gaming communities. The New York Times reports that Discord has a concentration of Neo-Nazis and other communities associated with online hate and harassment [35].

Telegram is a messaging application with direct message and group message features. It was included in this study due to its use as a key communication platform by extremists and those engaged in online harassment. Prior research shows that online extremists and terrorist organizations actively recruit and conduct harassment campaigns through Telegram [4, 34]. The data collection from Telegram spans 2,916 different channels with 126,432 users.

Gab. We also included Gab in our study to understand how a micro-blogging online social network is used to conduct coordinated online harassment. Gab has been extensively studied in the past in the context of hate speech, and prior work demonstrates the large amounts of coordinated harassment that occur on the platform [52].

Pastes. We included 41 different domains within the "pastes" category. The pastes sites are online content hosting services where users can post and store text. These text storage sites contain long-form text posts that are often only accessible with a direct link to the post. Posters are usually anonymous but can post with a username. In benign settings, they are often used for sharing code snippets. However, as Snyder *et al.* [37] showed, doxing content often finds its way to pastes sites, presumably because of their long-form and pseudo-anonymous functionality. Paste sites sometimes also contain posts with large database dumps in a technical format (e.g., SQL), but we do not include this type of post in the doxing category. These sites commonly provide rate-limited APIs that enable collection of all new posts, but old posts are only accessible with the random post ID number. Therefore, crawlers for these data sources have been running for several years to actively collect data, and are assumed to be incomplete.

5 METHODOLOGY

Our goal was to create a filtering pipeline that produces call to harassment and dox data sets precise enough for manual annotation. Initially, we focused on recall to capture a more diverse set of calls to harassment and doxes. We used active learning to reduce the data sets to a highly precise set. We then used these data sets as a basis for our empirical study. We built two separate pipelines for identifying calls to harassment and doxes, which are shown in Figure 1. Even though doxes are also implied calls for further harassment in addition to being a type of harassment, we treat calls to harassment and doxes separately because doxes are easier to detect with specialized methods.

5.1 Initial Annotations

For each task of creating a pipeline, we built an initial set of positive and negative labeled documents.

Doxes. For the doxing task, we started with the set of annotations that we received from the authors of prior work [37]. They consisted of 10,387 negative and 799 positive examples, all collected from `pastebin.com` between 7/20/2016 and 8/31/2016. We also included 428 positive examples from the "Doxbin" site, which (as the name implies) only hosts doxes. In total, we had 1,227 positive and 10,387 negative examples for the initial doxing task.

Calls to harassment. There were no existing sets of call to harassment annotations from prior studies, so we searched for a set of keywords and phrases that were likely to be used in calls to harassment, similar to the strategy in the prior doxing study [37]. Our queries were a combination of keywords indicating mobilizing language, a subclause for ingroup versus outgroup language, and a clause for specific text related to calls to harassment, such as "doxxing," "raiding," and "reporting." An example query can be found in Figure 4 in the appendix. Initially, we ran our queries only on the 4chan, 8chan and 8kun data sets, since we expected that they would have the highest concentration of calls to harassment. Three of the authors annotated the resulting posts to create a set of 424 negative and 947 positive examples of calls to harassment.

5.2 Classifiers

We used the annotated doxes and calls to harassment described above to train NLP classifiers to filter the posts from the raw data set into a data set small and precise enough for manual analysis.

We used a computationally faster and more compact implementation of neural transfer-learning model BERT [46] called distilBERT [36] for both of our classifiers. There are two main steps when implementing transformer classifiers, `pre-training` and `fine-tuning`. During the pre-training step, we provided the classifier with a large corpus of unlabeled text and pre-trained it to predict masked (missing) tokens. The pre-trained classifier is then adaptable to different Natural Language Processing (NLP) tasks during the second fine-tuning phase. Next, the fine-tuning process uses the weights learned during pre-training and trains over them using labeled training data. In our case, we had labeled training data for each classification task.

We tokenized the documents into sequences using both punctuation splitting and the WordPiece [50] sub-word segmentation algorithm. DistilBERT has a max-sequence length of 512 characters and so we employed a method of random spanning without overlap to reduce the size of longer documents. This method of dealing with text longer than the max-length ensured that we had spans of text from all areas of the input document. The challenge here is a balance between models with a small memory footprint that can process large amounts of data and ensuring the input text contains enough information about the sample.

We also experimented with other methods of handling longer text, such as taking spans of text from the beginning and end of the document, taking overlapping spans of text during the splitting phase, and selecting spans of random length. We found that taking random spans with no overlap resulted in the best performance for our sequence classification tasks. We chose not to use a more computationally complex classifier that can handle documents longer than 512 characters due to computational limitations.

5.3 Crowdsourced Annotations

To improve the performance of our classifiers, we obtained additional annotations using a third-party crowdsourced data annotation service. Prior studies of online hate, harassment, and toxicity have also used crowdsourced annotations to expand corpora of annotated data [3, 10, 23, 39, 41, 47, 51].

We developed two annotation guides detailing the tasks. Annotators were allowed to participate in the study if they received a score of 90% or above on an initial set of 10 randomly selected posts from our set of initial annotations, and annotators were retested every tenth document. We removed annotators from the task if their score fell below 85%. An example question and template for our crowdsourcing tasks can be found in Figure 3 in the appendix.

Table 2 shows a distribution of the over 100,000 documents we had annotated, which included over 79,000 for the doxing task and over 25,000 for the call to harassment task. At least two annotators annotated each document. Annotators disagreed on 3.94% and 18.66% of the raw documents for the doxes and calls to harassment annotation tasks, respectively. We calculated Cohen's Kappa over the two initial annotations for each task; the scores for the doxing and call to harassment tasks were considered moderate agreement (0.519) and fair agreement (0.350), respectively. These agreement scores are an indication of the difficulty of the tasks for non-domain experts. When the two annotators did not agree, the document was annotated by a third annotator to break the tie. The final annotations were a result of a collaborative process and several iterations with the third-party annotation service. We established a spot-checking process with the third-party service, which involved reviewing random samples of annotations in order to keep track of poor annotator performance. In addition, one

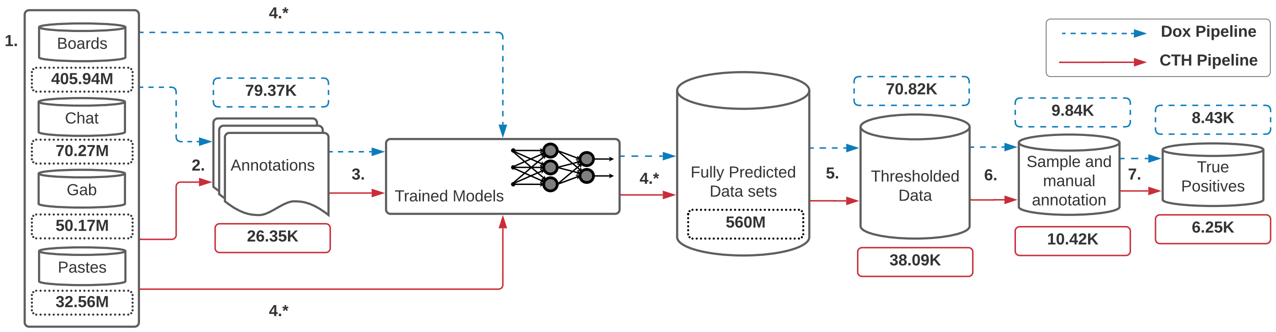


Figure 1: Call to Harassment (CTH) and doxing analysis pipeline with the number of documents in each step.

	Doxxing		Call to Harassment	
	Positive	Negative	Positive	Negative
Boards	163	797	967	8,751
Chat	536	19,943	401	8,314
Gab	216	35,166	356	7,564
Paste	2,955	19,598	-	-
Total	**3,870**	**75,504**	**1,724**	**24,629**

Table 2: Full annotated training data set sizes per task. This data was used to train a classifier for each task. The call to harassment task does not apply to the pastes data set because we sought to study this type of online harassment in collaborative online spaces; pastes do not enable this interactivity.

Classifier	Text length	Label	F1	Precision	Recall
Doxing	512	Dox	0.76	0.77	0.75
		No Dox	0.99	0.99	0.99
		Weighted Avg.	0.98	0.98	0.98
		Macro Avg.	0.88	0.88	0.88
Call to harassment	128	CTH	0.63	0.63	0.63
		No CTH	0.97	0.97	0.97
		Weighted Avg.	0.95	0.95	0.95
		Macro Avg.	0.80	0.80	0.80

Table 3: Performance of the best classifiers for each task during hyperparameter optimization evaluation.

of the authors reviewed all positive labeled annotations from the third-party annotation service after data set delivery.

We employed an active learning approach to sample data for crowdsourced annotation. At a high level, this cyclical process involved training fine-tuned classifiers with a subset of very precise data, using these fine-tuned classifiers to predict the entire data set, and then sampling from the fully classified data set across the distribution of the predicted scores. The sampling process separated the samples into distinct ranges based on the predicted positive class probability. We segmented the predicted data into 10 ranges between 0.0 and 1.0 and sampled evenly from each range. The crowdsourced annotators then annotated a sampled set, which we combined with the data from the prior sets to train a new classifier. We then repeated this process twice per data set for each task. We evaluated classifier performance for each task on every iteration of the active learning cycle. While we made use of crowdsourced annotations to train our filtering classifiers, all of the annotations in the later stages of the pipeline (used to confirm true positives) were performed by our domain expert annotators.

For qualitative analysis, two of the authors, who are domain experts, manually annotated 1,000 documents predicted as calls to harassment and 1,000 documents predicted as doxes at step 7 of our pipeline (Figure 1). Cohen's Kappa statistic showed strong agreement both for the call to harassment task (0.845) and for the doxing task (0.893). This suggests that having all of the annotations done by domain experts would likely have resulted in improved classifier performance. Unfortunately, it was infeasible to have this many domain expert annotations completed for our study.

5.4 Hyperparameter Optimization

When training and fine-tuning our two classifiers, one for detecting doxes and one for detecting calls to harassment, we combined a subset of annotated data from all of our data sources (i.e, boards, chat, pastes, and Gab). We preferred this approach over training individual classifiers for each task and data source for two reasons. First, the sparsity of positive examples for some data sources resulted in classifiers with poor performance. Second, the high dimensionality of the task meant that data from the different sources was unlikely to provide conflicting information to the classifier's weights.

We withheld evaluation sets of data annotations to use for hyperparameter tuning and to optimize our classifiers' parameters for better AUC-ROC scores. Table 3 presents the final classifier performance in AUC-ROC, F1, precision and recall for each task. Our final classifiers were then used to predict the entirety of all of our data sets for both tasks. We attribute the performance differences between tasks to the semantic nuance in calls to harassment compared to doxes. For example, a common false positive we observed in the call to harassment model trials were instances where

Classifier	Data set	Threshold t	Nr > threshold	Nr. annotated	True Positive
Doxing	Boards	0.9	14675	3300	2549
	Discord$^\diamond$	0.5	197	*197	153
	Gab	0.8	1905	*1905	1657
	Pastes	0.5	52849	3241	3118
	Telegram$^\diamond$	0.6	1194	*1194	948
	Total	-	**70823**	**9837**	**8425**
Call to harassment	Boards	0.935	30685	3016	2045
	Gab	0.935	2141	*2141	1335
	Discord$^\diamond$	0.5	1093	*1093	510
	Telegram$^\diamond$	0.7	4166	*4166	2364
	Total	-	**38085**	**10416**	**6254**

Table 4: Evaluation of annotated samples for all tasks. * marks data sets where every document above the threshold was annotated. We separated the "chat" data set into individual platforms with separate thresholds (indicated by $^\diamond$) in order to improve performance.

a user was encouraging the crowd to contact their local elected representative, which we do not consider harassment. We found that the best solution for accounting for the more diverse sets of call to harassment language was to gather more edge cases as training data. In addition, we experimented with training call to harassment models in individual data sets before deciding to combine training data from multiple sources. We found that the model had poorer performance when training on individual data sets as compared to using combined data. Lastly, the distilBERT architecture allows training data text of up to 512 characters. The length parameter is selected and fixed for training/testing, thus we hyperparameter optimized it to determine the best text length per task.

5.5 Threshold Selection

Our high-level goal when designing and implementing the classifiers was to narrow down the data set into a smaller set of likely true positives suitable for manual annotation. Thus, we selected thresholds that produced a set of samples that was manageable in terms of manually annotating, but still captured a diverse range of harassment attack types.

For our threshold selection, we prioritized recall in order to have as little bias as possible when selecting positive examples for empirical analysis. For each classifier, we chose a threshold t, where a predicted label $> t$ is considered the positive class (i.e., a call to harassment or dox), and a predicted label $< t$ is considered the negative class. We started by selecting a random sample of documents where $t = 0.5$ (the standard threshold), and manually annotated them to compute the precision of our classifiers. We increased t and re-evaluated if the precision was overly low to the point that we would not be able to manually annotate enough documents to create reasonably sized data sets of positive examples. As a way to ensure we were not risking recall, once the precision was sufficiently high, we lowered t and re-evaluated: if the precision remained similar to that at the higher t, we used the lower t value to obtain a higher recall. We repeated this process until selecting the final t values (one for each data source) shown in Table 4.

The "Nr > threshold" column in Table 4 counts how many documents received scores above that threshold. We refer to this selection of documents as "above the threshold" data sets for the remainder of our analysis. After selecting a threshold, we then manually annotated a sample of each data set for each task. In some instances, such as in the case of doxing on Telegram, we annotated all of the data above the threshold because its size was manageable. When the data set above the threshold was too large for manual annotation by the three domain expert authors, we annotated a random sample. As a reminder, the Cohen's Kappa statistic indicated strong agreement for our domain expert annotations. We show the total number of documents annotated in the "Nr. annotated" column in Table 4. Our annotations uncovered a total of 8,425 actual doxes and 6,254 actual calls to harassment; we refer to them as our "annotated" data sets in the remainder of our analysis.

5.6 PII and Gender Extraction Methods

We developed 12 regular expressions to programmatically extract types of PII included in doxes and calls to harassment. The types of PII we extracted include: addresses, credit card numbers, email addresses, Facebook profiles, Instagram profiles, phone numbers, U.S. Social Security Numbers (SSNs), Twitter handles, and YouTube channels. In order to optimize our regular expressions for precision, we chose to detect only U.S. phone numbers, addresses and SSNs and modified existing regular expressions from the CommonRegex Python library [28] to better fit our use case. To optimize for precision when detecting credit card numbers, we relied on a different regular expression for each type of card company. For the social media profiles, we implemented two types of regular expressions:

- Regular expressions to capture the URLs of the user profiles for each platform. We used stopwords to remove the keywords that are reserved for site functionalities, but follow the same style of URL as the user profile.
- Regular expressions to detect a combination of social media name or abbreviation, followed by the username (i.e., facebook/fb (case-insensitive): username), where the username fulfills the allowed username conditions described on each platform's website.

We evaluated the accuracy of our regular expressions on a subset of 98 true positive doxes from the pastes data set. All regular expressions had an accuracy of 95% or higher. The least accurate regular expressions were for phone numbers, street addresses, SSNs and Facebook profiles. Seven of the PII regular expressions had an accuracy of 100%.

Additionally, we attempted to identify the likely gender of dox and call to harassment targets by extracting gendered pronouns used in the text with regular expressions. This method can produce incorrect results when the attacker lacks knowledge of the self-identified gender of the target, or purposely refers to them by the wrong pronoun. The latter constitutes a type of harassment in its own, referred to as "deadnaming" [42]. In our approach, we inferred each target's likely gender based on the group of pronouns that occurred most frequently, either "he/him/his," or "she/her/hers." We manually evaluated this method in a sample of 123 doxes from the pastes data set that contained pronouns, and found that the extracted pronoun was associated with the target of the dox in 94.3% percent of the cases.

6 CALLS TO HARASSMENT

Unless otherwise stated, our analysis in this section is limited to the annotated set of 6,254 true positive calls to harassment. Table 4 shows that the majority (71%, 4,409) of the calls to harassment were from the boards (33%, 2,045) and chat (38%, 2,364) data sets.

6.1 Categorizing Calls to Harassment

To characterize the categories of calls to harassment that we discovered, we started with a taxonomy of harassment attack types from prior work [42] and adapted it based on the calls to harassment in our data set. Three of the authors coded 500 classified calls to harassment above the threshold, initially attempting to assign them to parent categories and more specific subcategories from the existing attack type taxonomy. Each call to harassment was to be assigned to one or more of these categories. In addition to assigning attack types, the authors indicated when a call to harassment either a) did not fit into any of the existing parent categories, b) fit into an existing parent category but not into a sub-category, or c) there was no clarity about which particular subcategory it fell into. The authors met several times throughout the annotation process to discuss any disagreements about specific calls to harassment.

When discussing instances of a), we found that the existing taxonomy did not include harassment by way of spreading an admittedly false narrative with the direct intent of manipulating public perception. To account for this, we added the "public opinion manipulation" parent category. In addition, we found enough specific examples of "hashtag hijacking" to designate it as a subcategory of "public opinion manipulation." In "hashtag hijacking," individuals plan to derail the message of an existing hashtag on Twitter with the intent of manipulating public perception. We also promoted the "purposeful embarrassment" subcategory into its own parent category, re-named it "reputational harm," and made a distinction between "public" instances when harmful narratives about the subject are posted publicly, and "private" instances when the harassers contacted individuals in the personal or professional networks of the target of the harassment to spread harmful information. We made this change because of the prevalence of this type of harassment, but also because it involves steps that are fundamentally different from "toxic content," the category it was grouped in previously [42]. In addition to being "toxic," instances of "reputational harm" include acts that intentionally threaten an individual's reputation, for example by contacting a place of work or family.

While discussing b), we discovered instances of calls to harassment where an individual encouraged the crowd to "bully" or "blackmail" a target, but without suggesting an explicit tactic. To accommodate these cases, we added a parent category for "generic" calls to harassment. We also observed instances of calls to harassment that fell into a parent category, but that lacked detail to warrant assigning any specific subcategory. To account for these cases, we created a "miscellaneous" subcategory for each attack type.

Lastly, while going through the examples of c), we decided to merge several subcategories because we found enough examples that encompassed both, and we saw no clear distinction between them. Previous work defines "raiding" or "brigading" as an attack where a large group of people overwhelm the comment feed of a targeted group or individual, and describes "dogpiling" as a situation where a person is targeted in order to recant an opinion or statement [42]. In our data set, we often lacked context to determine the motivation of the attacker or whether the intended target was an individual or group, for this reason we decided to merge "raiding" and "dogpiling."

We removed some types of online abuse and harassment from the taxonomy because they were either not relevant in the context of calls to harassment, or were not prevalent in the data sets we studied. For example, although "incitement" is a common form of online harassment, we found that it did not fit conceptually as a category in our taxonomy because calls to harassment are inherently considered "inciting" a crowd. In addition, we did not find any examples of "browser manipulation" or "IoT manipulation" in calls to harassment.

6.1.1 *Proposed taxonomy.*

Our final call to harassment attack type taxonomy includes 10 parent attack types (Table 5), and 28 subcategory attack types (Table 11 in the appendix). We define the 10 parent attack types as follows:

Content Leakage: Intentional leaking of personal information, media/imagery, or other PII. This category also includes doxing, e.g. *"[name] must be harassed, get her phone number and address."*

Impersonation: Intentionally pretending to represent a third party in order to do harm to the impersonated or another individual. Includes creating false imagery in order to present someone in a falsified context, e.g. *"make deep fakes of porn with them. send them to all their friends and parents and family."*

Lockout and Control: Hacking or gaining unauthorized access to a target's account, device or otherwise. Sometimes the attacker might also have an additional motive associated with gaining access, e.g. *"Physh his emails and find any info to blackmail with."*

Overloading: Attempting to put a target in a state where they are flooded with notifications, messages, calls or otherwise that they cannot manage. This category can sometimes co-occur with doxing if the targeted user accounts are included, e.g. *"Post FB & Twitter accounts so we can spam him with hate."*

Public Opinion Manipulation: Spreading narratives with the direct intent of manipulating public perception, e.g. *"We need to keep pushing that the LGBT flag is now a hate symbol. Use #ColorCulture on twitter and share on #DiversityWins, #LGBT, and any others to get people to see it. Use #NotOurFlag for the countermovement..."*

Reporting: Deceiving an online reporting system or institutional authority. Includes "SWATing" and mass account reporting to the platforms where the target holds accounts (for platform policy violations that may not actually have occurred), e.g. *"Let's mass-report his twitter and youtube..."*

Reputational Harm: Publicly or privately harassing an individual's family, employer or otherwise with the intent of damaging their reputation, e.g. *"Report him to the neighbours, he should be more careful with his atrocious beliefs if he doesn't want ostracism."*

Surveillance: Following or monitoring an individual and reporting the results online with the intent of exposing publicly otherwise private behavior, e.g. *"We should find all their yachts and stick trackers to them. And track them on gps."*

Toxic Content: A wide range of harassment including hate speech, unwanted explicit content or otherwise inflammatory remarks that are unwanted by the target, e.g. *"Did you send her a in game mail calling her out atleast? send her bleach and tell her she's trash and you'd rather a bad bitch than a fat one."*

6.2 Attack Type Analysis

Our analysis enables us to categorize the prevalence and proportion of different types of harassment incited in calls to harassment. Table 5 presents these proportions broken down by the different channels (Chat, boards, and Gab). Across all data sets, "reporting" appeared in the largest share of calls to harassment (3,193, or 51% of the total). This attack encourages co-harassers to report an individual to the authorities or to the online social networks where the individual holds accounts. The most prevalent subcategory was "mass flagging" (included in 1,496 calls to harassment), where the goal is to censor the target by inciting a group of co-harassers to use a platform's reporting features to have content removed and accounts banned. The next most prevalent reporting type was "false reporting to authorities" (found in 877 calls to harassment), where harassers report the target to immigration officials, law enforcement, employers, or parents. Potential harms can be economic (e.g., being fired from their job [16]), mental [17], and physical [22]. We suspect that the majority of calls to harassment fall into the "reporting" category because of the low effort and low cost of reporting. The next most prevalent attack category was "content leakage," which is defined as an attack that typically includes leaking not only the target's personal information (doxing), but also sensitive videos and images. In the posts we annotated, however, the overwhelming majority were doxing attacks, which are another form of harassment that does not require sustained effort.

To determine whether differences in subcategories of "reporting" attacks across data sets were statistically significant, we ran several one-way chi-square tests, while correcting for multiple testing. Nearly all differences were statistically significant ($p < 0.01$). The only subcategory of "reporting" that did not have a statistically significant difference was miscellaneous "reporting" when comparing Chat and Boards. The most frequent reporting subcategory on boards and Chat was "mass flagging," accounting for 36.20% on boards and 60.24% on Chat. On Gab, the most frequent subcategory was "misc.," which accounted for 40.18% of the reporting calls to harassment. We found that the reporting subcategories were the most balanced on the boards, where "misc." and "false reporting to authorities" accounted for 28.30% and 35.50% of the reporting calls to harassment, respectively. We manually investigated reporting subcategories on all data sets to provide qualitative insights into their differences. We found that many of the "mass flagging" examples on Gab appeared politically motivated, whereas on the boards, calls to harassment appeared to be more broadly motivated. The "mass flagging" messages in the Chat data set looked to be a counter-response to existing doxes or hate and harassment. Future work could explore in more detail common tactics and procedures used in sub-categories of "reporting" harassment.

Another significant difference ($p < 0.01$) is that compared to Chat and Gab, the boards did not have as large a focus on "overloading" (6.06%, 124), which includes the "raiding" sub-category [29, 33]. We suggest extending prior work on raiding—that was based on monitoring only boards [21, 29]—to chat (14.47%, 416) and Gab (19.85%, 265), where the attack type is more prevalent.

	Boards		**Chat**		**Gab**	
Attack Type \ Size	2,045		2,874		1,335	
Content Leakage	25.57%	(523)	21.09%	(606)	23.67%	(316)
Generic	7.14%	(146)	5.6%	(161)	4.57%	(61)
Impersonation	2.93%	(60)	1.43%	(41)	1.2%	(16)
Lockout And Control	0.24%	(5)	0.17%	(5)	0.0%	(0)
Overloading	6.06%	(124)	14.47%	(416)	19.85%	(265)
Public Opinion Manip.	6.94%	(142)	3.13%	(90)	1.72%	(23)
Reporting	56.33%	(1,152)	52.51%	(1,509)	40.82%	(545)
Reputation Harm	7.82%	(160)	12.87%	(370)	10.71%	(143)
Surveillance	0.73%	(15)	0.49%	(14)	0.37%	(5)
Toxic Content	7.63%	(156)	2.54%	(73)	4.57%	(61)

Table 5: Call to harassment parent attack types per data set. The columns do not sum to 100% since a call to harassment can include multiple attack types.

Gender. We found that 2,383 of the calls to harassment appeared to be targeting males, and 1,160 females (2,711 unknown). Table 10 in the appendix shows a breakdown of the type of calls to harassment per male and female pronouns where the pronouns could be determined. We found the largest significant gender difference in the category of "reputation harm." The subcategory of "reputation harm in a private setting" was present in 7.5% of calls to harassment labeled as "female," compared to 2.98% labeled as "male," chi-square test with $p < 0.01$. However, reputation harm in a public setting was found in only 4.66% of calls to harassment labeled as "female," and 5.96% of those labeled as "male."

When we manually investigated these differences, we found that many of the private reputation harm examples in the case of calls to harassment with "female" targets contained threats of leaking non-consensual explicit imagery to family. We could not find any such example in calls to harassment labeled as "male." This illustrates the potential disproportionate impact of gendered harassment—posting non-consensual explicit imagery ("revenge porn") is considered criminal in many parts of the world [48]. In this study, we cannot measure the impact of these particular calls to harassment, although prior studies indicate that women experience greater harms from online harassment [45]. Furthermore, we note that these results are subject to the limitations and accuracy of our pronoun-based target gender identification.

Co-occurrence. We calculated the co-occurrences of different attack types in order to measure trends in coordinated harassment. In total, 13% (831) of the annotated calls to harassment contained more than one attack type. Out of those, 767 (92.3%) contained two attack types, followed by 54 (6.5%) with three attack types, and only 10 (1%) with four or more attack types. Although a small percentage of the total annotated calls to harassment, this co-occurrence breakdown sheds light on some meaningful trends.

For example, more than half (64%) of the calls to harassment labeled as "surveillance" were also labeled as "content leakage." When manually investigating these examples, we typically saw that the crowd was encouraged to find personal information about a target and then use it to stalk them. Another co-occurrence tactic is the combination of "impersonation" and "public opinion manipulation." We saw that 30% of the "impersonation" calls to harassment were also instances of "public opinion manipulation." In these examples, the call to harassment encouraged the crowd to create fake representations of groups of individuals in order to spread a false narrative, similar to coordinated inauthentic behavior [40].

6.3 Call to Harassment Threads

To better understand how calls to harassment develop, we conducted an analysis of threads including calls to harassment on message boards. We aimed to discover (1) which attack types receive more responses from the community, (2) where in the thread calls to harassment originate, as this may help inform future work on detection and mitigation of calls to harassment, and (3) the degree to which calls to harassment and doxes co-occur in these threads.

We restrict our thread analysis to data from the boards because thread post ordering was not available to us in the other data sets. The boards data includes 4chan and other image board sites, which have a threaded structure. Users have the ability to start a thread (called an original post) or to reply to any individual message in an existing thread. We define the responses to calls to harassment as all messages in a thread after the call to harassment. As a baseline for comparison in this analysis, we selected a sample of 5,000 random posts from the boards data; we manually verified that they did not contain any calls to harassment.

Attacks that receive more responses. We ran a pairwise t-test on each call to harassment attack type and compared it to the baseline in order to measure any significant difference from normal posting activity. We used a mean difference comparison to ensure we were accounting for outliers in the distribution of thread sizes, and to measure the size of the difference for any statistically significant differences. We only ran the test on calls to harassment labeled with a single category to ensure independence of samples. In addition, we excluded "Lockout" and "Surveillance" because there were only 2 examples in each of those categories. In total, we ran the test on 1,541 calls to harassment. We ran a pairwise t-test on the log of the size of the threads in order to ensure symmetric distribution, and corrected for multiple comparisons using Benjamini Hochberg with a default error rate of 0.1.

The only call to harassment attack type with a significant difference in responses was "toxic content" with $p < 0.01$ and t-statistic of 2.8477, indicating that these threads receive a statistically significant larger response size. Upon manual investigation, we found that calls to harassment labeled as "toxic content" appear to be very low effort attacks. These threads typically call for the group to use racial slurs or other hate speech when messaging a target off-platform. Other instances of "toxic content" include sending unwanted explicit content to a target, which we also consider low effort relative to the other attack types. Figure 5 in the appendix shows a CDF of the thread size split by calls to harassment and the random baseline.

We also compared the number of responses based on gender detected in the call to harassment and found no statistically significant difference between genders when compared to one another and when compared to the baseline (one-way chi-square test).

Position in the thread where the calls to harassment appear. We found that calls to harassment rarely appear as the first (3.7%, 75) or last (2.7%, 55) post in a thread. Call to harassment posts are fairly evenly distributed over the length of the thread. The median, mean and standard deviation for thread position was 70th, 145th and 263 places, respectively. Therefore, it appears that threads tend to devolve into calls to harassment. Future work could explore the ways in which threads on the boards, or other platforms, progress into calls to harassment.

Overlap between calls to harassment and doxes in a thread. For this analysis, we used all calls to harassment and doxes above the threshold of our classifier, since our smaller annotated data sets would likely not capture much of the overlap. Unfortunately, this does introduce some error into our analysis due to false positives. We identified overlap by measuring the number of call to harassment documents above the threshold that shared a thread with a dox document above its respective threshold.

We calculated that 2,620 calls to harassment out of 30,685, or 8.53%, contained a dox. Because there are fewer doxes above the threshold, an even larger percentage (17.85%) of doxing threads contain a call to harassment. These co-occurrences of both attack types are extremely common relative to the probability that a call to harassment or a dox appear alone in a random thread, which is 0.20% and 0.10% respectively.

We manually investigated the threads that contained both doxes and calls to harassment, and noticed some themes. Sometimes, a call to harassment occurred before a dox, for example when an individual made a request for "content leakage." In other cases, the dox appeared before the call to harassment, and the information in the dox was used to promote harassment. For example, we observed doxes where PII was used in a call for false reporting to authorities or a raid. Future work could delve deeper into the dynamics of these threads to understand more nuanced instances of calls to harassment and doxes.

7 DOXING

Contrary to the calls to harassment we studied in the prior section, doxes do not necessarily contain mobilizing language to explicitly incite harassment of the individual whose PII is revealed in the dox. In this section, we study doxes as a harassment technique, based on the 8,425 annotated doxes above the threshold of the separate doxing classifier. We aim to understand the PII contained within doxes, and create a taxonomy of potentially elevated harm risks based on the PII contained in each dox.

7.1 Personally Identifiable Information

Table 6 shows the prevalence of PII found in doxes across the different platform types. We find that the doxes in our paste sample contain more types of PII on average than doxes from our board sample. This might be due to boards tending to have shorter posts than paste documents. Alternatively, a collaboration might be occurring elsewhere and the final dox is posted on the pastes.

PII \ Size	Boards 2,549	Chat 1,101	Gab 1,657	Paste 3,118
Addresses	29.34% (748)	29.61% (326)	18.04% (299)	45.67% (1,424)
Cards	0.16% (4)	4.27% (47)	0.0% (0)	4.94% (154)
Emails	14.87% (379)	14.71% (162)	20.04% (332)	45.35% (1,414)
Facebook	12.44% (317)	6.36% (70)	6.04% (100)	39.32% (1,226)
Instagram	4.2% (107)	3.27% (36)	0.6% (10)	9.97% (311)
Phones	22.17% (565)	26.98% (297)	30.24% (501)	45.51% (1,419)
SSN	0.71% (18)	1.36% (15)	0.42% (7)	3.98% (124)
Twitter	9.3% (237)	3.45% (38)	6.28% (104)	13.63% (425)
YouTube	8.24% (210)	2.0% (22)	1.09% (18)	11.8% (368)

Table 6: PII included in doxes, broken down by data set.

Harm Risk	PII
Online	Email, Instagram, Facebook, Twitter, YouTube
Physical	Address, Zip Code
Economic / Identity	Email, Credit card number, SSN
Reputation*	Family member names, place of employment

Table 7: We consider a doxing target to be at risk of certain harm types based on specific types of PII in the dox. For example, a target is at risk of "online" harm when a dox contains OSN or email address PII. *We used manual annotation for the "Reputation" risk category.

We also investigated which PII commonly co-occur with one another. We found that street addresses, phone numbers and email addresses co-occurred with all other types of PII more than 35% of the time. In addition, doxes with Facebook accounts were more likely to contain email addresses (39%), phones (25%) and street addresses (24%) compared to other OSN profiles. For example, these three categories co-occur with Youtube and Twitter accounts less than 15% and 20% of the time, respectively. A Facebook account might allow a doxer to obtain additional information that is not as easily discoverable from other OSN profiles.

7.2 Harm Risk Taxonomy

We developed what we call a "harm risk" taxonomy to contextualize how the PII included in a dox might increase the risk of specific types of harm to targets. Our taxonomy is based on a prior survey of online harassment [42] and our own analysis of calls to harassment. We use a combination of automatically extracted PII and manual analysis to determine the categories of harm that the doxing target might be at increased risk of experiencing. We define the following harm categories (summarized in Table 7):

Online Harm: Doxes that contain a social media profile can increase the risk of online harm.

Economic Harm: Doxes including credit card numbers, social security numbers, or email addresses[1] might increase the risk of economic harm.

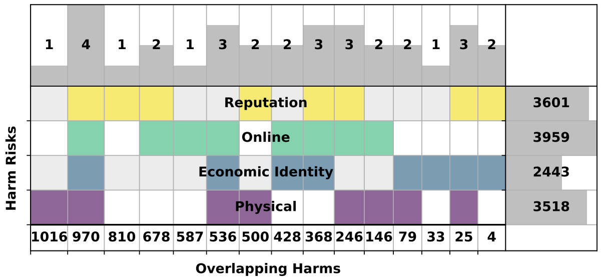


Figure 2: Venn diagram visualization of overlap between harm risk categories. Each column corresponds to a combination of harm risk categories (colored cells as opposed to the alternating grey/white cells) and each row refers to individual risk categories. The furthest column on the right-hand side is the total per category and the top row refers to the number of combinations in that particular column.

Physical Harm: Information about the physical location of an individual, such as an address or zip code, can increase the risk of physical harm.

Reputation Harm: Includes doxes that contain information about the target's family members or employer. This information is often used in reporting-based harassment. We manually annotated the samples for reputation risk since this cannot be inferred from the extracted PII.

We note that more than 50% of the Discord samples did not contain any harm risk indicators. Manual analysis showed that doxes in this data set included other types of PII not included in our extraction pipeline, such as birthday, age or nicknames. It is worth noting that Discord is the only data set with an explicit policy against doxing, and is also known for removing material violating platform policy [1].

In Figure 2, we visualize the overlap between various harm risk categories. The figure shows a type of venn diagram where each cell is shaded for each harm risk a dox contains. Categories are not mutually exclusive because doxes usually contain multiple types of PII, exposing the target to multiple types of harm. The top row indicates the number of harm risk categories in that combination and ranges between 1-4. The furthest column on the right-hand side shows the total number of doxes that were labeled with that particular row's harm risk. For example, 3,959 doxes were marked as containing "Online" harm risk. The bottom most row of the figure shows the total number of doxes per the combination of harm risks above it. For example, 970 (11.5%) of doxes contained all 4 harm risks. About 73% of doxes containing all harm risks were from the pastes data set, supporting our previous finding that doxes in this data set contain more types of PII. In contrast, indicators of online risk occur most often alone in the boards and Gab data sets. Manual analysis of entries from these data sets shows that some of them contain partial doxing information, such as an online profile, as a reply to a previous message.

Reputation risk occurs isolated, with no other risk indicators, in 23% of the cases in the chats data set. We manually looked at the

[1] We include email addresses since this might place the target at risk of their email address being compromised or the target being spear phished.

results and noticed that many of the Telegram entries containing reputation risk reveal an individual's participation in political or ideological organizations.

A major limitation of our taxonomy and analysis is that it only captures likely increased risks of harm traditionally associated with the information included in a dox. We do not measure the actual impact suffered by the targets. This is an area for future work.

7.3 Repeated Doxes

We found that there were often multiple doxes likely targeting the same person. We call these repeated doxes. Through manual investigation, we found that social media profile accounts (Facebook, YouTube, Twitter, Instagram) were the most reliable method of linking multiple doxes that were likely about the same target.

We performed our analysis of repeated doxes on the complete set of 70,820 documents above our dox classifier threshold for each data set. We used the complete set because in our smaller set of manually annotated doxes we only found 936 (11.12%) doxes that we could consider duplicate based on overlapping OSN PII. In the complete predicted set, we identified 14,587 (20.1%) doxes that contain social media accounts that appear in more than one dox. Of these repeated doxes, 98% were reposted to the same data set. Only 250 repeated doxes were cross-posted to multiple data sets. The majority of repeated doxes, 13,076 (89.64%), were posted to paste sites; 1,402 (9.61%) were posted on boards, 62 on chats, and 47 on Gab.

Our manual investigation of these repeated doxes found several potential explanations for them. One explanation appears to be that an aggressor will post a partially completed dox and update it periodically with additional information, which requires making a new post and therefore results in a repeated dox. Additionally, it appears some repeated doxes are the result of multiple different self-advertised doxing groups who are targeting the same person. Lastly, there were some instances where a dox that included multiple people was later split into individual doxes for each target.

7.4 Doxing Threads

Similar to Section 6.3, we restrict our thread analysis to only data from the boards and define the responses to doxes as all messages in a thread after the dox. We also use the same baseline of 5,000 annotated board posts that do not contain a dox.

We found that there was no significant difference in response volume based on a pairwise t-test comparing the log count of doxing and random baseline posts. We also investigated where in a board thread doxes are located. The median, mean and standard deviation for thread position were 142th, 59th and 236 places respectively. We also found that 248 doxes (9.7%) appeared as the first post in a thread, and only 69 (2.7%) appeared as the last post. This indicated that the response size would not be a good doxing detection feature. It also shows the need for future work to understand how threads escalate to doxing attacks.

8 BLOGS

Ideologically-driven blogs have become important vehicles of online participation across political ideologies [25]. There is increasing evidence that some of these communities facilitate coordinated harassment campaigns across the ideological spectrum [11, 30, 32].

	Daily Stormer	NoBlogs	The Torch
Total number of posts	36,851	78,108	93
Relevant Doxing Posts	3,072	668 (1,389*)	38
Actual Doxes (% Relevant)	90 (2.9%)	66 (9.8%)	23 (60.5%)

Table 8: Overview of the qualitative blog analysis data set. *includes entries that we could not analyze because they were in a foreign language.

8.1 Methodology

Blog entries tend to be longer than typical social media or imageboard posts. The classifiers from Section 5 did not perform well on the blog data, possibly due to blog posts extending over the 512 token mark that our distilBERT models are using. Instead of building larger and more expensive models, we chose to analyze the blog posts qualitatively. Initially, we used keyword-based parsing, beginning with simple queries containing PII terms, such as "phone" and "email," which were often indicators of harassing posts on other platforms. We used these simple queries to narrow down the number of blogs that we looked at, starting from 19 blogs that our data provider considered as high risk.

The majority of the doxes that we were able to identify were concentrated on three blogs: The Daily Stormer, The Torch Antifascist Network, and Noblogs. Our goal was to find higher-profile sites that participated in coordinated online harassment, thus this is not a comprehensive study of all blogs containing doxes.

To understand the forms and methods of harassment used in the posts in these three blogs, we performed additional keyword-based searches using keywords including: "phone," "email," "dox," and "dob:". To get an estimate for the efficacy of searches by these keywords, we evaluated them on the Torch Antifascist Network because it had a limited number of entries. We found that the query using these keywords missed 10 out of 33 doxes. We call the resulting entries "relevant" doxing posts. They are displayed in Table 8. They were annotated manually to understand the nature and prevalence of doxes on these platforms.

8.2 The Torch & NoBlogs

The Torch Antifascist Network is a blog created by a decentralized antifascist group with the same name. Noblogs, as described on their website, is a non-commercial, antifascist, antisexist, privacy-oriented blog platform [31]. The harassment methods and targets found in these two blogs are very similar to each other, for this reason we group them together in Table 9. For example, the targets of doxing in both of these platforms were individuals claimed to be members of far-right groups, participating in rallies or protests, or engaging in fascist or racist activity. NoBlogs contains many sub-blogs, spanning a variety of topics and languages but not all of them contain doxes. We found that 45% (66) of the doxes that we identified in the "relevant" data set were part of two specific blogs, one of which was focused on de-anonymizing and doxing targets from previously leaked chat logs [49].

The doxes on The Torch and NoBlogs generally start with a narration of who the target is, and the author's rationale for harassment. After the description of the target, the doxes include the target's PII.

While our data set includes no pictures, the entries almost always refer to photos of the targets, typically taken during protests and rallies, which were posted alongside other PII. The target's physical address or general location is very often also mentioned.

The authors express a goal of "alerting the community about the threat" and "reporting" to expose the far-right activity of the target through alerting neighbors, landlords or employers. These blogs seek to leverage the social stigmatization of participation in far-right organizations by using exposure of the target's identity in conjunction with evidence of their far-right activity as a tool of internet vigilantism [38].

8.3 The Daily Stormer

While The Daily Stormer contains doxes, they are different from doxes in the far-left blogs described above. Our analysis found that a dox often co-occurred alongside a call to overload a target, such as through raiding or spamming. Typically, these blog entries start with the description of an event or person, followed by a narration of the author's thoughts on the matter. The blog author often concludes the entry with the contact information of the target and a call to harass. These entries often contained less PII relative to the far-left blogs studied, such as just an email address or social media handle to reach the target, which could concentrate the overloading efforts on a specific communication channel. 60% (54) of doxes in our filtered subset of "relevant" posts from The Daily Stormer included a call to overload the target. Out of the remaining doxes, 26 (29% of the total) included a Twitter handle or email address, but not an explicit call to raid.

9 DISCUSSION

Our quantitative analysis of multiple types of harassment is only a first step towards understanding calls to harassment. While we cannot draw conclusions about the exact spread and impact of the types of attack, we can use it to inform what key stakeholders could do to mitigate these threats and build upon our improved understanding in future work.

9.1 Harassment Taxonomies

We begin our analysis of calls to harassment based on the taxonomy developed by prior online harassment research [42]. Through rounds of discussion among domain expert annotators, we discovered the need for the previous taxonomy to be adapted to the data sets we were studying. We added categories accounting for incitement of spread of false narratives, which we called "public opinion manipulation." Moreover, we promoted the "purposeful embarrassment" category into a new parent category named "reputation risk." We merged sub-categories of "overloading" to avoid the need to distinguish between "raiding" and "dogpiling" when information was insufficient. Additionally, we also introduced a "miscellaneous" category to account for occasional attacks that did not fit in our taxonomy. In addition to a more detailed overview of attacks, our study also serves as a warning to other researchers about the dynamic nature of online harassment and the need to adapt previous taxonomies to the data set being studied.

9.2 Suggestions for Future Research

We identify four stakeholders that could help mitigate the impacts of calls to harassment and further develop call to harassment and doxing detection mechanisms: the research community, online platforms, other authorities, and anti-harassment groups.

Researchers. Further qualitative and quantitative research is necessary to better understand the online harassment ecosystem. We suggest qualitative research to further study the impact of calls to harassment and doxing on targets—for example, by identifying which characteristics of a dox make it more likely to affect the intended target. As an area of future quantitative research, we recommend the development of techniques for automatic detection of a wider array of calls to harassment and doxes. Additional research could also extend our classifiers to detect each type of attack separately, in order to provide more accurate assessments of the call to harassment ecosystem. Additionally, future work can look at the dynamics of cross-platform calls to harassment and trends of growth. Longitudinal analysis of calls to harassment could provide insights into new attack types, and whether these online fringe communities are influenced by offline trends and events.

Online Platforms. Our findings also have implications on how online platforms can improve their anti-harassment efforts. For example, in Section 6 we found a high prevalence of calls to harassment encouraging abuse of account or content reporting systems, which might indicate that these systems—intended to mitigate harassment—might themselves be enabling harassment. The architecture of reporting systems themselves—generally low-friction and accessible to any user of the platform—can be used by attackers as a low-effort harassment tool. Attackers abusing platform reporting systems have been documented anecdotally [13, 20], but we are the first to show the prevalence of this type of attack. We recommend that platforms investigate their reporting systems to understand if they are being abused, and openly publish their findings. If these systems are being abused, then platforms should work with researchers to defend them, and publish openly about potential solutions to this issue. The platforms could also potentially use measurements of the call to harassment ecosystem to direct their analysis and defensive resources along with creating improved anti-harassment policies.

Other Authorities. Other authorities such as employers and law enforcement would also benefit from understanding the ecosystem of coordinated harassment and how these groups intentionally manipulate reporting mechanisms. Better understanding could help authorities better support individuals who are being harassed, while also inspiring caution in authorities to help them avoid being abused by communities of harassers to inflict greater harms.

Anti-harassment groups. Advocacy groups that work with targets of harassment can likely benefit from combining their target perspective with measurements of the call to harassment ecosystem to better identify emerging attack trends before they become widespread. We encourage advocacy groups to collaborate with researchers in conducting future research on communities that are most vulnerable to calls to harassment and doxing in order to help them develop greater awareness of online harassment strategies, and adopt online privacy best practices in mitigating harms.

	The Torch/No Blogs	Daily Stormer
Attack	**Doxing**	**Doxing**
	Invites readers to provide additional information	Often co-occurs with calls to overload
	Includes narration of activities of the target, along with PII	Includes narration of activities of the target
	Photos from rallies and protests	Contact information: Twitter handle or email
	Includes facts related to the target's physical location	**Overloading**
	Public Reputational Harm	Most common: raiding and spamming
	Distributing flyers/posters	Raiding often contains hate speech
	Alerting friends, neighbors, landlords	**Hate Speech**
	Private Reputational Harm	In the form of meme campaigns
	Alerting employer	In the form of hashtag hijacking

Table 9: Taxonomy of attacks in blogs.

10 CONCLUSION

We present the first holistic measurements of the call to harassment ecosystem. We developed a filtering pipeline that we used to identify 14,679 calls to harassment from four large-scale data sets, and also identified calls to harassment posted to ideology-based blogs. Based on our analysis of these calls to harassment, we refined and improved an existing taxonomy of harassment attack types. We used this taxonomy to categorize the preferred approaches of coordinated attackers and the proportion of incitements for various types of harassment on different platforms. Our analysis showed that over 50% of the incitements to harassment included calls to report the target to authorities or platforms. Finally, we provided suggestions for actions and future research that could be performed by researchers, platforms, authorities, and anti-harassment groups.

ACKNOWLEDGMENTS

We thank the anonymous reviewers and shepherd for their useful feedback. This research was supported by a contract from Google and National Science Foundation grants 1916126 and 2016061.

REFERENCES

[1] Bobby Allyn. 2021. Group-Chat App Discord Says It Banned More Than 2,000 Extremist Communities. Retrieved May 20, 2021 from https://www.npr.org/2021/04/05/983855753/group-chat-app-discord-says-it-banned-more-than-2-000-extremist-communities

[2] Dunja Antunovic. 2019. "We wouldn't say it to their faces": Online harassment, women sports journalists, and feminism. *Feminist Media Studies* 19, 3 (2019), 428–442.

[3] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1100–1105. https://doi.org/10.1145/3308560.3317083

[4] Jason Baumgartner, Savvas Zannettou, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Telegram Dataset. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 840–847. https://ojs.aaai.org/index.php/ICWSM/article/view/7348

[5] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Hate is Not Binary: Studying Abusive Behavior of #GamerGate on Twitter. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (Prague, Czech Republic) *(HT '17)*. Association for Computing Machinery, New York, NY, USA, 65–74. https://doi.org/10.1145/3078714.3078721

[6] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference* (Troy, New York, USA) *(WebSci '17)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/3091478.3091487

[7] Gina Masullo Chen, Paromita Pain, Victoria Y Chen, Madlin Mekelburg, Nina Springer, and Franziska Troger. 2020. 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism* 21, 7 (2020), 877–895.

[8] Mengtong Chen, Anne Shann Yue Cheung, and Ko Ling Chan. 2019. Doxing: What Adolescents Look for and Their Intentions. *International Journal of Environmental Research and Public Health* 16, 2 (2019), 218.

[9] Qiqi Chen, Ko Ling Chan, and Anne Shann Yue Cheung. 2018. Doxing Victimization and Emotional Problems among Secondary School Students in Hong Kong. *International Journal of Environmental Research and Public Health* 15, 12 (Nov. 2018), 2665.

[10] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9. AAAI, Palo Alto, CA, 61–70.

[11] Nigel Copsey and Samuel Merrill. 2021. *Understanding 21st-Century Militant Anti-Fascism: Full Report.* CREST, Lancaster, UK. https://crestresearch.ac.uk/download/3722/understanding_21st-century_militant_anti-fascism_full_report.pdf

[12] M. Dadvar, Rudolf Berend Trieschnigg, and Franciska M.G. de Jong. 2014. Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. In *Proceedings of the 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014 (Lecture Notes in Computer Science)*. Springer, New York, 275–281. https://doi.org/10.1007/978-3-319-06483-3_25

[13] Nicolas Kayser-Bril Daham Alasaad and Suniya Qureshi. 2021. *The Insta-mafia: How crooks mass-report users for profit.* Algorithm Watch. Retrieved April 27, 2021 from https://algorithmwatch.org/en/facebook-instagram-mass-report/

[14] Caitlin Dewey. 2015. This is how Twitter's new anti-harassment filter works. (Surprise! It works really well.). Retrieved May 19, 2021 from https://www.washingtonpost.com/news/the-intersect/wp/2015/03/31/this-is-how-twitters-new-anti-harassment-filter-works-surprise-it-works-really-well/

[15] David M Douglas. 2016. Doxing: a conceptual analysis. *Ethics and Information Technology* 18, 3 (June 2016), 199–210.

[16] Emma Grey Ellis. 2017. Whatever Your Side, Doxing Is a Perilous Form of Justice. https://www.wired.com/story/doxing-charlottesville/.

[17] Jerry Finn. 2004. A Survey of Online Harassment at a University Campus. *Journal of Interpersonal Violence* 19, 4 (2004), 468–483. https://doi.org/10.1177/0886260503262083 arXiv:https://doi.org/10.1177/0886260503262083 PMID: 15038885.

[18] April Glaser. 2019. Where 8channers Went After 8chan. Retrieved May 20, 2021 from https://slate.com/technology/2019/11/8chan-8kun-white-supremacists-telegram-discord-facebook.html

[19] Lauren Goldman. 2015. Trending Now: The Use of Social Media Websites in Public Shaming Punishments. *American Criminal Law Review* 52 (17 March 2015), 415–451.

[20] Shelby Grossman, Ross Ewald, Jennifer John, Asfandyar Mir, Kim Ngo, Natasha Patel, and A.R. 2020. *Reporting for Duty: How A Network of Pakistan-Based Accounts Leveraged Mass Reporting to Silence Critics.* Stanford Internet Observatory Cyber Policy Center. https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/20200901_pakistan_report.pdf

[21] Gabriel Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.

[22] Brian Krebs. 2019. Man Behind Fatal 'Swatting' Gets 20 Years. Retrieved May 19, 2021 from https://krebsonsecurity.com/2019/03/man-behind-fatal-swatting-

gets-20-years/

[23] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (Bellevue, Washington) *(AAAI'13)*. AAAI Press, Palo Alto, CA, 1621–1622.

[24] Emily Laidlaw. 2017. Online Shaming and the Right to Privacy. *Laws* 6, 1 (Feb 2017), 3. https://doi.org/10.3390/laws6010003

[25] Eric Lawrence, John Sides, and Henry Farrell. 2010. Self-Segregation or Deliberation? Blog Readership, Participation, and Polarization in American Politics. *Perspectives on Politics* 8, 1 (2010), 141–157. https://doi.org/10.1017/S1537592709992714

[26] Megan Lindsay, Jaime M. Booth, Jill T. Messing, and Jonel Thaller. 2016. Experiences of Online Harassment Among Emerging Adults: Emotional Reactions and the Mediating Role of Fear. *Journal of Interpersonal Violence* 31, 19 (2016), 3174–3195.

[27] Chen Ling, Utkucan Balci, Jeremy Blackburn, and Gianluca Stringhini. 2020. A First Look at Zoombombing. *CoRR* abs/2009.03822 (2020). arXiv:2009.03822 https://arxiv.org/abs/2009.03822

[28] madisonmay. [n. d.]. *CommonRegex*. https://github.com/madisonmay/CommonRegex

[29] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. In *Proc. ACM Hum.-Comput. Interact.*, Vol. 3. Association for Computing Machinery, New York, NY, USA, Article 207, 21 pages.

[30] Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. Retrieved April 27, 2021 from https://www.chinhnghia.com/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

[31] NoBlogs. 2021. *No Blogs*. Retrieved April 27, 2021 from https://noblogs.org/

[32] Abby Ohlheiser. 2017. The man behind the neo-Nazi Daily Stormer website is being sued by one of his 'troll storm' targets. Retrieved May 19, 2021 from https://www.washingtonpost.com/news/the-intersect/wp/2017/04/18/the-man-behind-the-neo-nazi-daily-stormer-website-is-being-sued-by-one-of-his-troll-storm-targets/

[33] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. In *ICWSM*.

[34] Nico Prucha. 2016. IS and the Jihadist Information Highway – Projecting Influence and Religious Identity via Telegram. *Perspectives on Terrorism* 10, 6 (2016). http://www.terrorismanalysts.com/pt/index.php/pot/article/view/556

[35] Kevin Roose. 2017. This Was the Alt-Right's Favorite Chat App. Then Came Charlottesville. Retrieved May 19, 2021 from https://www.nytimes.com/2017/08/15/technology/discord-chat-app-alt-right.html

[36] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*. IEEE, New York, NY.

[37] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. 2017. Fifteen Minutes of Unwanted Fame: Detecting and Characterizing Doxing. In *ACM Internet Measurement Conference*.

[38] Daniel J Solove. 2008. *The future of reputation*. Yale University Press.

[39] Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012. Automatic Identification of Personal Insults on Social News Sites. *Journal of the American Society for Information Science and Technology* 63, 2 (Feb. 2012), 270–285. https://doi.org/10.1002/asi.21690

[40] Kate Starbird. 2019. Disinformation's spread: Bots, trolls and all of us. *Nature* 571 (07 2019), 449–449. https://doi.org/10.1038/d41586-019-02235-x

[41] Joel Tetreault, Elena Filatova, and Martin Chodorow. 2010. Rethinking Grammatical Error Annotation and Evaluation with the Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Los Angeles, California, 45–48. https://www.aclweb.org/anthology/W10-1006

[42] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *IEEE Security & Privacy*.

[43] Twitch. 2021. *Our Plan for Addressing Severe Off-Service Misconduct*. Twitch. Retrieved May 19, 2021 from https://blog.twitch.tv/en/2021/04/07/our-plan-for-addressing-severe-off-service-misconduct/

[44] Twitter. 2019. Private information policy. Retrieved May 19, 2021 from https://help.twitter.com/en/rules-and-policies/personal-information

[45] The Economist Intelligence Unit. 2021. Measuring the prevalence of online violence against women. Retrieved May 20, 2021 from https://onlineviolencewomen.eiu.com/

[46] 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 *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[47] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media* (Montreal, Canada) *(LSM '12)*. Association for Computational Linguistics, USA, 19–26.

[48] Wikipedia. [n. d.]. *Revenge porn legislation*. Wikipedia. Retrieved May 19, 2021 from https://en.wikipedia.org/wiki/Revenge_porn#Legislation

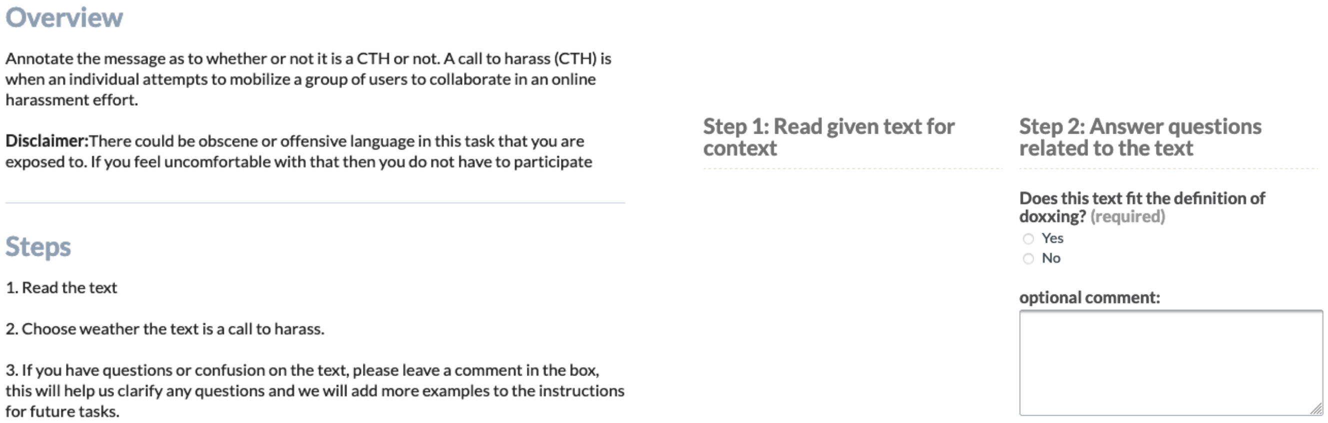
[49] Baynard Woods. 2017. How Unicorn Riot covers the alt-right without giving them a platform. *Columbia Journalism Review* (2017). Retrieved May 15, 2021 from https://www.cjr.org/united_states_project/charlottesville-alt-right-unicorn-riot.php

[50] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144 [cs.CL]

[51] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. ACM, New York, 1391–1399. https://arxiv.org/pdf/1610.08914.pdf

[52] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber. In *Companion Proceedings of the The Web Conference 2018*.

A APPENDIX



Overview

Annotate the message as to whether or not it is a CTH or not. A call to harass (CTH) is when an individual attempts to mobilize a group of users to collaborate in an online harassment effort.

Disclaimer: There could be obscene or offensive language in this task that you are exposed to. If you feel uncomfortable with that then you do not have to participate

Steps

1. Read the text

2. Choose weather the text is a call to harass.

3. If you have questions or confusion on the text, please leave a comment in the box, this will help us clarify any questions and we will add more examples to the instructions for future tasks.

(a) Call to harassment annotation guide

Step 1: Read given text for context

Step 2: Answer questions related to the text

Does this text fit the definition of doxxing? (required)

○ Yes
○ No

optional comment:

(b) Dox example annotation task

Figure 3: Example crowdsourcing annotation guide and example task (with information redacted).

```
SELECT
*
FROM `dataset`
WHERE
/*First clause: contains mobilizing lang.*/
(REGEXP_CONTAINS(LOWER(body), r'\Q we need to\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q we should\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q we need to\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q lets\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q we have\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q we will\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q we\E'))
/*subclause for in group mobilizing lang. v target*/
AND (REGEXP_CONTAINS(LOWER(body), r'\Q them\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q him\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q her\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q all\E')
OR REGEXP_CONTAINS(LOWER(body), r'\Q entire\E'))
```

Figure 4: SQL query to gather a sample of possible calls to harassment for manual annotation and initial model training.

	Unknown		**Female**		**Male**	
Attack type \ Size	2,711		1,160		2,383	
Content Leakage: Doxing	10.96%	(297)	18.53%	(215)	20.18%	(481)
Content Leakage: Leaked Chats Profile	0.15%	(4)	1.12%	(13)	0.42%	(10)
Content Leakage: Non-Consensual Media Exposure	2.69%	(73)	6.47%	(75)	2.01%	(48)
Content Leakage: Outing/Deadnaming	0.04%	(1)	0.17%	(2)	0.13%	(3)
Content Leakage: Dox Propagation	2.10%	(57)	1.64%	(19)	5.33%	(127)
Content Leakage (Misc.)	0.18%	(5)	0.34%	(4)	0.46%	(11)
Content Leakage – Total	**16.12%**	**(437)**	**28.27%**	**(328)**	**28.53%**	**(680)**
Impersonation: Impersonated Profiles	2.40%	(65)	1.29%	(15)	0.67%	(16)
Impersonation: Synthetic Pornography	0.07%	(2)	0.60%	(7)	0.08%	(2)
Impersonation (Misc.)	0.18%	(5)	0.26%	(3)	0.08%	(2)
Impersonation – Total	**2.65%**	**(72)**	**2.15%**	**(25)**	**0.83%**	**(20)**
Lockout And Control: Account Lockout	0.07%	(2)	0.00%	(0)	0.13%	(3)
Lockout And Control (Misc.)	0.00%	(0)	0.09%	(1)	0.17%	(4)
Lockout and Control – Total	**0.07%**	**(2)**	**0.09%**	**(1)**	**0.30%**	**(7)**
Overloading: Negative Ratings/Reviews	0.33%	(9)	0.09%	(1)	0.38%	(9)
Overloading: Raiding	10.44%	(283)	15.86%	(184)	9.90%	(236)
Overloading: Spamming	0.85%	(23)	0.60%	(7)	1.09%	(26)
Overloading (Misc.)	0.07%	(2)	0.26%	(3)	0.92%	(22)
Overloading – Total	**11.69%**	**(317)**	**16.81%**	**(195)**	**12.29%**	**(293)**
Public Opinion Manipulation: Hashtag Hijacking	2.55%	(69)	0.09%	(1)	0.34%	(8)
Public Opinion Manipulation (Misc.)	4.13%	(112)	2.07%	(24)	1.72%	(41)
Public Opinion Manipulation – Total	**6.68%**	**(181)**	**2.16%**	**(25)**	**2.06%**	**(49)**
Reporting: False Reporting to Authorities	13.68%	(371)	14.57%	(169)	14.14%	(337)
Reporting: Mass Flagging	30.17%	(818)	12.5%	(145)	22.32%	(532)
Reporting (Misc.)	15.75%	(427)	9.31%	(108)	12.55%	(299)
Reporting – Total	**59.60%**	**(1,616)**	**36.38%**	**(422)**	**49.01%**	**(1,168)**
Reputational Harm: Private	2.14%	(58)	7.50%	(87)	2.98%	(71)
Reputational Harm: Public	7.45%	(202)	4.66%	(54)	5.96%	(142)
Reputational Harm (Misc.)	0.66%	(18)	1.47%	(17)	1.01%	(24)
Reputational Harm – Total	**10.25%**	**(278)**	**13.63%**	**(158)**	**9.95%**	**(237)**
Surveillance: Stalking or Tracking	0.41%	(11)	0.60%	(7)	0.42%	(10)
Surveillance (Misc.)	0.15%	(4)	0.17%	(2)	0.00%	(0)
Surveillance – Total	**0.56%**	**(15)**	**0.77%**	**(9)**	**0.42%**	**(10)**
Toxic Content: Hate Speech	2.21%	(60)	3.45%	(40)	3.99%	(95)
Toxic Content: Unwanted Explicit Content	0.37%	(10)	2.41%	(28)	0.76%	(18)
Toxic Content (Misc.)	0.15%	(4)	0.43%	(5)	1.26%	(30)
Toxic Content – Total	**2.73%**	**(74)**	**6.29%**	**(73)**	**6.01%**	**(143)**
Generic	4.21%	(114)	8.53%	(99)	6.5%	(155)

Table 10: Complete call to harassment taxonomy with the prevalence of attacks per gender. Values do not sum up to 100% because multiple attack types can occur within the same call to harassment.

	Boards		**Chat**		**Gab**	
Attack Type \ Size	2,045		2,874		1,335	
Content Leakage: Doxing	17.46%	(357)	12.46%	(358)	20.82%	(278)
Content Leakage: Leaked Chats Profile	0.88%	(18)	0.10%	(3)	0.45%	(6)
Content Leakage: Non Consensual Media Exposure	5.09%	(104)	2.40%	(69)	1.72%	(23)
Content Leakage: Outing/Deadnaming	0.20%	(4)	0.07%	(2)	0.00%	(0)
Content Leakage: Dox Propagation	1.42%	(29)	5.78%	(166)	0.60%	(8)
Content Leakage (Misc.)	0.54%	(11)	0.28%	(8)	0.07%	(1)
Content Leakage – Total	**25.59%**	**(523)**	**21.09%**	**(606)**	**23.66%**	**(316)**
Impersonation: Impersonated Profiles	2.20%	(45)	1.32%	(38)	0.97%	(13)
Impersonation: Synthetic Pornography	0.44%	(9)	0.03%	(1)	0.07%	(1)
Impersonation (Misc.)	0.29%	(6)	0.07%	(2)	0.15%	(2)
Impersonation – Total	**2.93%**	**(60)**	**1.42%**	**(41)**	**1.19%**	**(16)**
Lockout And Control: Account Lockout	0.10%	(2)	0.10%	(3)	0.00%	(0)
Lockout And Control (Misc.)	0.15%	(3)	0.07%	(2)	0.00%	(0)
Lockout And Control – Total	**0.25%**	**(5)**	**0.17%**	**(5)**	**0.00%**	**(0)**
Overloading: Negative Ratings/Reviews	0.24%	(5)	0.31%	(9)	0.37%	(5)
Overloading: Raiding	4.35%	(89)	12.87%	(370)	18.28%	(244)
Overloading: Spamming	0.88%	(18)	0.77%	(22)	1.20%	(16)
Overloading (Misc.)	0.59%	(12)	0.52%	(15)	0.00%	(0)
Overloading – Total	**6.06%**	**(124)**	**14.47%**	**(416)**	**19.85%**	**(265)**
Public Opinion Manipulation: Hashtag Hijacking	0.78%	(16)	1.39%	(40)	1.65%	(22)
Public Opinion Manipulation (Misc.)	6.16%	(126)	1.74%	(50)	0.07%	(1)
Public Opinion Manipulation – Total	**6.94%**	**(142)**	**3.13%**	**(90)**	**1.72%**	**(23)**
Reporting: False Reporting to Authorities	20.00%	(409)	10.82%	(311)	11.76%	(157)
Reporting: Mass Flagging	20.39%	(417)	31.63%	(909)	12.66%	(169)
Reporting (Misc.)	15.94%	(326)	10.06%	(289)	16.4%	(219)
Reporting – Total	**56.33%**	**(1,152)**	**52.51%**	**(1,509)**	**40.82%**	**(545)**
Reputational Harm: Private	3.13%	(64)	4.45%	(128)	1.80%	(24)
Reputational Harm: Public	1.96%	(40)	8.35%	(240)	8.84%	(118)
Reputational Harm (Misc.)	2.74%	(56)	0.07%	(2)	0.07%	(1)
Reputational Harm – Total	**7.83%**	**(160)**	**12.87%**	**(370)**	**10.71%**	**(143)**
Surveillance: Stalking or Tracking	0.49%	(10)	0.49%	(14)	0.30%	(4)
Surveillance (Misc.)	0.24%	(5)	0.00%	(0)	0.07%	(1)
Surveillance – Total	**0.73%**	**(15)**	**0.49%**	**(14)**	**0.37%**	**(5)**
Toxic Content: Hate Speech	3.86%	(79)	1.98%	(57)	4.42%	(59)
Toxic Content: Unwanted Explicit Content	2.20%	(45)	0.31%	(9)	0.15%	(2)
Toxic Content (Misc.)	1.56%	(32)	0.24%	(7)	0.00%	(0)
Toxic Content – Total	**7.62%**	**(156)**	**2.53%**	**(73)**	**4.57%**	**(61)**
Generic	7.14%	(146)	5.60%	(161)	4.57%	(61)

Table 11: Complete call to harassment taxonomy with the prevalence of attacks per data set. Values do not sum up to 100% because multiple attack types can occur within the same call to harassment.

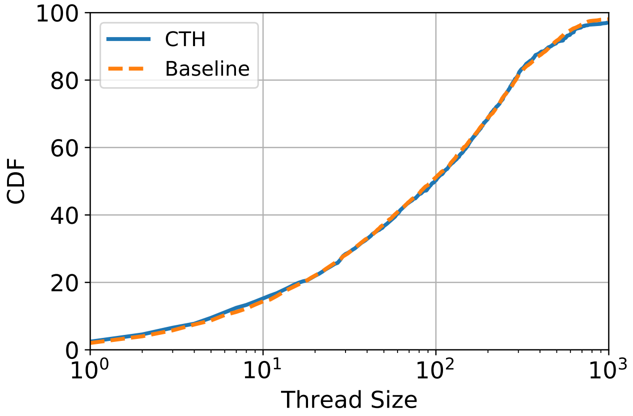


Figure 5: The length of the threads appearing after calls to harassment compared to a random baseline post.

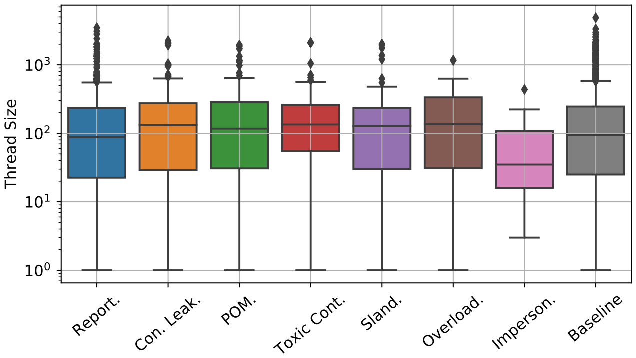


Figure 6: The length of the threads appearing after a call to harassment separately for each attack type.