# A Large Human-Labeled Corpus for Online Harassment Research

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L, Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, Derek Michael Wu
University of Maryland
College Park, MD

## ABSTRACT

A fundamental part of conducting cross-disciplinary web science research is having useful, high-quality datasets that provide value to studies across disciplines. In this paper, we introduce a large, hand-coded corpus of online harassment data. A team of researchers collaboratively developed a codebook using grounded theory and labeled 35,000 tweets. Our resulting dataset has roughly 15% positive harassment examples and 85% negative examples. This data is useful for training machine learning models, identifying textual and linguistic features of online harassment, and for studying the nature of harassing comments and the culture of trolling.

## KEYWORDS

online harassment; datasets

## 1 INTRODUCTION

*NB: This paper deals with violent online harassment. We include examples of tweets that use violent, including sexually violent, language; threats; vulgarity; hate speech; and degrading racist terms. We do not want readers to be surprised when they come upon this content, hence this note.*

Online trolling takes many forms, but at its core are posts that are harassing, offensive, threatening, and intimidating. It is not an isolated problem. The Pew Research Center found that, as of 2013, 73% of people had witnessed harassment online, and a full 40% of people had experienced harassment directly [**?** ]. They reported the following grim statistics:

- 60% of internet users said they had witnessed someone being called offensive names
- 53% had seen efforts to purposefully embarrass someone
- 25% had seen someone being physically threatened
- 24% witnessed someone being harassed for a sustained period of time
- 19% said they witnessed someone being sexually harassed
- 18% said they had seen someone be stalked

This behavior can take many forms, but online comments and other social media posts are a major source. Being able to identify the worst of these messages, understanding the motivation of those who post them, and integrating these insights into interfaces that can block or hide the content is a first step that could do a lot to improve people's experiences online.

To do this, we require a large dataset of high-quality data for analysis and model training. In this paper, we present the results of an 18-month project to create such a dataset. Our corpus contains 35,000 tweets labeled by a team of trained researchers.

## 2 RELATED WORK

Trolling is a term used to describe a very broad range of activities. Because the term is used colloquially, we will opt for "online harassment" as an overarching term in this work. We are specifically interested in the most aggressive, vile forms of online harassment. This includes threats of rape and other violence, intentionally offensive messages (racist, misogynistic, etc.), hate speech, and libelous personal insults. We want to identify messages that most would agree have firmly and completely crossed the line of "freedom of expression" or "encouraging debate". While there is a lot of research to be done in the broader trolling space, we believe that addressing these most egregious abuses will take us a solid step forward toward improving the way people interact online.

This type of trolling / online harassment has been addressed in the CMC and psychological literature. Hardaker [**?** ] offers the following:

> "A troller is a CMC user who constructs the identity of sincerely wishing to be part of the group in question, including professing, or conveying pseudo-sincere intentions, but whose real intention(s) is/are to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement."

Buckels et al. [**?** ] studied the psychological traits of trolls and, in their research, developed a Global Assessment of Internet Trolling (GAIT) scale, which builds on Hardaker's definition. The inventory uses four questions, rated on a 1-5 scale, and the mean is a user's GAIT score. Those questions are:

- I have sent people to shock websites for the lulz


ACM 978-1-4503-4896-6/17/06...$15.00.
DOI: http://dx.doi.org/10.1145/3091478.3091509

- I like to troll people in forums or the comments section of websites
- I enjoy griefing other players in multiplayer games,
- The more beautiful and pure a thing is, the more satisfying it is to corrupt

They also found that trollers scored extremely high on personality tests for narcissism, sadism, psychopathy, and Machiavellianism. In particularly, attention-seeking was a component of their personality, as was the sadistic impulse to harm others because it brought enjoyment.

Online harassment may extend beyond this, originating from hate groups or others who are not particularly entertained by their offense of others. However, this context of trolling motivation has been particularly helpful to us in this research when developing and applying codes.

## 3 CORPUS DEVELOPMENT

The lack of a good, large corpus of these kinds of messages has been a hindrance to this type of research in the past. [**? ? ?** ]. Thus, we set out to build one ourselves. We worked with some media organizations to collect and analyze blocked comments, but ultimately could not collect enough content from them to use in applications like machine learning. We turned to Twitter instead to find comments to label.

### 3.1 Initial Exploration

There is a vast world of trolling and harassment on Twitter. We were hoping to identify a representative sample of these tweets so our final corpus would be fairly representative of the types of harassing content one would find on the platform. We have access to an archive of the Twitter garden hose stream dating back to 2013. We began by randomly selecting tweets from this archive, but the density of harassing tweets was extremely low - only a few dozen tweets even rose to the level of "potentially offensive", and only a handful could have been considered truly harassing. Thus, we decided to move on to content that was more likely to be aggressive.

We considered sources of accounts that had been blocked by others. Block Together [1] is a site where Twitter users can share their block lists and others can subscribe to them. A list of block lists is not provided on the site, so we sought out lists that had been publicly shared. We collected a sample of tweets from each user on these lists and explored this set. Again, there were almost no harassing tweets in our dataset. We suspect this is because people may have been blocked for pestering the user who created the block list or otherwise being annoying without rising to the level of intense harassment or violence.

We then turned to The Block Bot [**?** ], a curated block list. Originally created by an atheist feminist community, the list contains three levels of blocks (1, 2, and 3) with Level 1 blocks being the most severe. These people have made documented threats or harassing tweets to others. Given the community who maintains the list, some issues like misogyny and Gamergate are more prominent while other racially and religious motivated issues are less common. That said, this list gave us our first set of real insights.

[1] https://blocktogether.org/

We took a sample of tweets from Level 1 blocks. As with the other sets, harassing posts were not very common. We identified only about 20 out of thousands we reviewed that rose to the level of harassment that interested us. Below are examples of some of these tweets, with the usernames of the targets anonymized. Please be forewarned that this begins the inclusion of language that may be upsetting.

```
@ANONYMIZED ROT IN HELL BITCH
@ANONYMIZED U FVCKING CVNT. YOU'LL GET URS
SOON ENOUGH
@ANONYMIZED Sorry, Canada. our hollywood
elites & its dumb models are idiots. wish
ISIS would give Chrissy a permanent haircut.
#CyberCaliphate Bloody Valentine's Day #MichelleObama!
We're watching you, you girls and your husband
@ANONYMIZED But instead, I'll drink your
blood out of your cunt after i rip it open
@ANONYMIZED i'm going to go to your apartment
and rape you to death. After I'm done, I'll
ram a tire iron up your cunt.
who ever said you could have to many slaves
was definitely one of mine, my sheds crowded
#whitepower
Black people are allowed on tv??? Wtf #fuckniggers
I'm proud to be white! #whitepower #WhitePplRule
#whitepride #whitesupremecy #whitepeoplearesupreme
#fuckniggers #FamousMelaniaTrumpQuotes
@ANONYMIZED teach me your knuckle ball technique
so i can shove my fist in your daughter
Obama is the jews nigger bitch and works
for Israel, not America. Why else would he
give the kikes 38 billion $$$ ?
```

While this set of tweets was not enough to constitute a collection, it gave us clues to begin an exploration of terms and language that could produce a much denser collection of harassing content. We began an exploratory search for terms that would produce a relatively high rate (with a minimum of around 25%) of offensive tweets. Simply searching for offensive words was not effective for this. Some of these have been reappropriated, are used in a relatively non-offensive way within communities, or are used with a much lower level of offense in other cultures. Instead, we turned our attention to hashtags and word structures that would produce a denser set of offensive tweets.

Note that this method abandoned the principle of creating a representative sample of harassing content. Whatever set of terms we defined may produce offensive content, but it would not be a true representation of all the offensive content that was out there. We accept this as a limitation of the work. It was a necessary step to producing a large enough sample.

### 3.2 Final Collection

To develop our final collection, we did an exploratory search that included derogatory terms for races and religions (e.g. "kike" included above, or "raghead"), hashtags we saw in earlier tweets or discovered in the exploration (e.g. #whitepower, #whitegenocide, #fuckniggers), and phrasing (e.g. "fucking BLANK" where the blank

is filled in with a religion or other derogatory term). In the process, we discovered other language that was not offensive by itself, but was used with high density in harassing tweets. For example "The Jews" was used about half the time in religious contexts and half the time in racist tweets at the time of our search. Indeed, simply searching for the word "feminist" produced a very high rate of harassing tweets.

In the end, we settled on the following list of search terms. It will produce a higher rate of tweets from alt-right / white nationalist tweeters, but we were willing to accept a corpus that was not necessarily representative of all harassing content in order to achieve higher density.

- #whitegenocide
- #fuckniggers
- #WhitePower
- #WhiteLivesMatter
- you fucking nigger
- fucking muslim
- fucking faggot
- religion of hate
- the jews
- feminist

We searched our archive of tweets for these terms and pulled all matching tweets. In some cases, if tweets were a response to another tweet, we included the original in the dataset. Some of these responses were harassing the original poster while others were agreeing with another harassing post. We randomly sorted the tweets and selected the first 35,000 from the list as our batch to label.

# 4 CODE BOOK DEVELOPMENT

## 4.1 Background and Context

Our team of researchers developed a code book to guide our labeling as we reviewed tweets and developed our final corpus. We made several decisions that are important to the end result in the labels. These were motivated by our desire for this corpus to be a stand-alone, text-based dataset appropriate for automated text analysis and training machine learning algorithms

First, we did not follow any links or look at images included in tweets. Without question, these would provide more context and information about the nature of the tweet. However, much of the media and links were inaccessible (and more have become inaccessible over time). Including such media in our analysis would also mean our results were not purely categorizations of the tweet text, and this is something we wanted to avoid. We believe an analysis of this media, especially images, is an open and important space for future research.

We also did not seek out other contextual elements. For example, a tweet where one user calls another a"fucking faggot" may be used between gay friends as a joke, reappropriating a term used by others as an insult. Determining that context would require a lot of digging into user relationships and guesswork by coders. We decided to ignore this context and analyze the text alone. This means that it is likely that some language which appears offensive without context was labeled as harassment when it was not indented or received that way. However, we believe there is still value in our analysis that says such language is, without special context, likely to be harassing.

We also spent time as a group learning common terms and acronyms likely to occur in this space, including things like Gamer Gate and SJW (social justice warriors), as well as the use of language or notation like white nationalist use of parentheses around Jewish names.

## 4.2 Coding Guidelines

An important part of our coding guidelines was collectively developing a line between offensive content and trolling / harassing content. We did not want to simply label any tweet that might be offensive. Indeed, we let many many offensive tweets by as "not harassment". This comes back to our main goal of building a corpus that identifies the worst of the worst content.

We broke down harassing tweets into a number of sub-categories. Because there is a lot of overlap amongst these sub-categories, the final corpus does not include these labels. However, they were useful guides for identifying specific types of content to look for. Below, we replicate the code book we developed for our internal use (without the many example tweets included)

*4.2.1 The Very Worst.* These tweets should be among the worst, most offensive or violent messages you will find. They will include content like:

- Deeply racist, misogynistic or homophobic, or otherwise bigoted. Not a little bit politically incorrect or slightly offensive. Something that would be very upsetting to a general reader.
- The use of shocking language primarily to upset the person who is reading. Words like nigger, cunt, etc. Note that these are used without that shock power by certain groups, so the simple presence of one of these words is not enough to fall into this category.
- Unapologetically or intentionally offensive this could be someone saying something with the intent of upsetting a group, or an extreme account (e.g. Neo nazis) using language that they approve of but they know the general public would disapprove of. Often this will be intended to upset people, but it could be from someone who does not care whether other people are upset and simply chooses to use what he knows will be offensive language.

*4.2.2 Threats.* These have language intended to make the target or a broader group fearful or to feel unsafe. The threats may be personal ("I saw you hitting on my boyfriend and I'm gonna cut you") or general ("we need to send the Jews back to the gas chambers"). Maybe explicit ("I'm going to a spray your brains all over the wall") or they may generally make the target feel unsafe without a specific threat of direct action (e.g. "someone needs to make you suck his dick" or "you just need the right man to fuck you").

*4.2.3 Hate speech.* These tweets express hate or extreme bias to a particular group. Could be based on religion, race, gender, sexual orientation, etc. Generally, these groups are defined by their inherent attributes, not by things they do or think.

Do not confuse political disagreement or political speech with hate. "I hate all Democrats" is not hate speech. "I hate all niggers" is. Political hatred is not hate speech because it's based on a political disagreement. The latter example is hate speech because it's

based on an inherent trait of a group of people. This is not just hate between two people. "I hate you so much!" would not count. This has to be hate toward a particular group.

*4.2.4 Directed Harassment.* Language directed at a particular person or group designed to upset them. This language may be milder than in other cases but should be part of the campaign (by one person or a group) to make the target feel threatened or intimidated. This could overlap with any of the above categories, but it does not have to. An @mention that says "you are a loser and I hope you die" would not count as any of the above but would count as directed harassment. There can be a lack of clarity here because sometimes statements are made in jest among friends. For the purposes of this project, imagine the comment coming from an unknown stranger.

Again, do not misconstrue a disagreement with harassment. Someone saying you are an idiot for, say, voting for Candidate X is likely not harassing the other person but disagreeing with them. Think carefully before you call it harassment - is the tweet intended to intimidate or upset the target, or is it intended to express disagreement with an opinion?

The target of Harassment should be a person or a group. People are usually targeted with @mentions ("@jengolbeck you need to die"). Groups may be targeted with a hashtag ("#blacklivesmatter <- no they dont") or by a name (e..g "The Jews").

*4.2.5 Potentially Offensive.* In our first round of coding, researchers had a hard time letting offensive content, like jokes in poor taste, go by without labeling them as harassment, even though they were relatively mild. To help us psychologically overcome that barrier, we introduced this category of Potentially Offensive content. These tweets will still be labeled as "non-harassing" in the final dataset, but it will help so we dont feel bad about letting offensive content go by unlabeled. For anything that doesnt rise to the level of clearly and unambiguously fitting into the categories above, you can label it Potentially Offensive.

*4.2.6 Non-Harassing.* This section of coding was for tweets that did not meet the aforementioned criteria. That did not mean these were considered "good" tweets. If a coder was uncomfortable with labeling a tweet with a disapproving label, then the coder could categorize it as potentially offensive. "Potentially Offensive" served as a label to make the coder feel better, though we did not distinguish these tweets from non-harassing tweets in the final dataset. It was critical that we did not apply the labels above too liberally. We made sure the text alone, without seeking extra content, was truly harassing. If the coder was unable to discern the category, then they marked it as 'non-harrasing'.

## 4.3 Coder Training

The coders on this project underwent extensive training on the codebook. Before labeling our final corpus, we spent three weeks reviewing and refining the codebook above, labeling sample sets of tweets as a group, and discussing the results. We went through four rounds of sample tweet coding, with the chosen examples becoming more ambiguous in each round. We followed each sample coding with extensive discussion to help highlight issues and make sure everyone had a similar calibration for the categories.

# 5 LABELED CORPUS

We spent 3 months labeling tweets in the corpus. Each tweet was labeled by two coders. If they agreed that it was Harassing or Non-Harassing (including "potentially offensive" labels), it was added to the final corpus. If they disagreed, a third coder was brought in to break the tie.

Of the 35,000 tweets, 2,711 required a third coder. This gives us an inter-rate agreement measured by Cohen's Kappa as 0.84 for the initial codes.

Given the inflammatory terms we used in our data set collection, we expected a high number of harassing tweets. However, only 15.7% of tweets rose to this level. Some were non-harassing because they were relating quotes or news stories, e.g.

`#DrudgeReport 'F*ck the Jews' scrawled at Jewish school in London...: http: / /t.co /ZaUB0piMN4 #News`

Others used the terms in non-offensive ways, such as the following:

`For the record, I see tax havens as the next world war, not fucking Muslims.`

Still others referred to someone else's opinion, so the tweet itself was not harassing:

`He hates all the Jews, he hates all the Jews... Samir Nasri, he hates all the Jews.`

We highlight these issues because they present research challenges for classification. How to differentiate between a tweet asserting a harassing or hateful position and one quoting it or attributing it to another will be a challenge. We hope that this dataset - with 5,495 positive examples and 29,505 negative examples - will be of use in discovering such distinguishing features.

# 6 DISCUSSION AND CONCLUSIONS

In this paper, we present a hand-coded dataset of 35,000 tweets labeled as Harassing or Non-Harassing. The research team spent months developing, refining, and training on a codebook designed to capture truly harassing content. We believe this dataset significantly contributes to the web science research community, both as a base of ground truth for training algorithms and as a source of data to understand and analyze the online harassment phenomenon.

Beyond its usefulness in machine learning, there is also important web science research to be done on the culture and patterns of online harassment. While this certainly will change as topics of discussion change, our codebook and dataset can provide a launching point for deeper qualitative research. For example, our work has many examples, positive and negative, that use the #WhiteGenocide hashtag. With these thousands of data points, researchers could define a much more detailed set of codes that describe the references and types of use (e.g. to refer to diversity initiatives broadly, as anti-immigration, as white nationalism, as anti-interracial marriage, etc.). Those can be developed, refined, and human coders can be trained on this stable dataset before they may choose to work on live data which presents its own complexities.

Because of Twitter terms of service restrictions and privacy concerns about individuals whose tweets are included, we are not posting the dataset in a public repository. However, this data can be shared with researchers who agree to a Data Terms of Use which

includes ethical considerations. Researchers can request access to the data via email to jgolbeck@umd.edu.

## 7 ACKNOWLEDGMENTS

This work was supported by NSF Award #1546829.

## REFERENCES

[1] Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. 2014. Detecting Online Harassment in Social Networks. (2014).
[2] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and individual Differences* 67 (2014), 97–102.
[3] Maeve Duggan and Aaron Smith. 2013. Social media update 2013. *Pew Internet and American Life Project* (2013).
[4] Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: from user discussions to theoretical concepts. *Journal of Politeness Research* 6, 2 (2010), 215–242.
[5] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual acm web science conference*. ACM, 195–204.
[6] 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 (2012), 270–285.