“Is No Thread Safe?”
Analyzing Harassment in Social Media Using Reddit Comments

1. Objectives

This project aims to use a data-driven computational linguistics approach to investigate harassment on social media, with a use-case of prototyping a formal linguistic model for identifying online harassment that can be used by social media companies to accurately identify harassing behavior that may be in violation of their use policies.

The principal questions this analysis aimed to address are: how do we define harassment on social media? what is the prevalence of harassment? how can a computational linguistic model identify harassment? what are the advantages and shortcomings of such a model?

2. Background

The pervasiveness of online harassment has been documented not only on the sites themselves, but also in research studies and mainstream media. For example, the Pew Research Center reported that 18% of all internet users report having experienced “being the target of physical threats, harassment over a sustained period of time, stalking, and sexual harassment.” and 73% of internet users report having witnessed online harassment (Duggan 2014). Mainstream media sources have also shed light on these issues. For example, John Oliver’s late night talk show “Last Night Tonight”, which included a segment on online harassment and threats of physical violence towards women on the internet in June 2015 (Faircloth, 2015).

There is a shortage of legislation around online harassment, and current legal processes fail to keep pace with the rapidly changing landscape of digital communities (Dewey 2016). A lack of adequate intervention by the administration of these privately-owned websites is also widely criticized (e.g Franzen, 2015).

In response to such criticism, several social media sites have made public announcements to better address these issues (e.g. Tambln, 2015). Other social media services have added features to discourage harassing behavior. For example, the dating app Tinder has limited the number of outgoing messages a user can send in a day, and another dating service only allows women to send first messages. However, criticisms continue that social media and online dating sites are not doing enough to protect users, specifically women (Urquhart & Doward, 2013).

The majority of the current techniques for addressing harassment are reactive in nature (e.g. reporting behavior that has already occured) require a prohibitive degree of manpower (e.g. dependence on moderators), or leave the responsibility of action to the victim (e.g. tools for blocking a user).

Comprehensive preventative approaches that increase the efficacy of identifying harassment without significantly increasing manpower are lacking, and text-based approaches are widely passed-over in favor of features such as flagging tools or reporting. This project aims to add to the current landscape of harassment mitigation techniques by prototyping a text-based model than can filter messages likely to be harassing to a manageable quantity, which can then be efficiently reviewed by staff. Such a model would allow site administration to be accountable for fostering appropriate user behavior without dedicating excessive resources to this cause.

3. Defining Harassment

Harassment is a concept that has grown more complex in the digital age, as new communication platforms and varying degrees of anonymity emerge. New digital techniques such as doxxing (distributing personal information about an individual with the intent to cause harm) and revenge porn (publishing explicit images of an individual in order to ruin their reputation) expand the already diverse set of approaches to harassment. Therefore, a comprehensive definition for online harassment aims to go beyond classic definitions of threatening actions to encompass a wide range of behaviors and their effects. For the purposes of this project, harassment was defined as:

> targeted behavior that would lead a reasonable person to conclude they are not free to express themselves without endangering their personal safety, security, or emotional well-being.

This definition was arrived at after a review of legal definitions, as well as a variety of social media sites' user policies regarding harassment/abuse. The aspect of "targeted" behavior was included to attempt to account for a difference between behavior that is generally hateful or intolerant and that which affects specific individuals or groups directly. The concept of endangering personal wellness was meant to differentiate between harassment and behavior that is insulting or annoying in nature, without more profound psychological effects. The self-expression element is meant to be context-specific to social media, whose primary model is community-based discussion and interaction

4. Data Source and Collection

4.1 Data Source

Reddit was used as a representative site for a preliminary analysis of harassment. It was deemed eligible for analysis based on its high traffic, strong reputation for harassing content, demonstrated interest in mitigating harassment, and publically accessible data.

Reddit is a major social media site self-reported to have over 3 million monthly registered users and 245 million unique monthly visitors. The platform is a bulletin board style where users post external or original content and discuss and evaluate posts through comments and a voting system. Their data policy states that all posts are public and may be viewed or used by any user or visitor.

The site has been widely criticized for its high instances of harassment, especially through the use of dedicated online communities organized against specific groups. Although traditionally known as an open-forum community with a large contingent of staunchly anti-censorship users, Reddit has taken recent steps to address harassment. For example, the administrators announced a major change to their user policy in January 2016 that included the principal "do not threaten, harass, or bully," and has since implemented a more active role in banning or restricting harassing or potentially harassing content. The site also expanded their blocking feature, to allow users to block others who directly engage with them without the aggressor's knowledge. These changes have been recent, and are not yet rigorous enough to represent a comprehensive change to the community at large.

4.2 Collection

Data was collected using an open-source program designed to scrape the entire content of an designated subreddit within any given data range. Subreddits and dates for two types of data are as follows:

4.2.1 Targeted data

A set of specific communities were for selected for scraping based on their high likelihood of containing harassing content. As banned content was not accessible, candidacy was based on communities that

were directly associated with acknowledged harassing communities (i.e. they were listed as allies to such a group, or were created as a replacement after a subreddit was banned to circumvent moderation.) Additionally, some subreddits were included because they were listed in forum discussions of the most harassing subreddits.

Seventy-five subreddits were investigated for analytical potential, but over half were disqualified on the basis they were either banned, outdated, or contained too few posts for analysis. Thirty-five subreddits were used in the final data collection. Appendix A contains a list of all subreddits used for this targeted collection. Data was collected once daily

Data was collected weekly on Sundays for all posts in each subreddit that week for five weeks February 7th-March 13th, 2016.

4.2.2 Wild Data

In order to collect an unbiased cross-section of the entire site, data was also collected from r/all, which contains every post on the entire site Data was collected once daily to retrieve all posts from the past 24 hours between the dates of March 16th-28th, 2016.

It is worth noting that that a daily/weekly frequency of collection decreased the likelihood of retrieving data before if was removed by moderators or edited/deleted by users, and thus the data may not demonstrate a truly representative sample of all harassment that occurs on the site.

4.3 Extracting Comments

After collecting all content from these pages, data was parsed to isolate comments only, with the rationale that the bulk of harassment is likely to be concentrated on user-to-user interactions, as opposed to post titles written publically as discussion topics. Comments represent the most direct (public) communication between users on the site, which most closely aligns with the current definition of harassment as targeted behavior.

5. Final Data and Labeling

5.1 Data Set

The final data set contained the following number of comments:

Wild Data: 7,402,341
Target Data: 640,191
Total: 8,042,532

5.2 Frequency of harassment

A total of 4283 lines of data were labeled by one labeler as harassing or non-harassing using the definition given above. The percentage of harassment vs. not harassment can be seen below:

Table 1. Harassment in Labeled Data

Label	Instances	% Total
Harassment	93	2.17%
Not Harassment	4,190	97.83%

Discussion: The frequency of 2.17% percent of all comments including harassment helps to elucidate that harassment is occurring on the site despite moderator efforts, at a rate of approximately 21 in every 1000 comments. Given that Reddit reports 725.85 million comments a year, this percentage would scale to a significant number (over 15 million/year) of instances if applicable sitewide.

5.3 Frequency of Moderator Activity

Moderator activity was defined as a moderator removing a post or banning a user from a subreddit, measured in instances of the text that occurs in those circumstances, [removed] and [you have been banned from posting to], respectively.

Table 2. Moderator Activity in Dataset

Type	Instances	% Total Data
Removed	111,432	1.39%
Banned	13,094	0.16%
Total	124,526	1.55%

Discussion: The moderator activity does not imply that the text or behavior removed was harassing per se, but it does give a sense of how actively the moderators are intervening in otherwise unmediated user-to-user interaction. It is worth noting that, despite moderators taking action on 1.5% of comments, there is a greater percentage (2.7%) of posts that are harassing that are either missed by or not reacted to by moderators.

6. Building a Model

The linguistic model for this analysis was based on a classifier comprised of regex patterns that returned positive and negative matches with text in the labeled dataset.

6.1.1 Dictionary of Lexical Tokens

The approach to building a rules-based linguistic model began with compiling a list of (primarily adjective and noun) tokens associated with harassing behavior, such as racial slurs and derogatory insults. An initial dictionary was built using external sources resources such as online collections of harassing messages on social media and forum discussions of well-known Reddit harassment. This lexicon was further refined through a visual inspection of a random unlabeled segment of the dataset. The revised dictionary included tokens specific to the Reddit community or otherwise previously unknown. For example, the word “autist,” a negatively connotated noun meant to convey low intelligence derived from the adjective ‘autistic,” was observed through data inspection alone. This token was observed 2169 times in the dataset.

6.1.2 Syntactic Patterns

Visual inspection revealed common syntactic patterns in harassing language in the data. The bulk of these constructions centered around overtly violent verbs such as rape, kill, beat, murder, die and the expression of desire or intent to carry out these actions on the interlocutor (e.g. “I hope you,” “I’m going to,” “I will”). Some verb phrases such “fuck off” and “fuck you” were observed as likely to be part of harassing patterns. Additionally, a significant portion observed harassment centered around the second person singular, especially in negative assertions of another’s identity (e.g. “You’re just a little ___,” “You __”). Therefore, multiple variations of this class of “you”-centric sentence structure were included in an attempt to match novel utterances fitting this pattern.

6.1.3 Accounting for Variation

Regex patterns for these items were then built, then adjusted to account for alternation such as spelling mistakes, internet slang, symbol swearing, and other variations on standard orthography For example,more than 3 symbols in a row was determined likely to be a symbol replacement for a swear word (e.g. “@$$!*&&”) a and set of symbols was used as an optional character class in place of the vowel in common swears (.e.g shit, fuck, cunt).

6.2 Classifier Development

The regex patterns developed using the lexical items and syntactic structures above were combined in myriad ways (as follows) and tested against random segments of the labeled data to determine their efficacy, measured by precision and recall. The following order of operations was used to refine the model:

6.2.1 The Dump Test

First, a regex pattern of all tokens was matched over the labeled data to determine how comprehensive the word list was:

Table 3. All Tokens Regex Test

VERSION	True Positives	False positives	True Negatives	False Negatives	Precision	Recall
DUMP	86	1512	2678	2	5.40%	97.70%

This test revealed that recall was high (98%), demonstrating the word list accurately returned nearly all instances of harassment. However, precision was an extremely low 5%, confirming that this method over-generates a large degree of text that contains similar items but is not harassing in nature.

Tokens were also grouped into parts of speech (adjectives, nouns, verb phrases) and tested separately:

Table 4. Part-of-Speech Regex Test

VERSION	True Positives	False positives	True Negatives	False Negatives	Precision	Recall
ADJ	47	205	3985	43	18.70%	52.20%
NOUN	68	682	3508	22	9.10%	75.60%
ADJ+NOUN	76	804	3385	15	8.60%	83.50%
VERB PHRASE	74	1025	3165	16	6.70%	82.20%

This test revealed a similar pattern than non-discriminated word lists produce a high recall but low precision, demonstrating a more complex pattern is required to separate the signal of harassment from the noise, as this method allows for a lot of noise to be returned.

6.2.2 Two Plus

Next, the same group of all regex patterns were used, with the specification that at least two of the tokens must be matched in order for a comment to be returned as positive for harassment. It returned the following results:

Table 5. “Two Plus” Regex Test

VERSION	True Positives	False positives	True Negatives	False Negatives	Precision	Recall
2PLUS	30	9	4181	60	76.90%	33.30%

This method had the opposite effect of being too restrictive, so that precision is high (77%), but too many relevant items are not retrieved by the classifier, as the criteria is too stringent.

6.2.3. YOU+

Due to the high instances of the second person singular in harassing text, a classifier was run requiring the verb phrase “You/You’re a ___” (and closely related variations) with any other token. It returned:

TABLE 6. “YOU __” Regex Test

VERSION	True Positives	False positives	True Negatives	False Negatives	Precision	Recall
YOU__	25	7	4182	66	78.10%	27.50%

This approach had higher precision and recall than expected, but returned too many false negatives to be considered a comprehensive model. However, it illustrates how common this type of construction is in harassing behavior.

These test made apparent the need for a tiered system wherein some tokens were prioritized over others for their likelihood to be harassing.

6.2.4. Robust Model

Finally, a model was built to only return text that contained either:

One or more items highly likely to be harassing on their own:

A. Complete or intense verb phrases (e.g. fuck yourself, kill yourself, die you __, I’ll rape you) dysphemistic words such as the most extreme racial slurs or sexualized language (e.g niggerfag, shitlord, autist, whore)

--OR--

At least two items (one or more of each class) likely to be harassing in combination:

B. Incomplete syntactic constructions (e.g you’re a __), relatively weak verb phrases (e.g shut up) and intensifiers (e.g. fucking)

C. Averagely offensive nouns/adjectives (e.g fatass, moronic, paki, pussy)

This model can be visualized as:

A | ((B)+.*(C)+)

Five versions of this model were created and tested, with adjustments made as to which constructions fit in which classes, the amount of flexibility for infixing left in the different syntactic structures, and the groupings of related words. The results from testing those versions are:

Table 7. Tuning Classifier Results

VERSION	True Positives	False positives	True Negatives	False Negatives	Precision	Recall
V1	69	317	3385	21	17.90%	76.70%
V2	73	305	3884	19	19.30%	79.30%
V3	77	270	3919	13	22.20%	85.60%
V4	71	110	4079	20	39.20%	78.00%
V5	61	44	4145	28	58.10%	67.80%

The best of these models (V5) was selected as the candidate classifier, which was tested over five random subsets of the labeled data:

VERSION	True Positives	False positives	True Negatives	False Negatives	Precision	Recall
V5_1	5	4	413	2	55.60%	71.40%
V5_2	7	5	399	4	58.30%	63.60%
V5_3	5	2	434	4	71.40%	55.60%
V5_4	4	3	414	2	57.10%	66.70%
V5_5	6	5	417	2	54.50%	75.00%
				Average	59.4%	66.5%

This classifier had an average of 59% Precision and 67% Recall.

6.2.5 Candidate Classifier on Wild Data

The candidate classifier was run on on the unlabeled dataset, and examined for consistency with the results above. A visual inspection revealed that the classifier's precision was lower, returning a large amount of false positives. This is likely due to a larger dataset allowing for more instances of creativity and new token generation than was present in the small sample labeled. Over-generating positives is preferred to generating too few or returning too many false negatives, given that the use case of this model. The use-case includes a step for human quality control on the model's output. Thus, it is preferred for a reviewer to see superfluous instances of non-harassment than to never be exposed to harassing text for review

7. Challenges to Accuracy

This type of rules-based model was moderately effective, especially in the case of spelling alternations and internet slang, as the regular expression constructions could account for a wide range of iterations of

a base token. For example, the model returned the following comment, which contains a alternations of the canonical forms of “nigger,” and “faggot.”

thanks nigr faget

However, there were a fair number of areas where the model could not effectively differentiate between signal and noise. Many of the false positives returned by the model are offensive, hateful, or insulting in nature, but not considered harassing because they are presented in an untargeted fashion (general expression of intolerance or anger, etc). A more robust model would be needed to account for the subtleties between these types of interactions more accurately, by taking into account the varied contexts in which one item can have different effects. Many false negatives were returned because the model is too rigid to account for the abundant creativity possible in the digital medium. The challenge, rooted in generative capacity of language, persists in all attempts to computationally address linguistic problems. However, it is especially relevant in the rapidly evolving linguistic space of online communities, where change is more rapid than spoken language. A rules-based approach is particularly limited in an exceptionally creative context wherein there is a high level of innovation and variability in the data.

8. Future Directions

There are several directions that are worth pursuing as follow-ups to the current analysis. Some examples include:

8.1 Further Analysis of Current Dataset

Using the current dataset, further analysis of different aspects of the data could be illustrative of relevant patterns. For example, a pattern was casually observed that the length of harassing comments tended to be shorter than non-harassing text, on average. This observation may merit further investigation to determine whether there is a correlation between hastiness and harassment, and therefore if harassment could be addressed by interventions aimed at impulse control. Alternatively,, different text fields could be analyzed for frequency of harassment to determine variations in different ways of addressing audiences.

Additionally, building a larger set of labeled data would increase the reliability of testing figures, and including using multiple labelers per line of text would help ensure external validity of labels.

8.2 Sexual harassment:

In the current analysis, sexual harassment proved difficult to definitively identify without context. For example, posts that comment on a woman’s figure in a positive but explicit way could be interpreted as consensual or abusive depending on the consent and interpretation of the woman. Analyzing such data would require a more complete account of conversations, in contrast to the current analysis of isolated comments alone. It would be worthwhile to explore how different approaches could better account for this type of data, such as including the original post and/or the woman’s responses to such sexualized comments.

8.3 Further Data Sources

Other social media sites should be analyzed to determine whether any findings are artifacts of the particular medium or community practice of this particular site. For example, a less anonymous community may demonstrate different patterns, such as harassment focused more towards outward identifying traits (race, gender, weight), or less overt harassment tactics (due to higher accountability).

9. Recap

In general, this analysis found that the prevalence of harassment is significant in reddit comments (2.1%), and although the moderators are demonstrating a fair amount of activity, it is insufficient to catch all the instances of harassing behavior. A rules-based classifier can narrow the instances of text likely to be harassing with some efficacy, but has barriers to accuracy as presented by a) the variable meanings of the same tokens in different contexts b) the unusually creative linguistic environment of the digital sphere, and social media in particular. Further, defining harassment as separate from other types of inappropriate behavior such as discourtesy, insult and hate presents a challenge for a computational model that cannot account for connotation as well as a natural speaker.

It is important to note that although the instances of harassment may seem low, there is a high degree of discussion surrounding the effects of harassment on the community. For example, one comment from the current dataset simply reads:

> IS NO THREAD SAFE?

Capturing the notion that the effects of harassment are amplified by the discussion of these instances, and its effect on the perception of community safety. Even though these behaviors are targeted, they have a ripple effect on everyone who uses the platform.

Sources:

Dewey, Caitlin."In The Battle Of Internet Mobs Vs. The Law, The Internet Mobs Have Won".*Washington Post*. N. p., 2016. Web. 10 Apr. 2016.

Duggan, M.
Duggan, M. (2014). *Online Harassment.Pew Research Center: Internet, Science & Tech*. Retrieved 20 April 2016, from http://www.pewinternet.org/2014/10/22/online-harassment/

Faircloth, Kelley. (2015). *Jezebel.com*. Retrieved 20 April 2016, from http://jezebel.com/john-olivers-internet-misogyny-rant-is-satisfying-as-he-1713019882

Tamblyn, Thomas. May 2015 Twitter CEO Admits 'We Suck At Dealing With Abuse'.HuffingtonPost.http://www.huffingtonpost.co.uk/2015/02/05/twitter-ceo-we-suck-at-dealing-with-abuse_n_6619534.html

Urquhart, C. and Doward, J.
Urquhart, Conal, and Jamie Doward. 2013. "Anti-Stalking Group Has Received Many Calls From Women Who Met Violent Men Online". *The Guardian*. Accessed April 20 2016. http://www.theguardian.com/lifeandstyle/2013/sep/07/stalking-harassment-online-dating-sites

Appendix A: Subreddits Used for Targeted Data Collection

/mensrights
/r/AwfullyPunchableFaces
/r/becomeaman
/r/BlackPeopleTwitter
/r/changemyview
/r/fatlogic
/r/fatpeoplestories
/r/imgoingtohellforthis
/r/jokes
/r/misogyny
/r/offensive
/r/pics
/r/picsofniggers
/r/polacks
/r/shemales
/r/StruggleFucking
/r/SubredditDrama
/r/TRPOffTopic
/r/videos
gaming
pettyrevenge
r/4chan
r/askTRP
r/Awfulthoughts
r/Drama/
r/funny
r/ImAWhinyLiberalBitch
r/Punchablefaces
r/RedPillWomen
r/ShitRedditSays
r/TheRedPill
r/thiscrazybitch
r/unexpectedjihad
r/whiterights
r/worstof