PLOS ONE

RESEARCH ARTICLE

Analyzing and learning the language for different types of harassment

Mohammadreza Rezvan[1☯]*, **Saeedeh Shekarpour**[2☯]*, **Faisal Alshargi**[4], **Krishnaprasad Thirunarayan**[3], **Valerie L. Shalin**[3], **Amit Sheth**[5]

1 University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **2** University of Dayton, Dayton, Ohio, United States of America, **3** Knoesis Center, Wright State University, Dayton, Ohio, United States of America, **4** University of Leipzig, Leipzig, Germany, **5** AI Institute, University of South Carolina, Columbia, South Carolina, United States of America

☯ These authors contributed equally to this work.
* mohammadrezarezvan94@gmail.com (MR); sshekarpour1@udayton.edu (SS)

Citation: Rezvan M, Shekarpour S, Alshargi F, Thirunarayan K, Shalin VL, Sheth A (2020) Analyzing and learning the language for different types of harassment. PLoS ONE 15(3): e0227330. https://doi.org/10.1371/journal.pone.0227330

Editor: Kazutoshi Sasahara, Nagoya University, JAPAN

Received: August 17, 2018

Accepted: October 18, 2019

Published: March 27, 2020

Data Availability Statement: Data cannot be shared publicly because of Twitter data policy. Data are available from the Institutional Data Access / Ethics Committee (contact via Email) for researchers who meet the criteria for access to confidential data.

Funding: We acknowledge support from the National Science Foundation (NSF, https://www.nsf.gov) award CNS 1513721: Context-Aware Harassment Detection on Social Media. Dr. Amit Sheth, Dr. Krishnaprasad Thirunarayan, and Dr. Valerie Shalin received this award. Any opinions,

Abstract

THIS ARTICLE USES WORDS OR LANGUAGE THAT IS CONSIDERED PROFANE, VULGAR, OR OFFENSIVE BY SOME READERS. The presence of a significant amount of harassment in user-generated content and its negative impact calls for robust automatic detection approaches. This requires the identification of different types of harassment. Earlier work has classified harassing language in terms of hurtfulness, abusiveness, sentiment, and profanity. However, to identify and understand harassment more accurately, it is essential to determine the contextual type that captures the interrelated conditions in which harassing language occurs. In this paper we introduce the notion of contextual type in harassment by distinguishing between five contextual types: (i) sexual, (ii) racial, (iii) appearance-related, (iv) intellectual and (v) political. We utilize an annotated corpus from Twitter distinguishing these types of harassment. We study the context of each kind to shed light on the linguistic meaning, interpretation, and distribution, with results from two lines of investigation: an extensive linguistic analysis, and the statistical distribution of uni-grams. We then build type- aware classifiers to automate the identification of type-specific harassment. Our experiments demonstrate that these classifiers provide competitive accuracy for identifying and analyzing harassment on social media. We present extensive discussion and significant observations about the effectiveness of type-aware classifiers using a detailed comparison setup, providing insight into the role of type-dependent features.

Introduction

Disclaimer: *This article uses words or language that is considered profane, vulgar, or offensive by some readers. Owing to the topic studied in this article, quoting offensive language is academically justified but neither we nor PLOS in any way endorse the use of these words or the content of the quotes. Likewise, the quotes do not represent our opinions or the opinions of PLOS, and we condemn online harassment and offensive language.*

Although social media has enabled people to connect and interact with each other, it has also made people vulnerable to insults, humiliation, hate, bullying–facing threats from

findings, and conclusions, recommendations expressed in this material are those of the author (s) and do not necessarily reflect the views of the NSF. And the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

individuals who are either known (e.g., colleagues, friends) or unknown (e.g., fans, clients, anonymous entities). Please note that in this work, cyberbullying and harassment are used interchangeably. A Pew research study reports [1] that one-in-five (18%) victims of cyberbullying characterized their exposure as severe. The resulting negative impact from emotional distress, privacy concerns and threats to physical safety and mental health, affect individuals online and offline. This calls for tool-based, automatic detection, monitoring, and analysis of hurtful language to protect online users. The prior state-of-the-art is limited to detecting specific hurtful language such as hate speech [2], abusive language [3], and profanity [4], collectively termed Negative Affective Language (NAL). In the following, we present the definitions and terms for variants of harassing language:

- **Hate speech** is "speech that denigrates a person because of their innate and protected characteristics" [5]. Furthermore, it is divided into two categories: *directed* and *generalized*, depending upon whether there is an explicit target or not.
- **Abusive Language** is "the collection and misuse of private user information, cyberbullying and the distribution of offensive, misleading, false or malicious information" [6].
- **Offensive Language** employs profanity, is strongly impolite, rude or vulgar red expressed with fighting or hurtful words to insult a targeted individual or group [7–10].
- **Aggressive Language** shows overt, angry and often violent social interaction with the intention of inflicting damage or other unpleasantness upon another individual or group of people [11, 12].
- **Harassing (Cyberbullying) Language** is the use of force, threat, or coercion to abuse, embarrass, intimidate, or aggressively dominate others. It typically denotes repeated and hostile behavior performed by a group or an individual [11–13].

These definitions are highly subjective and overlap, making them hard to differentiate. For example, the definition of harassing language is similar to aggressive language. We posit that all of these NALs are **hurtful** and thus **harassing**. But they might vary in their level of severity, presence or absence of target (victim), contextual interpretation and purpose. In this paper, we frame harassing language as *offensive language where a given post/message contains "profanity, strongly impolite, rude, vulgar or threatening language*".

State-of-the-art harassment detection fails to exploit the **contextual type** of harassing language. Webster's dictionary [14] provides the following definition for context: "the parts of a discourse that surround a word or passage and can throw light on its meaning". Here, we describe the notion of contextual type as the linguistic or statistical conditions that help in differentiating the type of harassment. For example, the circumstance of a student who has been subjected to sexual harassment by her ex-partner differs from a student racially harassed because of her/his color. We suggest that *contextual type* influences the linguistic characteristics of harassment. We propose five contextual types of harassment in online communication on social media: (i) sexual harassment, (ii) racial harassment, (iii) appearance-related harassment, (iv) intellectual harassment, and (v) political harassment. This categorization is represented in Fig 1. Below, we define each type of harassment using illustrative examples from the Twitter corpus we have created.

1. *Sexual harassment* is offensive sexual speech that usually targets females. E.g., the harasser might comment on the victim's body in a vulgar manner or mention sexual relationships in an aggressive way. Note that using sexually profane words is not sufficient to indicate offensive sexual harassment [15, 16].

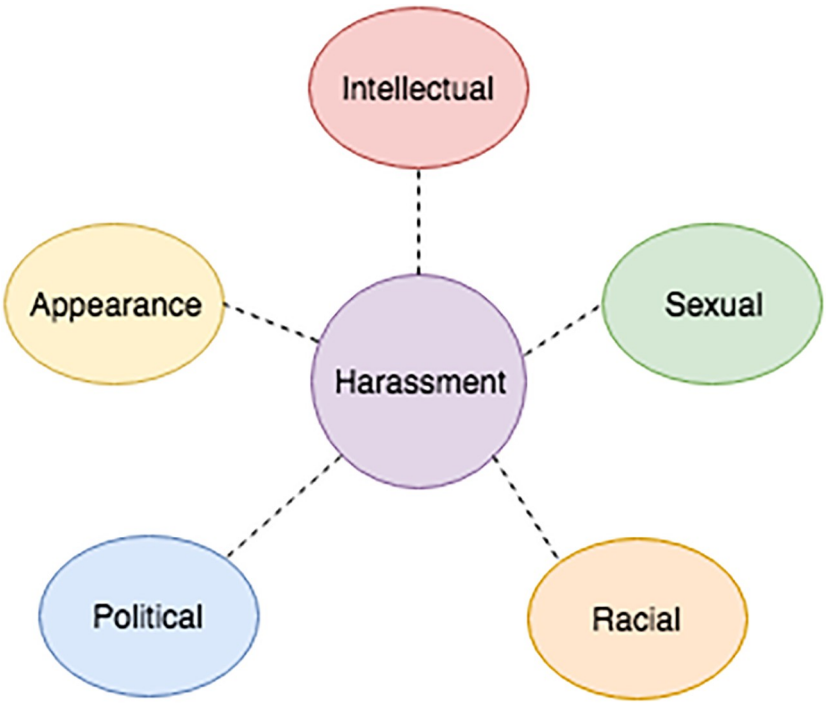

Fig 1. Five contextual types of harassment.

https://doi.org/10.1371/journal.pone.0227330.g001

2. *Racial harassment* targets race and ethnicity characteristics of a victim such as skin color, country of origin, culture, or religion, in an offensive manner [17].
3. *Appearance-related harassment* uses embarrassing language referring to body appearance. Fat shaming [18] and body shaming are key subtypes of this type of harassment.
4. *Intellectual harassment* offends the intellectual power or opinions of individuals [19].
5. *Political harassment* is related to someone's political views [20]. Typical targets are politicians and politically inclined individuals who receive threatening messages [21].

Determining the real intent behind a tweet regarding the type of harassment can have serious implications for public perception. Reliable assessment of the type of harassment can have significant repercussions. We are unaware of any prior work on studying harassment concerning these five types. We summarize our contributions as follows: (i) We introduce five contextual types of harassment. Then, we provide a systematic, and comparative analysis to assess offensive language from linguistic and statistical perspectives for each contextual type. This allows us to exploit relevant features for developing classifiers to identify these critical types of harassment on social media. (ii) We develop type-aware classifiers and capture their effectiveness using a detailed comparative study. This paper is organized as follows. The next section reviews the related literature. We then present the type-aware corpus that we have developed. Subsequently, we analyze our compiled corpus linguistically as well as statistically, which shows us the significant type-specific features for various types of harassment. We then discuss supervised learning approaches and classifiers for detecting the harassing language in comparative settings. We also provide an error analysis study regarding the pitfalls and challenges of our strategy. We close with the conclusions and our future plans.

Table 1. Summery of the related research.

Paper	Goal	Data	Conclusions
[7]	Detecting offensive and hateful speech language	85.4 Million Tweets Collected from 33458 twitter user using profane words. 25000 tweets are selected.	Collected discriminating terms for hate speech and offensive language
[11]	Detecting aggressors and their behavior on social media	1.6 million tweets collected in 3 months, using crowd sourcing for annotation.	Determined that posts of aggressor profiles are more negative
[8]	Detecting offensive language and identifying its sender.	The data set includes comments from 2,175,474 Youtube users in reaction to the top 18 videos on different Topics.	(i) Conceptualized offensive content, and (ii) enhanced features using lexical, style, structural, and context-specific features.
[13]	Predicting cyberbullying incidents on Instagram social media	41K users that are cyber bullied according to the random seed nodes. 3165K tweets collected from 25K public users while 697K Tweets labeled as profane tweets	Classifier designed, trained, and applied for collecting data. Logistic regression classifier
[22]	Detecting harassment based wrt. content, sentiment, and context	∼11K tweets used in experiments Fundacio'n Barcelona Media (FBM): Kongregate, Slashdot and MySpace. Totally 10,951 tweets collected and nearly 167 labeled offensive.	Improving accuracy in detecting harassing language using discussion-style and chat-style language
[23]	Detecting harassers and victims in cyberbullying incidents	Collected twitter data using profane words Twitter data contains 180,355 users and 296,308 tweets.	Accuracy improved wrt. network features.
[24]	Detecting instigators and victims of bullying	180K profile on Twitter and ∼300K tweets using profane words as seed	scoring level of cyber bully and victim.
[25]	Detecting cyber bullying in the Japanese community.	Data from Japanese secondary schools	Automatically extract new vulgarities from the Internet to keep their offensive lexicon up to date.
[9]	Understanding behavior and actions of individuals using emotion detection	∼2.5M tweets	tweets dataset using harassment-related and emotion hashtags
[35]	Detecting bullying incident on social networks	∼2M tweets collected in 4 weeks	Developed a practical method of text mining, clustering, dimensionality reduction and classification.
[26]	Classifying cyberbullying activities on social network	Collected data from 18,554 users data from Formspring.Me and MySpace.	predicting cyber bullying using fuzzy logic
[31]	investigating the correlation of harassment on Facebook	555 Facebook users in the United States (59% female; Mage = 30.90, SDage = 9.19)	Results show most of the updating posts related to the intellect people, children, and who they are in the romantic relation.
[32]	Identifying narcissism, activities on Facebook social media	256 Facebook users from locations around the world.	Text mining for narcissistic using on the comment likes
[36]	Automatic cyber bulling detection on social media text	English and Dutch corpora from *ASKfm*social media sites.	detecting signals of cyber bulling on social media, about bullies, victims, and bystanders.
[37]	Decompose the overall detection problem into detection of sensitive topics, lending itself into text classification.	corpus contain 4500 YouTube comments.	Concluded binary classifier for individual labels outperform multiclass classifier.
[38]	Cyberbulling detection with in multi modal content.	∼K entries from Instagram and Vine Dataset	proposed cyberbulling detection framwork **XBully** based on network representation leaning.
[39]	Identification of fake content in online news.	980 entries from fakeNewsAMT and celebrity Dataset	Linguistic analysis shows the importance of the lexical, syntactic, and semantic of content.

https://doi.org/10.1371/journal.pone.0227330.t001

State-of-the-art in harassment research

The previous research studies targeted various social media sources such as Twitter, Instagram, and Facebook. In Table 1, we summarize the prior literature with their corresponding goals, conclusions, and underlying data sets. Here, we specifically note particularly prominent related work. In [13], the authors seek to predict cyberbullying incidents on Instagram. They built a predictive model for the incidence of cyberbullying using features from initially posted data, a social graph, and temporal properties. The work in [22] proposed an approach for detecting harassment features based on the content, sentiment, and context. Using Slashdot and MySpace data, they showed significant improvement using TFIDF supplemented with

sentiment and contextual features. The authors of [23] proposed an approach to spotting harassers as well as victims on social media. They considered the social structure and infer which user is a likely instigator and which user is expected to be a victim. This model is based on social interactions and the language of users in social media. Similarly, [24] proposes a method that simultaneously discovers instigators and victims of bullying incidents. It extends an initial bullying vocabulary using twitter and ask.fm. In [25], the authors proposed a supervised learning method for detecting cyberbullying in Japan. In [26], the authors propose a supervised learning method based on *fuzzy logic* and *genetic algorithm* to identify the presence of cyberbullying terms and classify activities, such as flaming, harassment, racism, and terrorism on social media. Fuzzy rules were used to classify data, and a genetic algorithm was used for optimizing the parameters.

[9] explores the correlation of behaviors and actions of people and their emotions. The authors developed a large emotion-labeled dataset of harassing tweets. They applied 131 emotion hashtag keywords categorized into seven groups and collected 5 million tweets. To find useful features for emotion identification, they applied LIBLINEAR [27] and Multinomial Naive Bayes [28] algorithms. They extracted n-gram features [29] to analyze the emotion, and they applied *Linguistic Inquiry and Word Count (LIWC)* to expand the feature set with related emotional words. Interestingly, the authors of [12] target cyber-aggression and cyberbullying in a multi-modal context with text comments and media objects on Instagram. They concluded that non-text features are not able to substantially improve the performance of cyberbullying detection compared to text-based features.

Different from the previous work, some literature examines the **psychological implications** of harassment incidents [30]. The authors in [31] sought the reasons behind the updates of posts on Facebook. They noticed that: (i) the majority of posts are about social activities and everyday life, (ii) people with low self-esteem updated their status on relationship whereas those with high self-esteem update their status with respect to their children. Moreover, people with narcissistic personality disorder updated their status through their achievements. Furthermore, they observed a correlation between the number of likes and comments with esteem level of people (e.g., the people with the low self-esteem receive fewer likes and comments because their status expresses greater negative affect). Similarly, the authors of [32] discuss narcissism personality disorder in Facebook users and its implications in harassing incidents. Our own past work [33, 34] focused on (i) using a conversation between a sender and a receiver to better capture its normal linguistic nature (e.g., base rates for curse word usage) and the nature of the relationship between participants (e.g., friends vs. strangers), and (ii) analyze comments/review threads to better identify offensive content in non-text media such as YouTube videos [34], to reliably detect harassment between participants.

Type-aware harassment corpus

We published a type-aware annotated corpus and lexicon in [40]. Our corpus consists of 25,000 annotated tweets for the five types of harassment content and is available on the Git repository [41]. In the following, we discuss our strategies for corpus compilation and annotation. The identification of cyberbullying typically begins with a lexicon of potentially profane or offensive words. We created a lexicon (compiled from online resources [42] [43] [44] [45] [46]) containing offensive words covering five different types of harassment context. 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.

Corpus development and annotation

We employ Twitter as the social media data source because of its extensive public footprint. Twitter reports 313 million monthly active users that generate over 500 million tweets per day [47]. Although the size of a tweet is restricted (140 characters at the time of corpus collection), 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 API between December 18th, 2016 to January 10th 2017 [47] (This date was close to the US presidential election. Then our political sub-corpus has many tweets with the subject of Trump). 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 who were the undergraduate students with a major in computer science and minor in psychology or sociology were employed for our annotation task. The annotators determined whether or not a given tweet was 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. Ultimately, we acquired ≈24,000 annotated tweets represented in Table 2. Note that the annotation task was done on a per tweet basis although it can be improved using the entire conversation history.

Agreement rate

Although the annotators employed three labels, i.e., *yes*, *no*, and *other*, the eventual corpus excluded all tweets without a consensus label of "yes" or "no". That is, the corpus contains only those tweets that received at least two "yes" or two "no" labels. Cohen's kappa coefficient [35] measures the quality of annotation by category in Table 3. The appearance-related context shows the highest agreement rate whereas political and sexual contexts have the lowest, indicating that they are more challenging to judge due to higher ambiguity.

Annotating Golbeck corpus

The public state-of-the-art harassment-related corpus is the Golbeck corpus [40] 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. Since we require context-aware annotations, we re-annotated the harassing tweets of Golbeck. The agreement rate (Cohen's kappa) between the two annotators is 86%. As shown in Table 4,

Table 2. Annotation statistics of our categorized corpus.

Contextual Type	Annotated Tweets	Harassing ✓	Non-Harassing ✗
Sexual	3,855	230	3,619
Racial	4,976	701	4,275
Appearance-related	4,828	678	4,150
Intellectual	4,867	811	4,056
Political	5,663	699	4,964
Combined	24,189	3,119	21,070

https://doi.org/10.1371/journal.pone.0227330.t002

Table 3. Agreement rate.

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

https://doi.org/10.1371/journal.pone.0227330.t003

more than 75% of the 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 contexts may have different consequence. An imbalanced corpus at the foundation of our research effort could result in misses of particular practical import to teenage mental health, concerning sexuality, appearance and intellect.

LIWC analysis for different types of harassment

Linguistic analysis of our corpus sheds light on the differences between the harassing corpus versus non-harassing corpus for each type. Furthermore, it provides a comparison between various types of harassment. We divided our corpus into 12 sub-corpora: (i) *one generic corpus* containing all harassing tweets regardless of their type, called the **combined harassing corpus**, (ii) *one generic corpus* containing all non-harassing tweets irrespective of the type called the **combined non-harassing corpus**, (iii) *five contextual type-aware corpora* including only harassing tweets per type, (iv) *five contextual type-aware corpora* including only non-harassing tweets per type. For linguistic analysis, we utilized LIWC [48] [49]. This tool tallies 96 linguistic features using a multiword lexicon for each feature. We individually analyzed each of the 12 sub-corpora using LIWC. An effect size, statistic estimates the magnitude of an effect (e.g., mean difference, regression coefficient, Cohen's *d*, and correlation coefficient) [50] metric was used to determine significant discriminators [51]. Conventionally, a proportion (feature) f_i is considered moderately discriminating when its effect size is more than 0.5 (i.e., $|e_{f_i}| > 0.5$), and is considered unhelpful if $|e_{f_i}| \approx 0$. The effect size for each feature is calculated as follows:

$$e_{f_i} = \frac{\overline{experimentalgroup} - \overline{controlgroup}}{std} \quad (1)$$

where, $\overline{experimentalgroup}$ is the mean of the experimental group on the given feature f_i, $\overline{controlgroup}$ is the mean of the control group wrt. the given feature f_i and *std* is the standard deviation. For each content corpus as well as for the combined corpus, we consider the

Table 4. Statistics for the Golbeck corpus after our annotation wrt. contextual type.

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

https://doi.org/10.1371/journal.pone.0227330.t004

harassing corpus as the experimental group and the non-harassing corpus as the control group. We compared the prevalence of the 96 LIWC features in the harassing corpus to their prevalence in the corresponding non-harassing corpus. Out of the 96 original features, we removed features that were not significant in any of the contextual types and retained 38 of the most discriminating features as shown in Fig 2. The extreme red (green) color represents significance (regarding effect size) of the corresponding feature in the harassing (non-harassing) corpus. In the following, we highlight specific significant features to make three points. First, a feature is often diagnostic of the *non-harassing* corpus. Second, feature significance is type dependent. The third is related to both points: a given feature, such as "you", can be a positive indication of harassment for one type and a negative indication of harassment for another. In the following, we indicate **highly significant linguistic features** derived from Fig 2 for each individual type. Note that our corpus is already biased towards curse words because curse words are present as seeds for crawling. Thus, our observations on discriminatory features are conditional on a "high recall curse word-laden corpus".

Sexual corpus

The pronoun *"I"* is prevalent in the sexually non-harassing corpus with $e = -1.2$, which is highly significant, e.g., `i'm le***an kiss`. Furthermore, the feature *"MONEY"* is prevalent in the harassing corpus with $e = 2.9$. E.g., `send me free money b**ch h*es i won't give you anything to dance to you h*e a*s industry b*t*h d*cks**king p*r*star people`.

Racial corpus

The pronoun *"YOU"* is prevalent in the harassing corpus with $e = 0.9$, e.g., `Vishalp sikanda, Quideazam hahahaha u p**i can block u cant debat u p**i I***an.` The *"COMPARATIVE"* feature is prevalent in racial non-harassing corpus with $e = -0.84$, e.g., `save block p**i like po yung comment ni richard fronda` (the word 'like' is an indicator of comparison in LIWC). Thus, these features can be used to discriminate between harassing and non-harassing tweets.

Political corpus

The pronoun *"SHE"* and *"HE"* with $e = -0.9$ and the pronoun *"WE"* with $e = -0.8$ are prevalent in the non-harassing corpus, e.g., `realdonaldtrump putin a**hat just like word can express displeasure leader god help us` (us indicates the pronoun 'WE'). The *"RISK"* feature is significant in non-harassing with $e = -1.9$, e.g., `f*** wrong democratic senators`. The word *'wrong'* represents a risk feature in LIWC dictionary. Other sample risk related words are *'danger'*, *'doubt'*, etc. Furthermore, the *"ANXIETY"* feature with $e = -0.92$ is significant in the non-harassing corpus. E.g., `well i'm true dumb f*** democrat wouldn't doubt`.

Appearance-related corpus

"NEGATION" with $e = 2.3$ is prevalent in the harassing corpus (probably because of the negative language used for referring to the body and appearance-related subjects). E.g., `Taylor swift cant shake c**el toe.` The other significant feature in the harassing corpus is the "PAST TENSE". E.g., `Ugli a** didn't go run yesterday get work f*t**s.` Furthermore, the *"COMPARATIVE"* feature is prevalent in appearance-related

		Appearance	Inttellectual	Political	Racial	Sexual	Combined
Linguistic Dimension	I	0.00	0.00	0.00	0.00	-1.21	0.00
	We	0.07	0.29	-0.89	-0.06	0.00	0.19
	You	0.00	0.12	0.00	0.93	0.00	-0.80
	She/he	-0.16	-0.38	-0.98	-0.49	0.00	0.54
	They	-0.43	0.00	0.00	0.00	-1.00	0.00
	Ipron	-0.54	-0.59	0.13	0.30	-0.12	0.17
	Conj	-0.90	0.58	0.45	0.06	0.16	-0.06
	Negation	2.34	0.17	0.24	-0.31	0.00	-0.16
Grammer Dimension	Comparison	0.63	0.00	0.15	-0.85	0.55	-0.07
	Interrogtives	-0.98	-0.56	-0.27	-0.06	0.22	0.48
	Number	-0.57	0.00	0.45	0.22	1.35	-0.08
	Quantifier	-0.94	0.07	-0.23	0.48	0.53	-0.08
Feel	Anxeity	-0.37	0.21	-0.93	-0.15	0.35	0.31
	Sad	-0.81	0.28	-0.57	0.02	0.69	0.19
Psychology	Family	0.00	0.60	1.84	0.24	0.34	-0.34
	Friend	0.70	0.59	0.00	0.23	1.97	-1.00
	Female	0.00	2.33	0.00	0.00	0.00	0.00
Emotion	Insight	0.76	0.41	0.04	-0.02	0.08	-0.29
	Cause	0.00	0.00	-2.38	0.75	0.00	0.00
	Discrep	0.00	0.00	0.00	-0.52	0.00	-1.52
	Tentat	-1.11	-0.18	-1.00	-1.50	0.60	0.66
	Certain	-0.96	-1.03	-0.44	-0.09	0.82	0.45
Perceptual	Differ	-1.01	1.20	0.00	0.00	0.25	0.00
	See	0.03	-0.16	-0.51	-0.30	-0.36	0.24
	Hear	-1.58	0.30	-0.90	-2.06	0.86	0.00
	Feel	-0.66	0.00	0.00	0.25	0.58	1.24
Personal Activity	Health	0.00	1.23	-0.13	-0.47	-0.19	0.42
	Ingest	0.00	1.51	1.38	0.81	-0.52	0.00
	Risk	-0.91	-1.37	-1.99	-0.13	0.36	0.81
	Past	2.52	-0.02	-0.09	-0.37	0.00	-0.03
Personal Concern	Future	0.00	0.00	0.00	-0.57	0.00	0.40
	Motion	0.53	-0.02	0.09	-0.19	0.07	-0.06
	Home	-0.71	0.13	0.93	-0.21	0.73	-0.09
	Money	0.00	0.00	0.00	1.01	2.96	0.56
	Religion	0.32	-0.10	0.21	0.00	0.60	0.32
	Assent	-0.85	0.07	0.43	0.03	-0.12	0.05
	Nonflu	-0.66	0.14	0.08	0.06	0.00	0.06
	Filler	-0.94	-0.12	1.12	0.32	0.00	0.22

Fig 2. Significant LIWC features in comparing harassing corpus to non-harassing corpus for six categories. The extreme red (green) color indicates the significance of a given feature in the harassing corpus (non-harassing corpus). E.g. the negation feature with the value 2.34 in the appearance harassing corpus is significantly higher than non-harassing corpus. The white color indicates a lack of difference for a given feature when comparing two corpora.

https://doi.org/10.1371/journal.pone.0227330.g002

harassing corpus with $e = 0.63$. E.g., `hey lardass notice your look pizza perhaps like f*** salad a**hole.` The word *'like'* indicates a comparative feature.

Intellectual corpus

The *"FEMALE REFERENCE"* feature with $e = 2.3$ is highly significant in intellectual harassing corpus (perhaps because girls are harassed more wrt. intellectual issues.) E.g., `She is dumb f***.`

Combined corpus

"DISCREPANCY" with $e = -1.5$ is prevalent in the non-harassing corpus e.g., `boss brought drunken sugar cook explain there alcohol just sh**face.`

Statistical analysis of different types

We investigate the relationship between the offensive words employed in collecting our corpora and the specific lexical items in the crawled corpora. We determine **Q1**: whether or not offensive words are observed as frequent words, **Q2**: whether or not the frequent words in harassing corpora differ from those in non-harassing corpora, and **Q3**: whether or not frequent words are type-sensitive, in other words, whether the frequent words vary with type of context. Fig 3 shows the 2D visualization of the word embeddings of the top-25 most frequent words for the harassing corpora, whereas Fig 4 represents a similar display for the top-25 most frequent words for the non-harassing corpora (the following section presents the details of word embedding). The prevalence of curse words in the non-harassing corpora is comparable to the harassing corpora. This confirms that the presence of curse words is not a sufficient indicator of harassment. In the following, we mention our key observations.

Key observations

Regarding Q1, as expected, we observed that offensive words are commonplace in both harassing and non-harassing corpora across types (cf. Figs 3 and 4). In addition, we observed some emerging, frequent offensive words, such as "grab" and "camel" that can now be added to our initial offensive lexicon [41]. Furthermore, there are frequent words that are not necessarily offensive. E.g., consider "look" or "eat" in the appearance-related type where they are implicitly related to the associated type, applicable to the appearance of a subject. Regarding Q2, we observe that the frequent words in the harassing corpora are different from those in the non-harassing corpora. The particular words in the harassing corpora also can be added to the initial lexicon of seed words. The result of this analysis can be utilized for weighting the severity of offensiveness for every single word included in our lexicon.

To reply quantitatively to Q3, we ran an annotation task on the top-15 most frequent words for each type of harassing corpus as well as the corresponding non-harassing corpus. The description of this task is as follows: we asked the human annotators (i.e., graduate students) to determine whether or not a given frequent word is related to the associated type either explicitly or implicitly. E.g., the words "eat" or "food" are implicitly related to appearance while they seem far from the type racial. The results of this exercise appear in Table 5. In the harassing corpora, the percentage of relatedness of words to the associated type is higher than 67% and in sexual and racial types, it even reaches 80%. This percentage fluctuates for non-harassing corpora. E.g., in appearance-related type, it is higher than 93% while in racial it reaches 53%. In sum, we conclude that the frequent words are mostly type-sensitive. Moreover, the

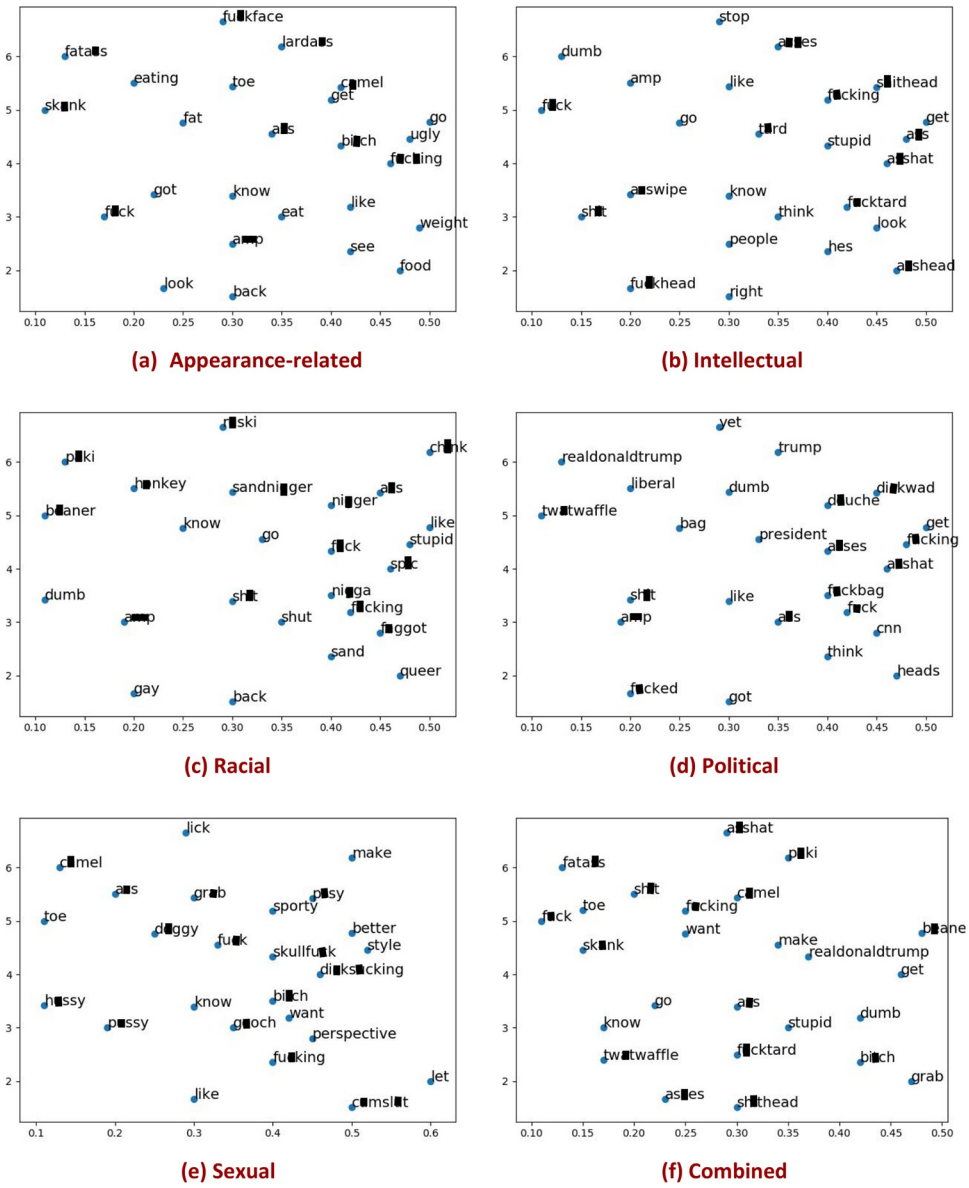

Fig 3. Top-25 frequent words within each harassing corpora.

https://doi.org/10.1371/journal.pone.0227330.g003

prevalence of apparently offensive language in the non-harassing corpus reinforces our claim that offensive language *per se* is not necessarily harassing.

One caveat is that the most frequent words appearing in the sub-corpus associated with each type are predominantly stop-words or curse words, as our initial seed terms are biased to an offensive lexicon. Ignoring these words, whose presence cuts across different types of harassment, revealed that the following prominent word groups are associated with various harassment types, shedding light on the possible features that may elicit harassment: (i) In the appearance-related harassment corpus, target words such as "eat", "ugly", "fat", "gym", and "weight", are present. (ii) In the intellectual harassment corpus, target words such as "dumb", "stupid", "work", and "head", are present. (iii) In the political harassment corpus, the target words such as "realdonaldtrump", "libtard", "dumb", "touch bag", "stupid", and "cnn", are

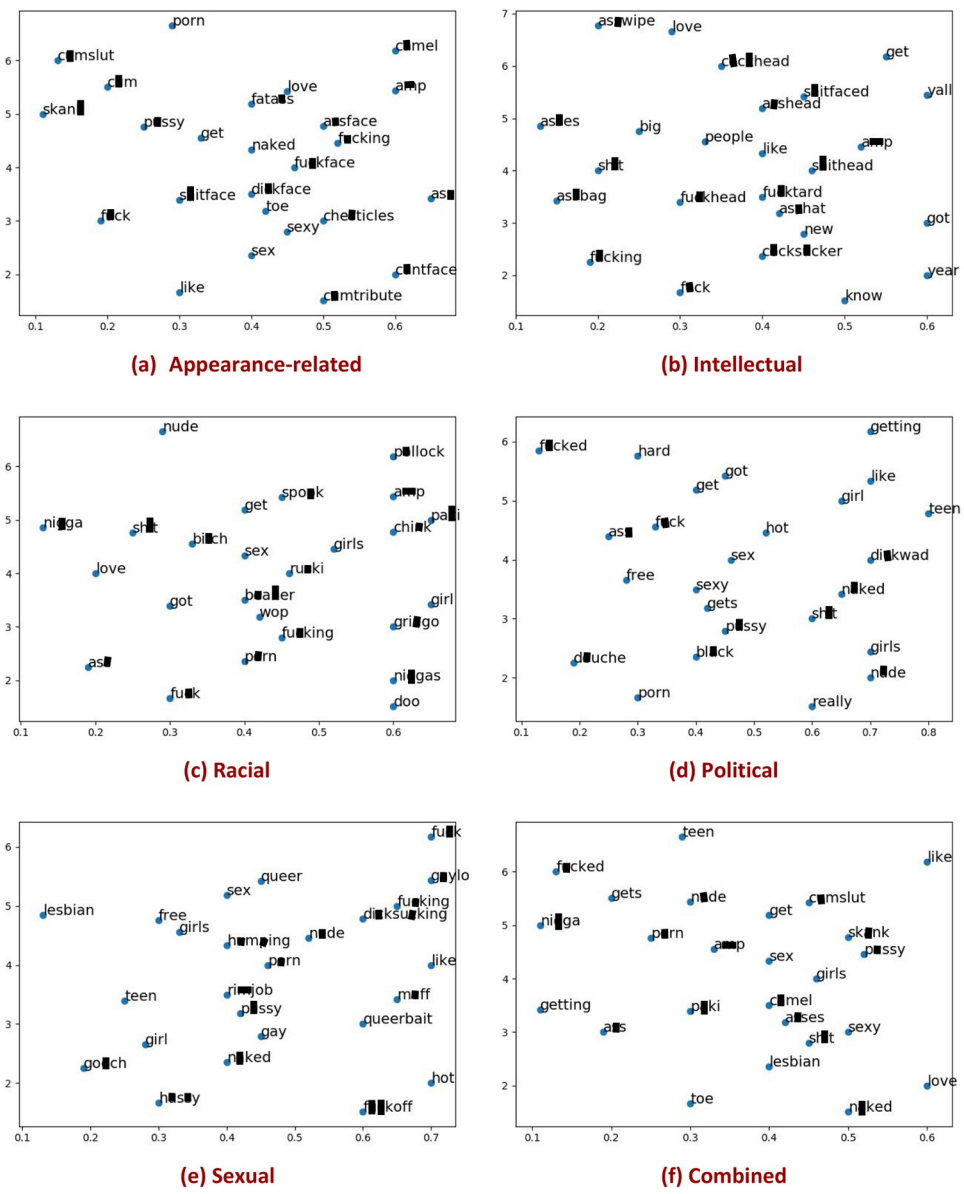

Fig 4. Top-25 frequent words within each non-harassing corpora.

https://doi.org/10.1371/journal.pone.0227330.g004

present. (iv) In the racial harassment corpus, target words such as "m*ki", "n**ger", "b**ner", "ch**k", "muslim", "i**ian", "moron", and "jew", are present. (iv) In the sexual harassment corpus, target words such as "hump", "hussy", "l**k", and "grab", are present.

Predicting different types of harassing posts

We aim to develop effective supervised learning methods to detect harassing language automatically and distinguish it from non-harassing language for each contextual type. The state-of-the-art contains various approaches for detecting harassing content from non-harassing content but not for discriminating the type of harassment. We approach this gap in two ways. The first is to build individual binary classifiers that identify a particular type of harassment,

Table 5. Percentage of type-dependent of top-15 frequent words within each sub-corpus. H stands for the harassing corpus and NH stands for the non-harassing corpus.

Category	Type	Percentage
Appearance-related	H	66.6%
	NH	93.3%
Intellectual	H	73.3%
	NH	73.3%
Political	H	80%
	NH	73.3%
Racial	H	80%
	NH	53.3%
Sexual	H	80%
	NH	60%

https://doi.org/10.1371/journal.pone.0227330.t005

e.g., a binary classifier that identifies only racial content or a binary classifier that classifies just offensive political content. The second approach uses the state-of-the-art methods to detect harassing language; after such recognition, we can employ a type-aware classifier to predict the associated type for that harassment incident. We implemented both approaches. Initially, we trained the individual classifiers for each type. In another approach, we built up a binary classifier that differentiates harassing content from non-harassing content regardless of their type. Note that any classifier from the state-of-the-art can substitute for this part. Then, we built up a multi-class classifier that predicts the type of harassment incident. The results of our experiments for both approaches reveal high accuracy. Furthermore, to verify the effectiveness of our classifier, we apply transfer learning by running our classifier on the Golbeck corpus and assess its performance for how successfully it predicts the type of harassment. In the following, we present the details of our experiments.

Transforming tweets to vectors

We utilized four approaches for transforming tweets to numerical representations (i.e., vectors): (i) the conventional vectorization approach TFIDF, (ii) word2vec, (iii) fastText and (iv) a LIWC vector. We feed our classifiers with each of these individual vectors or a combination of them.

The Term Frequency and Inverse Document Frequency (TFIDF). We use this approach [52] to transform each given tweet into a weighted vector T.

Distributional semantics (i.e., word2vec and fastText). Distributional semantics (so-called embedding models) [53] play a vital role in many Natural Language Processing (NLP) applications. They capture the semantics of text units (e.g., words, characters, tweets, paragraphs or documents) from the underlying corpus and represent them in a low dimensional vector space. We use two major embedding models for representing each tweet. The first one is word2vec [54] and the second one is fastText [55] [56]. The first one learns a dense representation at the unigram level and the second one learns at the character level. Both of these approaches have two models, i.e., skip-gram model and CBOW model [53, 57] that are roughly similar. The skip-gram model (CBOW model) computes the probability of the target word w_k (i.e. context word) appearing in the neighborhood of the context word w_i (i.e. target word), $\mathcal{P}(w_k|w_i)$. In this work, the vector representation of a tweet is computed as the concatenation of the vector of all tokens within the tweet. In the rest of this paper we rely on the following notations to specify a vector. W(S) and W(C) denote the low dimensional vector obtained

respectively by the skip-gram model and CBOW model of the word2vec approach. F(S) and F (C) denote the low dimensional vector obtained respectively by the skip-gram model and the CBOW model of the fastText approach. We compiled a corpus containing 15,999,557 sentences from the Twitter and Leipzig Collection Corpora [58] leveraging our offensive lexicon presented in [59] as the underlying seed words. Then, we trained the embedding models on this accumulated corpus using the learning parameters reported in [53, 57]. Our dimension size equals 300, the window size is 3, and the minimum count equals 10.

LIWC vector. The vector obtained by running the LIWC tool is denoted by L.

Evaluation of the harassment classifiers

Preparing training datasets. As the number of harassing tweets is not equal to the number of non-harassing ones in our corpus—in fact, it varies for each type—we prepared balanced datasets for training the classifier. We prepared five type-aware training data sets using an under-sampling approach taking all of the harassing tweets with an equal number of non-harassing (randomly sampled). Also, we prepared a combined training data set considering all of the harassing tweets regardless of their type and an equal number of non-harassing tweets. Table 6 shows the size of the training data sets for each type. Each data set contains an equal number of harassing tweets versus non-harassing tweets. Later, we employ the remaining tweets to test the robustness of the classifiers against unseen data.

Training binary classifiers. In our experimental study, we trained four types of classifiers, using (i) *Support Vector Machine* (SVM) [60], (ii) *K-Nearest Neighbors* (KNN) [60], (iii) *Gradient Boosting Machine* (GBM) [61], and (iv) *Naive Bayes* (NB) [60]. We rely on the following settings for the GBM classifier: the learning rate is 0.1, loss function is logistic regression, the number of trees is 100, sub-sample is 1.0, the criteria function is Friedman MSE, the minimum sample is 2, the minimum number of samples required to be at a leaf node is 1, and the maximum depth of the individual regression estimators is 3. We ran 10-fold cross-validation with re-sampling and iteration strategies (repeated five times). Fig 5 shows the performance of the classifiers based on an F-score measure using a TFIDF vector. Generally, the results of the NB classifier in all of the cases were inferior whereas the GBM classifier outperforms others in the majority of settings except for a few instances comparable to the SVM classifier. Thus, in the following experiments, we rely on the GBM classifier.

Feature engineering. To gain insight over the effectiveness of various features, we feed the GBM classifier with various feature settings. The fine-grained results of our experiment are listed in Fig 6. We employed a various combination of vectors, for example, F(S)+W(S) means the input features were the skip-gram models of fastText and word2vec. In addition to the typical precision, recall and F-score measures, we provide specificity (True negative rates) and accuracy rates. We offer the following observations: (i) The tweet representation using F(S) +W(S) vector is the most effective input representation as it provides high and balanced rates

Table 6. Size of the training datasets for each type.

Category	Number of tweets
Appearance-related	1,344
Intellectual	1,622
Political	1,397
Racial	1,401
Sexual	461
Combined	6,225

https://doi.org/10.1371/journal.pone.0227330.t006

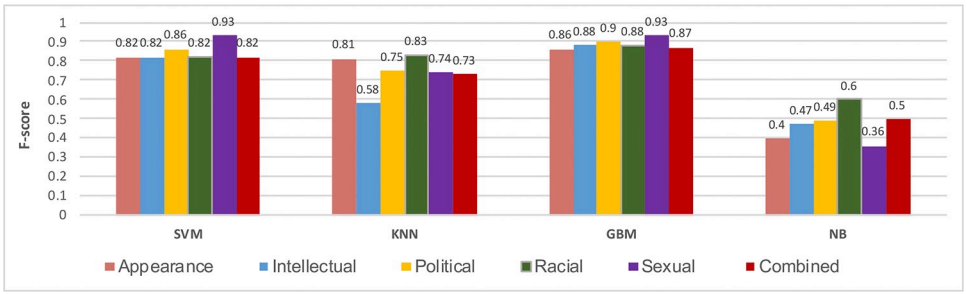

Fig 5. Comparative study of the F-score from four major classifiers i.e., SVM stands for support vector machine, KNN = K-Nearest Neighbor, GBM = Gradient Boosting Machine, NB = Naive Bayes, NN = Nueral Network).

https://doi.org/10.1371/journal.pone.0227330.g005

for all measures including precision, recall, F-score and specificity. Note that in multiple settings such as F(S)+L+T, precision, recall, and f-score are high whereas specificity is low meaning that the classifier is biased towards one of the classes and does not perform reasonably on both classes. (ii) In the settings for which the LIWC vector L is included, typically the specificity rate is low. This probably means L vector does not provide a discriminative representation for the classifier. (iii) Generally learning the representation of tweets using the fastText approach either with skip-gram or CBOW shows high performance. This might come from the fact that encoding tweets at the character level is more effective for detecting harassment. (iv) The sexual type resulted in the classifier with the highest accuracy (with F-score 96% and specificity 94%), racial and intellectual are in the next positions (respectively with F-score 88%, 86% and specificity 83%, 79%).

Binary classifier for harassment detection. We also trained a binary classifier on our combined corpus where it can differentiate the harassing language from non-harassing regardless of the contextual type. In situations that the type of harassment does not play a role, or type detection must occur after the harassment detection, using such a generic classifier is necessary. Table 7 shows the detailed results of this classifier in various settings of input features. Generally, the vector of FastText F shows an effective role, especially when it is coupled with the W vector; the specificity score reaches its optimum.

Type prediction using a multi-class classifier. Apart from building binary classifiers for predicting types, we trained a multi-class classifier to predict the type of harassment incidents. We trained several multi-class classifiers, among them the GBM classifier outperformed others. Herein we report the result for GBM classifier only. We used W(S)+F(S) vectorization approach as the input feature. Then, we trained this classifier on a corpus containing all of the sub-corpora from the previous step. This corpus has samples with six various labels where five labels indicate a particular type of harassment and the last label indicates "non-harassing" implying there is no harassing language. Table 8 shows the details of the evaluation on the performance of this classifier where the micro F-score is 0.92 and the macro F-score is 0.82. Note that in the macro-level, we calculate the performances of each class and then average whereas, in the micro-level, we calculate the performance for all classes, as computing contingency table and then evaluate precision/recall and F-score [62]. Digging into fine-grained efficiency shows that the accuracy across various classes holds similar behaviors except for a decrease in the precision and recall of the sexual type. As we will discuss in error analysis below, this type is prone to mis-classification with the other types particularly the racial type. However, comparing the performance of multi-class classifier and binary classifiers shows that the multi-class classifier mostly outperforms the binary classifiers by as much as ≈10%. Note that the accuracy of our classifier will improve on a generic tweet corpus because our current corpus has been crawled

	Feature	Precision	Recall	F-Score	Accuracy	Specificity
Apperance	T	0.85	0.89	0.86		
	T+L	0.89	0.92	0.9		
	F(S)+L+T	0.95	0.97	0.95	0.93	0.28
	F(C)t+L+T	0.94	0.93	0.90	0.88	0.19
	F(S)	0.80	0.83	0.81	0.79	0.74
	F(C)	0.77	0.76	0.75	0.73	0.70
	F(S)+L	0.94	0.92	0.89	0.87	0.28
	W(S)+L+T	0.94	0.96	0.93	0.91	0.30
	W(S)+L	0.95	0.97	0.94	0.92	0.28
	F(S)+W(S)	0.82	0.82	0.80	0.79	0.74
Intellectual	T	0.92	0.86	0.88		
	T+L	0.94	0.89	0.91		
	F(S)+L+T	0.94	0.93	0.90	0.88	0.35
	F(C)+L+T	0.91	0.89	0.85	0.82	0.37
	F(S)	0.82	0.83	0.81	0.79	0.74
	F(C)	0.79	0.84	0.81	0.79	0.73
	F(S)+L	0.94	0.92	0.91	0.89	0.76
	W(S)+L+T	0.94	0.93	0.91	0.90	0.72
	W(S)+L	0.93	0.93	0.91	0.89	0.70
	F(S)+W(S)	0.87	0.88	0.86	0.84	0.79
Political	T	0.91	0.9	0.9		
	T+L	0.94	0.94	0.94		
	F(S)+L+T	0.95	0.96	0.93	0.91	0.52
	F(C)+L+T	0.93	0.91	0.88	0.86	0.41
	F(S)	0.82	0.81	0.80	0.78	0.75
	F(C)	0.75	0.75	0.73	0.71	0.66
	F(S)+L	0.95	0.95	0.93	0.91	0.47
	W(S)+L+T	0.94	0.95	0.92	0.90	0.49
	W(S)+L	0.95	0.94	0.92	0.90	0.48
	F(S)+W(S)	0.84	0.83	0.81	0.79	0.73
Racial	T	0.77	0.69	0.72		
	T+L	0.87	0.84	0.84		
	F(S)+L+T	0.95	0.93	0.91	0.89	0.43
	F(C)+L+T	0.94	0.89	0.87	0.85	0.41
	F(S)	0.87	0.85	0.84	0.84	0.83
	F(C)	0.77	0.75	0.75	0.74	0.72
	F(S)+L	0.94	0.95	0.93	0.91	0.72
	W(S)+L+T	0.95	0.95	0.93	0.91	0.53
	W(S)+L	0.95	0.94	0.92	0.90	0.51
	F(S)+W(S)	0.90	0.89	0.88	0.87	0.83
Sexual	T	0.95	0.91	0.93		
	T+L	0.97	0.95	0.96		
	F(S)+L+T	0.98	0.98	0.98	0.97	0.53
	F(C)+L+T	0.97	0.97	0.96	0.95	0.56
	F(S)	0.87	0.83	0.81	0.78	0.67
	F(C)	0.78	0.80	0.77	0.73	0.61
	F(S)+L	0.97	0.95	0.94	0.93	0.71
	W(S)+L+T	0.97	0.96	0.95	0.94	0.59
	W(S)+L	0.98	0.97	0.96	0.95	0.82
	F(S)+W(S)	0.95	0.96	0.96	0.95	0.94

Fig 6. Comparative study of the various feature settings on the performance of the GBM classifier using measures such as precision, recall, F-score, accuracy, and specificity. The extreme colors, i.e., purple, yellow, green, olive, and pink show the higher values versus the white color that shows a lower value.

https://doi.org/10.1371/journal.pone.0227330.g006

Table 7. Performance of the GBM binary classifier on the combined corpus.

Feature	Precision	Recall	F-Score	Accuracy	Specificity
T	0.84	0.81	0.82		
T+L	0.9	0.87	0.88		
F(S)+L+T	0.94	0.92	0.90	0.88	0.37
F(C)+L+T	0.94	0.88	0.86	0.84	0.44
F(S)	0.83	0.83	0.82	0.80	0.75
F(C)	0.78	0.76	0.76	0.75	0.73
F(S)+L	0.94	0.95	0.93	0.91	0.69
W(S)+L+T	0.94	0.93	0.91	0.89	0.70
W(S)+L	0.93	0.94	0.92	0.90	0.74
F(S)+W(S)	0.90	0.89	0.88	0.87	0.83

https://doi.org/10.1371/journal.pone.0227330.t007

using curse words with a significantly higher proportion of harassing tweets compared to that in a generic tweets corpus, which is predominantly non-harassing and devoid of curse words. On the downside, it will miss harassment conveyed through "clean" words. However, to demonstrate the effectiveness of the current version of this classifier, in the next step we apply it on an unseen corpus to predict the type of harassment incident.

Comparison to the state-of-the-art. Since this work was the first to introduce contextual type for harassment, comparison to the state-of-the-art that relies only on two or three variants of harassment, is unfair. However, to verify the effectiveness of our type-oriented multi-class classifier, we tested it on the harassing tweets from the Golbeck corpus (an external corpus unseen to our classifier) that is a publicly available state-of-the-art harassment-related corpus [9]. This corpus contains 20,428 annotated tweets of which only 5,277 are labeled as harassing. It does not distinguish the nature of the harassment. In Table 4, we represented our annotations for the harassing tweets of the Golbeck corpus with respect to our types using human judges which yielded in an agreement rate of 86%. The proportion of harassing tweets per type is represented in the last column of Table 9. We ran our type-aware multi-class classifier (GMB classifier) to predict the associated type of harassing tweets on Golbeck corpus. Table 9 shows the precision, recall and F-score for each type. We observe an F-score of more than 94% for all types except for the type appearance. In the case of the racial type, the F-score reaches 98%. This high performance exceeds the state-of-the-art where they are mostly concerned about detecting the general harassing language (the reported accuracy ranges between 70% and 85%) [30, 63–65]. In addition, it shows robustness with unseen data. Note that the racial type is

Table 8. Performance of our multi-class classifier for predicting type of harassment incident.

Category	Precision	Recall	F-score
Appearance-related	0.84	0.85	0.84
Intellectual	0.87	0.85	0.86
Political	0.81	0.84	0.83
Racial	0.82	0.83	0.82
Sexual	0.58	0.62	0.60
Nonharassing	0.98	0.97	0.98
Micro Precision	0.92	**Macro Precision**	0.82
Micro Recall	0.92	**Macro Recall**	0.83
Micro F-score	0.92	**Macro F-score**	0.82

https://doi.org/10.1371/journal.pone.0227330.t008

Table 9. Performance of our classifier for predicting tweets for Golbeck corpus.

Category	Precision	Recall	F-score	Proportion Rate
Appearance-related	0.74	0.63	0.68	2.7%
Intellectual	0.91	0.92	0.91	7.2%
Political	0.90	0.95	0.92	3.0%
Racial	0.99	0.97	0.98	78.6%
Sexual	0.94	0.96	0.95	7.2%
Nonharassing	0.99	0.98	0.98	
Micro Precision	0.97		**Macro Precision**	0.91
Micro Recall	0.97		**Macro Recall**	0.90
Micro F-score	0.97		**Macro F-score**	0.91

https://doi.org/10.1371/journal.pone.0227330.t009

dominant in the Golbeck corpus. We also ran our classifier on a portion of 5,000 non-harassing tweets from Golbeck corpus, which resulted in the F-score > 98% (cf. Table 9). The last three rows of Table 9 show micro and macro precision, recall, and F-score. The closeness of the micro and macro measures shows that the classifier is not biased towards a dominant class.

Error analysis. To make sense of classifier errors, we examined a couple of tweets classified as sexual. E.g., for `@usr you deserved to be raped by a thousand Muslims in your c**t a**hole`, our classifier classified that as sexual harassment and not racial because of the word 'rape'. Similarly, the tweet `@usr @usr lol it's not against women. It's against f***ing feminist c***s like you. #feminazi #womenagainstfeminism` was classified as sexual. Such cases are ambiguous because even manual annotation is highly subjective. In other words, categorizing harassment is highly subjective and the boundary between types is not rigid. In majority of the overlapping cases (racial and sexual), the tweets were classified as sexual rather than racial. We also analyzed errors in political tweets and concluded that harassment signal can be: (i) implicit, e.g., `John Boehner blames Democrats for #shutdown. He better stop drinking cuz a few more drinks and he starts blaming the J*ws f`, (ii) ambiguous `??? You're a wh*** to the telecom industry, i hope your constituents vote you out.`, (iii) unreliable, e.g., `It's going to be a republican government in the US next term. Democrats can kiss their presidency bid goodbye. Let the J*ws rule!`, (iv) poorly captured through annotation, e.g., the tweet `@TrueNugget @FeministPeriod @Oregon-State Man college is becoming more and more a mistake.` in the Golbeck corpus. Our classifier misses them as they are weak cases of harassment.

Ethics

Our project involves analysis of Twitter data that is publicly available and that has been anonymized. It does not involve any direct interaction with any individuals or their personally identifiable data. So our work does not meet the Federal definition for human subjects research, specifically, "a systematic investigation designed to contribute to generalizable knowledge" and "research involving interaction with the individual or obtains personally identifiable private information about an individual". Thus, this study was reviewed by the Wright State University IRB and received an exemption determination.

Conclusion and future plans

In this paper, we introduced five contextual types for harassment, namely, (i) sexual, (ii) racial, (iii) intellectual, (iv) appearance-related and (v) political. We presented experiments with a

type-aware tweets corpus to analyze, learn, and understand harassing language for each type. Our contribution lies in providing a systematic and comparative approach to assessing harassing language from linguistic and statistical perspectives. Furthermore, we built type-specific classifiers, and the results of our experiments show the importance of considering the contextual type for identifying and analyzing harassment on social media.

In general, a single tweet identified as "harassing" may not provoke the same intense negative feeling that we associate with that word in the real-world scenario. However, in practice, "conversational" exchanges containing a sequence of such tweets can rise to the level of harassment causing mental and psychological anguish, and fear of physical harm. Nevertheless, our current Twitter dataset is limited to annotating single tweets in isolation for harassment. Furthermore, the reliable assessment of the type of harassment is a difficult problem because it requires significant knowledge of current events and common-sense. We plan to extend this work by learning the language of harassers as well as victims, and further study the contribution of non-verbal cues (i.e., conversational features, network features, and community features) for identifying online harassment activities, particularly on social media.

Acknowledgments

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.

Author Contributions

Conceptualization: Saeedeh Shekarpour.

Data curation: Mohammadreza Rezvan, Saeedeh Shekarpour, Faisal Alshargi.

Formal analysis: Saeedeh Shekarpour.

Funding acquisition: Krishnaprasad Thirunarayan, Valerie L. Shalin, Amit Sheth.

Investigation: Krishnaprasad Thirunarayan, Valerie L. Shalin, Amit Sheth.

Methodology: Mohammadreza Rezvan, Saeedeh Shekarpour, Faisal Alshargi.

Project administration: Saeedeh Shekarpour.

Resources: Faisal Alshargi.

Software: Mohammadreza Rezvan.

Supervision: Saeedeh Shekarpour.

Validation: Saeedeh Shekarpour.

Visualization: Faisal Alshargi.

Writing – original draft: Mohammadreza Rezvan, Saeedeh Shekarpour.

Writing – review & editing: Mohammadreza Rezvan, Saeedeh Shekarpour, Krishnaprasad Thirunarayan, Valerie L. Shalin, Amit Sheth.

References

1. DUGGAN M. Online Harassment 2017; 2017. Available from: https://www.pewinternet.org/2017/07/11/online-harassment-2017/.

2. Herz M, Molnár P. The content and context of hate speech: rethinking regulation and responses. Cambridge University Press; 2012.
3. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web. International World Wide Web Conferences Steering Committee; 2016. p. 145–153.
4. Mangaonkar A, Hayrapetian A, Raje R. Collaborative detection of cyberbullying behavior in Twitter data. In: 2015 IEEE international conference on electro/information technology (EIT). IEEE; 2015. p. 611–616.
5. ElSherief M, Kulkarni V, Nguyen D, Wang WY, Belding EM. Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. In: Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA; 2018. p. 42–51.
6. Talukder S, Carbunar B. AbuSniff: Automatic Detection and Defenses Against Abusive Facebook Friends. In: Twelfth International AAAI Conference on Web and Social Media; 2018. p. 385–394.
7. Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In: Eleventh International AAAI Conference on Web and Social Media; 2017.
8. Chen Y, Zhou Y, Zhu S, Xu H. Detecting offensive language in social media to protect adolescent online safety. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing. IEEE; 2012. p. 71–80.
9. Wang W, Chen L, Thirunarayan K, Sheth AP. Harnessing twitter "big data" for automatic emotion identification. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing. IEEE; 2012. p. 587–592.
10. Founta AM, Djouvas C, Chatzakou D, Leontiadis I, Blackburn J, Stringhini G, et al. Large scale crowdsourcing and characterization of twitter abusive behavior. In: Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA; 2018. p. 491–500.
11. Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A. Mean birds: Detecting aggression and bullying on twitter. In: Proceedings of the 2017 ACM on web science conference. ACM; 2017. p. 13–22.
12. Hosseinmardi H, Rafiq RI, Han R, Lv Q, Mishra S. Prediction of cyberbullying incidents in a media-based social network. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE; 2016. p. 186–192.
13. Hosseinmardi H, Rafiq RI, Han R, Lv Q, Mishra S. Prediction of cyberbullying incidents in a media-based social network. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE; 2016. p. 186–192.
14. Internet. Online Dictionary; 2017. Available from: https://www.merriam-webster.com/dictionary/context.
15. Lawyers NYED. Sexual Harassment; 2018. Available from: https://www.friedmanhouldingllp.com/sexual-harassment.html.
16. Commission UEEO. Sexual Harassment; 2018. Available from: https://www.eeoc.gov/laws/types/sexual_harassment.cfm.
17. Lawyers NYED. Racial Slurs and Racial Harassment; 2018. Available from: https://www.friedmanhouldingllp.com/racial-slurs-and-racial-harassment.html.
18. Berne S, Frisén A, Kling J. Appearance-related cyberbullying: A qualitative investigation of characteristics, content, reasons, and effects. Body image. 2014; 11(4):527–533. https://doi.org/10.1016/j.bodyim.2014.08.006 PMID: 25194309
19. Correlations. Ranking Bully Types; 2018. Available from: http://www.corrections.com/news/article/26649-ranking-bully-types-the-points-system.
20. Hub B. Guide to Dealing with Political Harassment in the Workplace; 2010. Available from: http://www.brighthub.com/office/career-planning/articles/89787.aspx.
21. Internet. Turns Out, There Is Political Discrimination and Harassment Too; 2018. Available from: https://www.performanceicreate.com/political-discrimination-harassment/.
22. Yin D, Xue Z, Hong L, Davison BD, Kontostathis A, Edwards L. Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB. 2009; 2:1–7.
23. Raisi E, Huang B. Cyberbullying detection with weakly supervised machine learning. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. ACM; 2017. p. 409–416.
24. Raisi E, Huang B. Cyberbullying identification using participant-vocabulary consistency. arXiv preprint arXiv:160608084. 2016.
25. Ptaszynski M, Dybala P, Matsuba T, Masui F, Rzepka R, Araki K. Machine learning and affect analysis against cyber-bullying. the 36th AISB. 2010; p. 7–16.

26. Nandhini BS, Sheeba J. Online social network bullying detection using intelligence techniques. Procedia Computer Science. 2015; 45:485–492. https://doi.org/10.1016/j.procs.2015.03.085
27. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. Journal of machine learning research. 2008; 9(Aug):1871–1874.
28. Mishne G, et al. Experiments with mood classification in blog posts. In: Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access. vol. 19; 2005. p. 321–327.
29. Tokuhisa R, Inui K, Matsumoto Y. Emotion classification using massive examples extracted from the web. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics; 2008. p. 881–888.
30. Raisi E, Huang B. Weakly Supervised Cyberbullying Detection Using Co-Trained Ensembles of Embedding Models. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE; 2018. p. 479–486.
31. Marshall TC, Lefringhausen K, Ferenczi N. The Big Five, self-esteem, and narcissism as predictors of the topics people write about in Facebook status updates. Personality and Individual Differences. 2015; 85:35–40. https://doi.org/10.1016/j.paid.2015.04.039
32. Younus A, Qureshi MA, Griffith J, O'Riordan C, Pasi G. A Study into the Correlation between Narcissism and Facebook Communication Patterns. In: 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). vol. 1. IEEE; 2015. p. 511–514.
33. Edupuganti V. Harassment Detection on Twitter using Conversations. Master Thesis, Department of Computer Science and Engineering, Wright State University, USA; 2017.
34. Kandakatla R. Identifying Offensive Videos on YouTube. Master Thesis, Department of Computer Science and Engineering, Wright State University, USA; 2016.
35. Sanchez H, Kumar S. Twitter bullying detection. ser NSDI. 2011; 12:15–15.
36. Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, et al. Automatic detection of cyberbullying in social media text. PloS one. 2018; 13(10):e0203794. https://doi.org/10.1371/journal.pone.0203794 PMID: 30296299
37. Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual cyberbullying. In: fifth international AAAI conference on weblogs and social media; 2011.
38. Cheng L, Li J, Silva YN, Hall DL, Liu H. XBully: Cyberbullying Detection within a Multi-Modal Context. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM; 2019. p. 339–347.
39. Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R. Automatic detection of fake news. arXiv preprint arXiv:170807104. 2017.
40. Golbeck J, Ashktorab Z, Banjo RO, Berlinger A, Bhagwan S, Buntain C, et al. A large labeled corpus for online harassment research. In: Proceedings of the 2017 ACM on Web Science Conference. ACM; 2017. p. 229–233.
41. Mohammadreza Rezvan SS. Harassment-Corpus; 2018. Available from: http://www.bannedwordlist.com.
42. Internet. Banned Word List; 2009. Available from: https://github.com/Mrezvan94/Harassment-Corpus.
43. von Ahn's Research Group L. Offensive/Profane Word List;. Available from: https://www.cs.cmu.edu/~biglou/resources/bad-words.txt.
44. Internet. Swear Word List;. Available from: https://www.noswearing.com/dictionary.
45. Users I. The Racial Slur Database; 1999. Available from: http://www.rsdb.org/races#iranians.
46. Users I. Macmillan Dictionary; 2009-2019. Available from: https://www.macmillandictionary.com/us/thesaurus-category/american/offensive-words-for-people-according-to-nationality.
47. Users I. Twitter; 2019. Available from: https://about.twitter.com/en_us/company.html.
48. Internet. LIWC2015; 2015. Available from: https://liwc.wpengine.com.
49. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015; 2015.
50. Coe R. It's the effect size, stupid: What effect size is and why it is important. 2002.
51. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological reviews. 2007; 82(4):591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x PMID: 17944619
52. Wu HC, Luk RWP, Wong KF, Kwok KL. Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS). 2008; 26(3):13. https://doi.org/10.1145/1361684.1361686

53. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held 2013, Lake Tahoe, Nevada, United States.; 2013. p. 3111–3119.

54. Google. Word2Vec; 2013. Available from: https://code.google.com/archive/p/word2vec/.

55. Facebook. FastText; 2013. Available from: https://fasttext.cc/docs/en/support.html.

56. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A. Advances in Pre-Training Distributed Word Representations. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, 2018.; 2018.

57. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. CoRR. 2013;abs/1301.3781.

58. Goldhahn ETQ D. Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In Proceedings of LREC 2012 (pp. 759-765); 2012.

59. Rezvan M, Shekarpour S, Balasuriya L, Thirunarayan K, Shalin VL, Sheth A. A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research. In: Proceedings of the 10th ACM Conference on Web Science. ACM; 2018. p. 33–36.

60. Mahdavinejad MS, Rezvan M, Barekatain M, Adibi P, Barnaghi P, Sheth AP. Machine learning for Internet of Things data analysis: A survey. Digital Communications and Networks. 2018; 4(3):161–175. https://doi.org/10.1016/j.dcan.2017.10.002

61. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002; 38(4):367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

62. Jurafsky D, Martin JH. Speech and language processing. vol. 3. Pearson London; 2014.

63. Marwa T, Salima O, Souham M. Deep learning for online harassment detection in tweets. In: 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS). IEEE; 2018. p. 1–5.

64. Zois DS, Kapodistria A, Yao M, Chelmis C. Optimal online cyberbullying detection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 2017–2021.

65. Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishra S. Detection of cyberbullying incidents on the instagram social network. arXiv preprint arXiv:150303909. 2015.