

A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research

Mohammadreza Rezvan
Kno.e.sis Center
Dayton, Ohio, USA
rezvan@knoesis.org

Saeedeh Shekarpour
University of Dayton
Dayton, Ohio, USA
sshekarpour1@udayton.edu

Lakshika Balasuriya
Kno.e.sis Center
Dayton, USA, Ohio
lakshika@knoesis.org

Krishnaprasad Thirunarayan
Kno.e.sis Center
Dayton, Ohio, USA
tkprasad@knoesis.org

Valerie L. Shalin
Kno.e.sis Center
Dayton, Ohio, USA
valerie@knoesis.org

Amit Sheth
Kno.e.sis Center
Dayton, Ohio, USA
amit@knoesis.org

ABSTRACT

A quality annotated corpus is essential to research. Despite the recent focus of the Web science community on cyberbullying research, the community lacks standard benchmarks. This paper provides both a quality annotated corpus and an offensive words lexicon capturing different types of harassment content: (i) sexual, (ii) racial, (iii) appearance-related, (iv) intellectual, and (v) political¹. We first crawled data from Twitter using this content-tailored offensive lexicon. As mere presence of an offensive word is not a reliable indicator of harassment, human judges annotated tweets for the presence of harassment. Our corpus consists of 25,000 annotated tweets for the five types of harassment content and is available on the Git repository².

CCS CONCEPTS

• **Applied computing** → **Annotation**; *Document analysis*; Document management and text processing; • **Social and professional topics** → *Hate speech*; *User characteristics*;

KEYWORDS

Annotated corpus, context, sexual, racial, political, appearance-related, intellectual, cyberbullying, harassment, offensive Lexicon, profane word.

ACM Reference Format:

Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin, and Amit Sheth. 2018. A

¹Disclaimer: This paper is concerned with violent online harassment. To describe the subject at an adequate level of realism, examples of our collected tweets involve violent, threatening, vulgar and hateful speech language in the context of racial, sexual, political, appearance and intellectual harassment. While these examples are shared to portray the reality, the readers are alerted in advance and may wish to avoid reading this material if it could cause discomfort and disagreement.

²<https://github.com/Mrezvan94/Harassment-Corpus>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WebSci '18, May 27–30, 2018, Amsterdam, Netherlands

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5563-6/18/05...\$15.00

<https://doi.org/10.1145/3201064.3201103>

Quality Type-aware Annotated Corpus and Lexicon for Harassment Research. In *WebSci '18: 10th ACM Conference on Web Science, May 27–30, 2018, Amsterdam, Netherlands*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3201064.3201103>

1 INTRODUCTION

Social media is being used extensively by people from various age-groups (e.g., 80+% usage for young adults (18-49) and 45+% usage for old adults (50+)³). Despite the communication advantages, participants may experience insult, humiliation, bullying, and harassing comments from strangers, colleagues or anonymous users. One-in-five, around 18% are affected⁴, posing numerous challenges to social engagement and trust, resulting in emotional distress, privacy concerns and threats to physical safety. All instances of harassment necessarily reflect a combination of sender intentionality and recipient experience. Our focus here is on the sender, whose messages are intended to harass. We study harassment[7] in five content areas: (i) sexual, (ii) racial, (iii) appearance-related, (iv) intellectual, and (v) political. Below, we briefly describe each type.

- **Sexual harassment** concerns sexuality and often targets females. The harasser might refer to a victim's sex organs with slang or describe sexual relations with slang. However, slang itself is not sufficient to indicate sexual harassment⁵.
- **Racial harassment** targets race and ethnicity characteristics of a victim such as color, country, culture, faith, and religion⁶.
- **Appearance-related harassment** is related to body appearance apart from sexuality. All dimensions of appearance are candidates, for example, hair style or looks. Fat shaming [1] and body shaming are critical sub-types.
- **Intellectual harassment** concerns intellectual power or the merits of individual opinion. Sub-types include level of formal education and grammar. Victims may in fact be intellectually gifted⁸.

³Observed statistics on January 8, 2018, from Pew research at <http://www.pewinternet.org/fact-sheet/social-media/>.

⁴<http://www.pewinternet.org/2017/07/11/online-harassment-2017/>

⁵<https://www.joshuafridmanesq.com/sexual-harassment.html>

⁶https://www.eeoc.gov/laws/types/sexual_harassment.cfm

⁷<https://www.joshuafridmanesq.com/racial-slurs-and-racial-harassment.html>

⁸<http://www.corrections.com/news/article/26649-ranking-bully-types-the-points-system>

- **Political harassment** relates to political views⁹, regarding issues under governmental influence such as global warming, the opioid epidemic, immigration or gun control. Typical targets are politicians and politically active individuals¹⁰.

The absence of a quality, annotated corpus of online harassment impedes comparative research on harassment detection. Our work [7] pioneers the content-specific study of cyberbullying. We publish here (i) our annotated content-specific lexicon and (ii) our content-specific annotated corpus validated by inter-rater reliability statistics. This paper is organized as follows: Section 2 reviews the related work. Section 3 explains the process for developing the five content-specific lexicons. In Section 4, we present the strategies for collecting and annotating our corpus. Section 5 provides examples of harassing as well as non-harassing tweets. We close with concluding remarks and our future plans.

2 RELATED WORK

Cyberbullying refers to the use of abusive language in social media or online interactions. While the majority of the prior research focuses on methods for detecting cyberbullying, there is no standard benchmark to evaluate and compare the performance of the existing approaches. The publicly available Golbeck corpus [3] contains 25,000 unique tweets with the binary annotation labels (i.e., harassing H or non-harassing N). There, authors use harassment hashtags such as #whitegenocide, #fuckniggers, #WhitePower, and #WhiteLivesMatter as crawling seeds. Human judges annotate the tweets using the binary labeling scheme.

Another harassment related dataset [8] focuses on racism and sexism. This dataset was collected during two months when the authors manually identified related hateful terms targeting groups based on aspects such as, ethnicity, sexual orientation, gender, and religion. Another corpus [2] distinguished between cyberbullying and cyber-aggression. Collection occurred from June 2016 till August 2016 with snowball sampling. This dataset contains 9,484 tweets from 1,303 users. Crowdsourcing workers labeled tweets according to four categories: 1) bullying, 2) aggressive, 3) spam, and 4) normal.

With respect to the methods for detecting harassment, [4] predicts cyberbullying incidents in Instagram. They extract features from text content, and the neighboring network along with temporal attributes to feed the predictive model. Another approach employed in [9] detects harassment features using content, sentiment, and context. These contextual features extracted from discussion and conversation improve the accuracy of harassment detection. Extending the context focus, [6] applies machine learning for detecting harassers and victims in a given cyberbullying incident. Their method considers social connections and infers which participants tend to bully and which participants are victimized. Their model is based on the connectivity of the users (network), the user interactions and the language of the active users.

Category	Count	Example
Sexual	453	assfuck, ball licker, finger fucker, Anal Annie, ass blaster
Appearance-related	15	assface, dickface, fatass, fuckface, shitface
Intellectual	34	assbag, asshat, asshit, dickbrain, dumbbitch
Racial	168	assnigger, beamer, Bigger, mulatto, mosshead
Political	23	Cockmuncher, towelhead, dickwad, propaganda, demon
Generic	44	arsehole, cockknoker, dick, fucker, sextoy

Table 1: Lexicon Statistics and Examples.

3 COMPILING AN OFFENSIVE WORDS LEXICON

The identification of cyberbullying typically begins with a lexicon of potentially profane or offensive words. We created a lexicon (compiled from online resources^{11 12 13 14 15}) containing offensive (i.e., profane) words covering five different types of harassment content. The resulting compiled lexicon includes six categories: (i) sexual, (ii) racial, (iii) appearance-related, (iv) intellectual, (v) political, and (vi) a generic category that contains profane words not exclusively attributed to the five specific types of harassment. A native English speaker conducted this categorization. Table 1 represents the statistics and examples of offensive words in each category.

4 CORPUS DEVELOPMENT AND ANNOTATION

We employ Twitter as the social media data source because of its growing public footprint¹⁶. Although the size of a tweet is restricted to 140 characters, once we consider a more extensive aggregation of tweets on a specific topic, mining approaches reveal valuable insights. We utilized the first five categories of our lexicon as seed terms for collecting tweets from Twitter between December 18th, 2016 to January 10th 2017. Requiring the presence of at least one lexicon item, we collected 10,000 tweets for each contextual type for a total of 50,000 tweets. As shown in Table 2, nearly half of these tweets were annotated. However, the mere presence of a lexicon item in a tweet does not assure that the tweet is harassing because the individuals might utilize these words with a different intention, e.g., in a friendly manner or as a quote. Therefore, human judges annotated the corpus to discriminate harassing tweets from non-harassing tweets. Three native English speaking annotators determined whether or not a given tweet is harassing with respect to the type of harassment content and assigned one of three labels “yes”, “no”, and “other”. The last label indicates that the given tweet either does not belong to the current context or cannot be decided. Finally, we can conclude 75,000 annotation work had been done totally.

¹¹ <http://www.bannedwordlist.com/lists>.

¹² <https://www.cs.cmu.edu/~biglou/resources>.

¹³ <http://www.noswearing.com/dictionary>.

¹⁴ <http://www.rsd.org/races#iranians>.

¹⁵ <http://www.macmillandictionary.com/us/thesaurus-category/american/offensive-words-for-people-according-to-nationality-or-ethnicity>.

¹⁶ Twitter reports 313 million monthly active users that generate over 500 million tweets per day <https://about.twitter.com/company>.

⁹ <http://www.brighthub.com/office/career-planning/articles/89787.aspx>

¹⁰ <https://www.performancecreate.com/political-discrimination-harassment/>

Contextual Type	Annotated Tweets	no. ✓	no. ✗
Sexual	3855	230	3619
Racial	4976	701	4275
Appearance-related	4828	678	4150
Intellectual	4867	811	4056
Political	5663	699	4964
Combined	24189	3119	21070

Table 2: Annotation statistics of our categorized corpus.

Agreement Rate. Although the annotators employed three labels, i.e., “yes”, “no”, and “other”, the eventual corpus excluded all of the tweets that did not have a consensus label of “yes” and “no”. In other words, the corpus only contains tweets that receive at least two “yes” or two “no” labels. Cohen’s kappa coefficient [5] measures the quality of our annotation by category in Table 3. The appearance-related context shows the highest agreement rate whereas the political and sexual contexts have the lowest indicating that these contents are more challenging to judge (ambiguity is higher).

Content Type	Agreement Rate
Sexual	0.70
Racial	0.84
Appearance-related	1.00
Intellectual	0.80
Political	0.69

Table 3: Agreement rate.

Comparison to Golbeck Corpus. The public state-of-the-art harassment-related corpus is Golbeck corpus [3] that only provides generic annotation, i.e., (i) harassing and (ii) non-harassing. This corpus contains 20,428 **non-redundant** annotated tweets of which only 5,277 are labeled as harassing. As we require a content-sensitive corpus, we created our own corpus. In the following, we present the principles and strategies employed in collecting, categorizing, annotating and preparing our corpus. We also categorized the Golbeck corpus according to our lexicon. It can be observed in Table 4 that more than 75% of harassing tweets are racial. This statistic confirms Golbeck’s observation. While this may be an accurate reflection of the base rate, our view is that different harassment content may have different consequence. An imbalanced corpus at the foundation of our research effort could result in misses of particular import to teenage mental health, concerning sexuality, appearance and intellect.

Contextual Type	#of Tweets
Sexual	380
Racial	4148
Appearance-related	145
Intellectual	381
Political	163
Non Harassing	41
Total	5277

Table 4: Statistics of Golbeck corpus after our annotation w.r.t. contextual type.

5 SAMPLES FROM OUR CORPUS

Below we provide some examples from our corpus, by content area. For each content area, we first show examples annotated as harassing with respect to the content in question. Then we show examples with similar content which are not annotated as harassing with respect to the content in question.

Sexual Harassing.

- @user: and you don't gotta pay none of ya bills baby ima do all that just don't fuck another nigga or ima shoot you
- to the dumbass bitch who tried opening my front door at 4 am nigga i'll kill you if i hear you again bro. i ain't a (URL)

Sexual non-harassing.

- make up is a form of art. i do not want to be a girl with real boobs or a vagina. i may want to do drag but two very different things.
- hot lesbian gets a pussy pounding with toy.
- three awesome teen babes licking each other pussies in absolute lesbian sex.

Appearance-related Harassing.

- @user @user we started killing you because our backs couldn't handle the weight of your fatass anymore.
- @user @user you dat skank postn pix of my girl sab-so show ya fuckin ugly greasy mug ya getto bitch

Appearance-related Non-harassing.

- think it's funny when girls finish a tweet with babes like they're talking down to them when really they're the same level of skank.
- competition time follow me & amp; retweet & amp; you can win a petite spaffnoshing oaty camel toe blotter out of my bin.
- @user: @user most insulting thing a skank can do to a woman who is worth having is mock her to a woman who isn't.É

Intellectual Harassing.

- @user what a complete disgrace of human u r.real cool wish death. no surprise from a washed up fucktard really__
- shoutout to dumb asses who go around clicking like or rt all the hot chicks posts no matter how stupid the shit it be.

Intellectual non-harassing.

- maybe this isn't sadness maybe this is just being a fuck-head __
- @user oh no i'm so sorry to hear that another one of your family members is a shithead
- @user is doing so well and his finally feeling happy we know his been lonely and we know how tired he is ? why can't our asses be happy.

Racial Harassing.

- @user shut the fuck up chink frog nigger.
- @user go back off private you chink.
- @user @user @user shut up you stupid paki.

Racial Non-harassing.

- @user do you know which exactly are the reasons for the police to release the paki? was it only cause of the 1 day period?!

- rt @user: coming up on gmb odious man-child @user interviews racist pathological lying assbat @user.
- @user 90% of paki names are islamic hence they are not in urdu. while urdu itself is a mixture/copy of other languages even in urdu

Political Harassing.

- # thanksdonald for getting rid off that assbat who has been president for 8 yrsb.
- @user how are u a jr high dickwad and president. a true leader doesn't taunt citizens who don't support him. pathetic. sad!
- @user: you're passive aggressive petty fuckbag who values a murderer fascist like putting over our own president. you're of.

Political non-harassing.

- @user yep and that's how the democrats do it. you know they pretend to know what their doing but really couldn't tell their asses.
- @user: liberals still continue to develop conspiracy theories in order to blame everyone else for having their asses hande f.
- @user those 4 trump supporters we're bad asses to jump 20 black lives. i call bs.

6 CONCLUSION AND FUTURE WORK

In this paper, we discussed the creation of a quality tweet corpus related to harassment and annotated that with respect to the five types of harassment content (i) sexual, (ii) racial, (iii) appearance-related, (iv) intellectual, and (v) political. This is the first corpus that takes content type into account. Furthermore, we have also developed a lexicon of content-specific offensive words along with a generic category of offensive words. We are making this dataset available to encourage comparative analysis of harassment detection algorithms. In future, we plan to employ this corpus for advancing our research on studying harasser and victim language.

7 ACKNOWLEDGEMENT

Thilini Wijesiriwardene assisted in the preparation of the corpus. We acknowledge support from the National Science Foundation (NSF) award CNS 1513721: Context-Aware Harassment Detection on Social Media. Any opinions, findings, and conclusions, recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

REFERENCES

- [1] Sofia Berne, Ann Frisén, and Johanna Kling. 2014. Appearance-related cyberbullying: A qualitative investigation of characteristics, content, reasons, and effects. *Body image* 11, 4 (2014), 527–533.
- [2] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. *CoRR abs/1702.06877* (2017). arXiv:1702.06877 <http://arxiv.org/abs/1702.06877>
- [3] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al. 2017. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 229–233.
- [4] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 186–192.
- [5] Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.
- [6] Elaheh Raisi and Bert Huang. 2017. Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the IEEE/ACM International Conference on Social Networks Analysis and Mining*.
- [7] Mohammadreza Rezvan, Saeedeh Shekarpour, Thirunarayan Krishnaprasad, Valerie Shalin, and Amit Sheth. 2018. Analyzing and Learning Language for Harassment in Different Contexts. In *Submitted to THE 12TH INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA (ICWSM-18)*.
- [8] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.
- [9] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB 2* (2009), 1–7.