Analyzing and learning the language for different types of harassment

Mohammadreza Rezvan[1‡], Saeedeh Shekarpour[2‡],
Krishnaprasad Thirunarayan[1], Valerie L. Shalin[1], Amit Sheth[1],

1 Knoesis Center, Dayton, USA,{rezvan, tkprasad, valerie, amit}@knoesis.org
2 University of Dayton, Dayton, USA, sshekarpour1@udayton.edu

‡These authors contributed equally to this work.

Abstract

The presence of a significant amount of harassment in user-generated content and its negative impact calls for robust automatic detection approaches. This requires that we can identify different forms or 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 context that represents the interrelated conditions in which they occur. In this paper[1], we introduce the notion of contextual type to harassment involving five categories: (i) sexual, (ii) racial, (iii) appearance-related, (iv) intellectual and (v) political. We utilize an annotated corpus from Twitter distinguishing these types of harassment. To study the context for each type that sheds light on the linguistic meaning, interpretation, and distribution, we conduct two lines of investigation: an extensive linguistic analysis, and a statistical distribution of unigrams. We then build type-ware 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 major observations about the effectiveness of type-aware classifiers using a detailed comparison setup providing insight into the role of type-dependent features.

Introduction

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 either known (e.g., colleagues, friends) or unknown (e.g., fans, clients, anonymous ones) individuals. A Pew research center report[2] shows that 41% of Americans have personally experienced cyberbullying (e.g., offensive name-calling, shaming [3]). One-in-five (18%) victims 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 automatic detection, monitoring, and analysis of hurtful language to protect online users with the help of tools. The prior state-of-the-art is limited to detecting hurtful language such as hateful speech [1], abusive language [2], and profanity [3], collectively termed Negative Affective Language (NAL). In the following, we present the definitions and terms for variants of harassing language:

[1]Disclaimer: 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 reality, readers are alerted in advance and may wish to avoid reading this material if it could cause discomfort and disagreement.

[2]Observed January 2018 at `http://www.pewinternet.org/2017/07/11/online-harassment-2017/`

[3]In this work, cyberbullying and harassment are used interchangeably.

- **Hate speech** is the "speech that denigrates a person because of their innate and protected characteristics." [4] 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." [5]
- **Offensive Language** is about profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words to insult a targeted individual or group [6–9].
- **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 [10, 11].
- **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 [10–12].

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

A limitation of the state-of-the-art is the failure to exploit the **contextual type** of harassing language. Webster's dictionary[4] 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 context 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 the *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.

Figure 1. Five contextual types of harassment.

1. *Sexual harassment* is an 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 speech [13, 14].
   - **Sexually harassing tweet:** `@user can i know how old is that mouth that i'm gonna skullfuck hardly?`
   - **Sexually non-harassing tweet:** `three awesome teen babes licking each other pussies in absolute lesbian sex`
2. *Racial harassment* targets race and ethnicity characteristics of a victim such as color, country, culture, religion, in an offensive manner [15].

[4] `https://www.merriam-webster.com/dictionary/context`

- **Racially harassing tweet:** `@user shut the fuck up chink frog nigger`
- **Racially non-harassing tweet:** `rt @user: coming up on gmb odious man-child @userinterviews racist pathological lying asshat @user.`

3. *Appearance-related harassment* uses embarrassing language referring to body appearance. Fat shaming [16] and body shaming are key subtypes of this type of harassment.

   - **Appearance-wise harassing tweet:** `@user @user we started killing you because our backs couldn't handle the weight of your fatass anymore`
   - **Appearance-wise non-harassing tweet:** `@user: @user most insulting thing a skank can do toa woman who is worth having is mock her to a woman who isn't`

4. *Intellectual harassment* offends the intellectual power or opinions of individuals. Even smart people may be ridiculed and become victims [17].

   - **Intellectually harassing tweet:** `@user what a complete disgrace of human u r.real cool wish death. no surprise from a washed up fucktard really…`
   - **Intellectually non-harassing tweet:** `@user oh no i'm so sorry to hear that another one of your family members is a shithead`

5. *Political harassment* is related to someone's political views [18]. Typical targets are politicians and politically inclined individuals who receive threatening messages [19].

   - **Politically harassing tweet:** `@user: you're passive aggressive petty fuckbag who values a murderer fascist like putting over our own president. you're oƒ.`
   - **Politically non-harassing tweet:** `@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`

Recent examples illustrate that determining the real intent behind a tweet regarding the type of harassment can have serious implications for public perception. Consider the controversial tweet from Roseanne Barr targeting Valerie Jarett: `muslim brotherhood & planet of the apes had a baby=vj`, characterizing Jarrett — an African-American woman born in Iran — as a child of the Muslim Brotherhood and an ape [20]. Twitter Users regarded the tweet as racist, while Barr defended herself as making a bad joke about Jarrett's politics and looks. Thus, whether the tweet is considered to be racist or regarded as appearance-related or political makes a significant difference. 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 below: (i) We introduce five contextual types for harassment. Then, we provide a systematic, 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 we have employed. 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 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.

State-of-the-art in Harassment Research

The previous work in the area of detecting cyberbullying is targeted by various social media sources such as Twitter, Instagram, and Facebook. In Table 1, we succinctly present the prior literature with their goals, conclusions, and underlying data set. Here, we specifically describe particularly prominent related work. In [12], 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 [21] 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 [22] proposed an approach to spotting harassers as well as victims on social media. They consider 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, [23] proposes a method that simultaneously discovers instigators and victims of bullying incidents. It extends an initial bullying vocabulary using twitter and ask.fm. In [24], the authors proposed a supervised learning method for detecting cyberbullying in Japan. In [25], 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. [8] 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 [26] and Multinomial Naive Bayes [27] algorithms. They extracted N-gram features [28] to analyze the emotion, and they applied *Linguistic Inquiry and Word Count (LIWC)* to expand the feature set with the related emotional words. Interestingly, the authors of [11] target cyber-aggression and cyberbullying in a multi-modal context with text comments and media objects on Instagram. They concluded that the non-text features are not able to substantially improve the performance of cyberbullying detection compared to the text-based feature. Different from the previous work, from a psychological perspective, the authors in [29] seek reasons behind the updates of posts on Facebook. They noticed that: (i) the majority of posts are about social activities and everyday life, (ii) the people with low self-esteem update their status on relationship whereas the ones with high self-esteem update their status with respect to their children. Moreover, people with narcissistic personality disorder update their status through their achievements. Furthermore, they observe 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 more negative affect). Similarly, the authors of [30] discuss narcissism personality disorder of Facebook users. Our past work 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 nature of the relationship between participants (e.g., friends vs strangers) [31], and (ii) analyze comments/review threads to better identify offensive content in non-text media such as YouTube videos [32], to reliably detect harassment between participants.

Paper	Goal	Data	Conclusions
[6]	Detecting offensive and hateful speech language	85.4 Million Tweets Collected from 33458 twitter user. 25000 tweets selected by random sampling and these selected tweets contain profane words.	Collected discriminating terms for hate speech and offensive language
[10]	Detecting aggressors and their behavior on social media	1.6 Million tweets collected in 3 months, crowd sourced method used for annotating collected tweets.	Determined that posts of aggressor profiles are more negative
[7]	Detecting offensive language and identifying its sender.	The dataset includes comments from 2,175,474 distinct 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.
[12]	Predicting cyberbullying incidents on Instagram social media	Data collection was based on the 41K users that are cyberbullies according to the random seed nodes. 3165K tweets collected from 25K public users while 697K Tweets labeled as profane tweets because the tweets contain at least one profane word.	Classifier designed, trained, and applied for collecting data. Logistic regression classifier applied on data to train the predictive model with the forward feature selection approach
[21]	Detecting harassment based on the content, sentiment, and context	$\sim$11K tweets used in experiments; Use three of Fundacio'n Barcelona Media (FBM): Kongregate, Slashdot and MySpace. Totally 10,951 tweets collected and nearly 167 of them labeled as containing offensive language.	Improving accuracy in detecting harassing language using discussion-style and chat-style language
[22]	Detecting harassers and victims in cyberbullying incidents	Collecting twitter data with at least one profane word in specific tweet between November 1, 2015 and December 14, 2015. Twitter data contains 180,355 users and 296,308 tweets.	Accuracy improved by adding some network features.
[23]	Proposed the method that simultaneously discovers instigators and victims of bullying	180K profile on Twitter and $\sim$300K tweets. Seed terms: Profane words	Defines functions that scores cyberbullies and victims.
[24]	Detecting cyberbullying in Japanese community.	Data from Japanese secondary schools	Automatically extract new vulgarities from the Internet to keep their offensive lexicon up to date.
[8]	Understanding behavior and actions of individuals using emotion detection	$\sim$2.5M tweets	Developed a large tweets dataset using harassment-related emotion hashtags available in tweets.
[33]	Detecting bullying incident on social networks	$\sim$2M tweets collected in 4 weeks	Developed practical method of text mining, clustering, dimensionality reduction and classification used to find the harassment on social media.
[25]	Classifying cyberbullying activities on social network	18,554 users data was collected from Formspring.Me and MySpace.	Proposed a system that focuses on detecting cyberbullying activity in social networks using fuzzy logic which helps government to take action before many users become a victim of cyberbullying
[29]	Proposing a method to figure out the correllation of harassment and updated post on Facebook	Data was collected from 555 Facebook users currently residing in the United States (59% female; Mage = 30.90, SDage = 9.19)	These results show most of the updating posts related to the intellect people, children, and who they are in the romantic relation.
[30]	Identifying narcissism, activities on Facebook social media	256 Facebook users with different locations around the world.	Measuring the comment likes and compare it with the narcissistic score and performing some text mining methods to extract the user behavior

Table 1. Summery of the related research.

Type-aware Harassment Corpus

The publicly available state-of-the-art harassment-related corpus [34] contains tweets with two labels: (i) harassing and (ii) non-harassing. This corpus contains 20,428 **non-redundant** annotated tweets of which only 5,300 are labeled as harassing. In contrast, our recently published corpus [35] contains 24,189 **non-redundant** annotated tweets of which only 3,119 are labeled as harassing. Furthermore, these tweets are categorized into five types: (i) sexual, (ii) racial, (iii) appearance-related, (iv) intellectual, and (v) political. Our earlier paper [35] presented the details about the data collection and annotation of our corpus. The current paper explores issues associated with different types of harassment and their analysis and illustrates them using a comparative study of the two corpora. Table 2 provides general statistics about our corpus. The sign ✓denotes the harassing label, and the sign ✗denotes the non-harassing label.

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

Table 2. Annotation statistics of our categorized corpus.

LIWC Analysis for Different Types of Harassment

Linguistic analysis of our corpus sheds light on the linguistic 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 the 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 harassing tweets per type, (iv) *five contextual type-aware corpora* including non-harassing tweets per type. For linguistic analysis, we utilized LIWC[5] [36]. This tool tallies 96 linguistic features using a multiword lexicon for each feature. We individually analyzed each of the 12 sub-corpora using the LIWC tool. An effect size[6] [37] metric was used to determine significant discriminators [38]. Conventionally, a proportion (feature) f_i is considered moderately discriminating in case its effect size is more than 0.5 (i.e., $|e_{f_i}| > 0.5$), and is considered unhelpful if $|e_{f_i}| \approx 0$. We compared the prevalence of the 96 LIWC features in the harassing corpus against their prevalence in the associated 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 most discriminating features shown in Fig. 2. The extreme white (blue) color represents the 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, the 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. Please be noted that our corpus has already biased to curse words (which are used for crawling). Thus, our observations on the discriminatory features is subjected to a ”high recall curse word-laden corpus.”

[5] `https://liwc.wpengine.com/`

[6] It is a statistic that estimates the magnitude of an effect (e.g., mean difference, regression coefficient, Cohen’s *d*, and correlation coefficient)

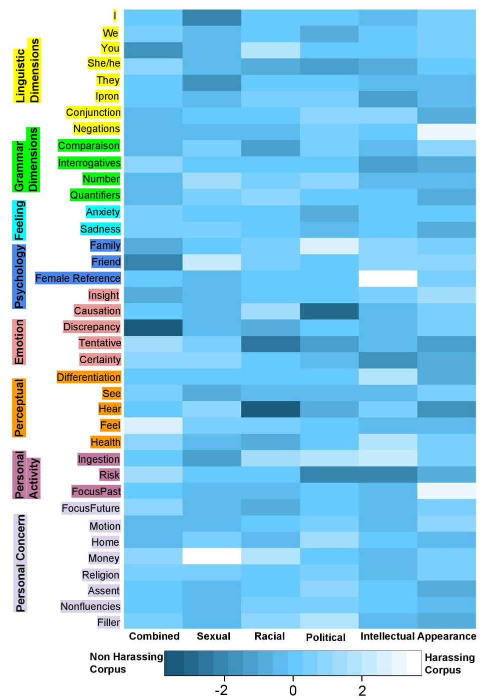


Figure 2. Significant LIWC features in comparing harassing corpus to non-harassing corpus from six categories of the corpus (Effect size scale is given at the bottom). The extreme blue (white) color implies the significance of a given feature in the non-harassing corpus (harassing corpus).

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 lesbian kiss`. Furthermore, the feature *"MONEY"* is prevalent in the harassing corpus with $e = 2.9$. E.g., `send free money bitch hoe wont give dance hoe ass industry bitch dicksuck porn star people`.

Racial Corpus The pronoun *"YOU"* is prevalent in the harassing corpus with $e = 0.9$, e.g., `Vishalp sikanda, Quideazam hahahaha u paki can block u cant debat u paki Indian.` The *"COMPARATIVE"* feature is prevalent in racial non-harassing corpus with $e = -0.84$, e.g., `save block paki 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 asshat 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., `fuck 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 non-harassing corpus. E.g., `well i'm true dumb fuck 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 camel toe.` The other significant feature in harassing corpus is the ''PAST TENSE''. E.g., `Ugli ass didn't go run yesterday get work fatass.` Furthermore, the *"COMPARATIVE"* feature is prevalent in appearance-related harassing corpus with $e = 0.63$. E.g., `hey lardass notice your look pizza perhaps like fuck salad asshole.` 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 w.r.t. intellectual issues.) E.g., `She is dumb fuck`.

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 shitface`.

Statistical Analysis of Different Types

We investigate the relationship between the offensive words employed for collecting our corpora and specific lexical items in the crawled corpora. We answer queries such as **Q1**: whether or not the 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 the content. Fig. 3 shows the word clouds of the top frequent words for harassing corpora, whereas Fig. 4 represents the word clouds of the top frequent words for non-harassing corpora. As the prevalence of curse words in the non-harassing corpora is higher, 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. Besides, we observed some emerging, frequent offensive words, such as "grab" and "camel", which can now be added to our initial offensive lexicon[7]. Furthermore, there are frequent words that are not necessarily offensive. E.g., consider "look" or "eat" in the appearance-related type. However, these are implicitly related to the associated type, applicable to the appearance of a subject. Regarding Q2, we observe that the frequent words in harassing corpora are different from those in non-harassing corpora. The particular words in harassing corpora again 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. Concerning Q3, we evaluated the top-15 frequent words for each type of harassing corpus as well as the non-harassing corpus. We asked the annotator to determine whether or not a given frequent word is related to the associated type either explicitly or implicitly. The results of our evaluation are represented in Table 3.

In harassing corpora, the percentage of type-dependency of words 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%. Thus, we conclude that the frequent words are mostly type-sensitive. Moreover, the prevalence of apparently offensive language in the non-harassing corpus reinforces our claim that offensive language *per se* is not necessarily harassing.

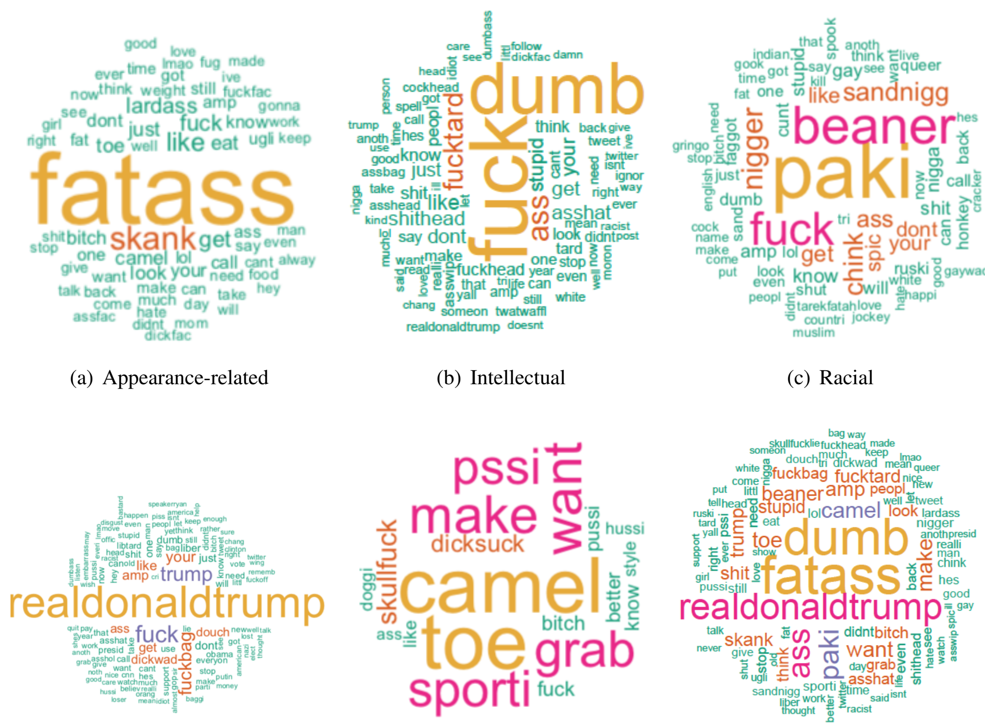
A caveat on the word clouds in Figures 3 and 4 is that the most frequent words appearing in the sub-corpus associated with each type are predominantly stop-words or curse words; these were employed as the seeds in our initial crawl of the twitter harassment dataset. Ignoring these words, whose presence cuts across different types of harassment, revealed the following

[7] `https://github.com/Mrezvan94/Harassment-Corpus`

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 intellectual harassment corpus, target words such as “dumb”, “stupid”, “work”, and “head”, are present. (iii) In political harassment corpus, target words such as “realdonaldtrump”, “libtard”, “dumb”, “touch bag”, “stupid”, and “cnn”, are present. (iv) In racial harassment corpus, target words such as “maki”, “nigger”, “beaner”, “chink”, “muslim”, “indian”, “moron”, and “jew”, are present. (iv) In sexual harassment corpus, target words such as “hump”, “hussy”, “lick”, and “grab”, are present.

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%

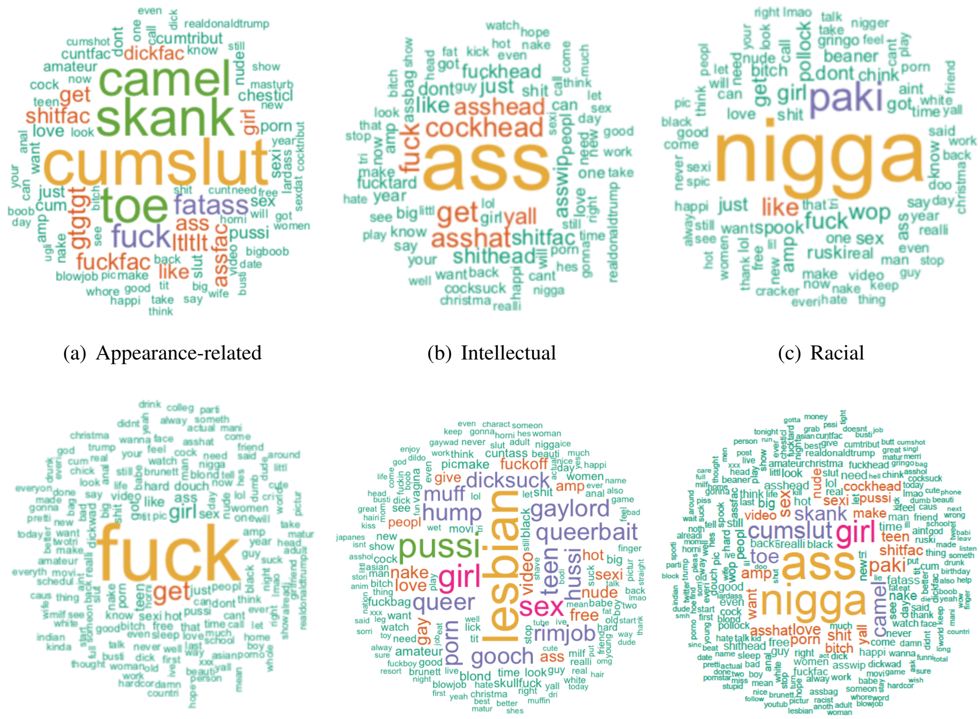
Table 3. Percentage of type-dependent lexicon words in top-15 frequent words (H stands for harassing corpus and NH stands for non-harassing corpus).



(a) Appearance-related (b) Intellectual (c) Racial

(d) Political (e) Sexual (f) Combined

Figure 3. Word clouds of harassing corpora.



(a) Appearance-related (b) Intellectual (c) Racial

(d) Political (e) Sexual (f) Combined

Figure 4. Word Clouds of non-harassing corpora.

Predicting Different Types of Harassing Posts

We aim to develop effective supervised learning methods to automatically detect harassing language and distinguish it from the non-harassing language for each contextual type. To address that, we pose two questions: (i) whether having individual binary classifiers for each type is more effective than a single multi-class classifier? (ii) whether training classifiers for specific types improves accuracy compared to the generic classifier (to recognize harassing vs. non-harassing tweets)? Regarding the first question, we trained both individual binary classifiers for each type and multi-class classifier to permit comparison. Concerning the second question, we compare the accuracy of binary classifiers for each specific types versus a baseline corpus disregarding type. In the following, we present the details, observations, and analysis of our experimental study.

Transforming Tweets to Vectors

We utilized four approaches for transforming tweets to numerical representation (i.e., vectors): (i) the conventional vectorization approach TFIDF, (ii) word2vec, (iii) paragraph2vec and (iv) 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 [39] to transform each given tweet into a weight vector T.

Distributional semantics (i.e., word2vec and paragraph2vec). Distributional semantics (so-called embedding models) [40] play a vital role in many Natural Language Processing (NLP) applications. They capture the semantics of text units (e.g., words, tweets, paragraphs or documents) in the underlying corpus and represent them in a low dimensional vector space. We use two types of embedding models for representing each tweet. The first one is **word2vec**[8] which transforms each individual word into a vector. The vector representation of a tweet is computed as the concatenation, summation or average vector of all vectors associated with words in the tweet. The second embedding model, called **paragraph2vec** [41], learns an individual vector for any tweet. W denotes the low dimensional vector obtained by word2vec approach, and P denotes the low dimensional vector obtained by paragraph2vec approach.

Training embedding models. To train both models (i.e., word2vec and paragraph2vec), we collected a corpus of 9 million tweets using our offensive lexicon as the underlying seed words. Then, we trained the embedding models on this accumulated corpus. Note that the dimension of the learned embeddings is 200 with the window size of 5.

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

Evaluation of the Harassment Classifiers

Set up. In our experimental study, we trained four types of classifiers (i) *Support Vector Machine* (SVM) [42], (ii) *K-Nearest Neighbors* (KNN) [42], (iii) *Gradient Boosting Machine* (GBM) [43], (iv) *Naive Bayes* (NB) [42]. We ran 10-fold cross-validation with re-sampling and iteration strategies (repeated five times).

Preparing training datasets. Since the number of harassing tweets are not equal to the number of non-harassing ones in our imbalanced corpus and varies for each type, we must prepare balanced datasets for training. We prepared five type-aware training data sets using an under-sampling approach (taking all of the harassing tweets with the equal number of non-harassing). Also, we prepared a combined training data set considering all the harassing tweets regardless of their type and an equal number of non-harassing tweets. Table 4 shows the the size of the training data sets for each type. Each data set contains a balanced number of harassing tweets versus non-harassing tweets.

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

Table 4. Size of the training datasets for each type.

Key General Observations. Table 5 represents the accuracy for all the settings involving the binary classifiers. Since the results of the NB classifier in all of the cases were inferior, we skip reporting them. In this table, P, R, and F stand for precision, recall and F-score, respectively. We offer the following observations: (i) The tweet representation using TFIDF vector is the most effective input representation, that is, it is better than the word2vec and paragraph2vec representations. This can be due to the small size of our training corpus and its inability to capture the semantics of data (tweets) adequately [44]. (ii) Adding LIWC vectors to the three basic vectorization approaches (i.e., T, W, P) improves the accuracy, sometimes as large as $\approx 10\%$. (iii) The combination of the three vectors T, L and W resulted in the best accuracy, implying that these three vectors are complementary. (iv) GBM classifier outperforms others in majority of the settings (except a few where SVM does well). (v) The best accuracy results (F)

[8] `https://code.google.com/archive/p/word2vec/`

observed for each type are as follows: 92% in appearance-related context, 93% in intellectual context, 94% in political context, 88% in racial, and 96% in sexual.

Key Comparison Observations. With respect to the effectiveness of the multi-class classifier, see Table 6. The general observations about binary classifiers are still valid here as (i) The T vectors outperform W and P vectors. (ii) Adding LIWC vectors improves the accuracy. (iii) The combination of the three vectors T, L and W results in the best accuracy. (iv) the GBM classifier outperforms classifiers. However, comparing the multi-class classifier and binary classifiers shows that the binary classifiers outperform the multi-class classifiers by as much as $\approx 20\%$.

To characterize the role of types, we compared the performance of binary type-specific classifiers with the binary classifier trained on the type-unaware data set, i.e., combined data set. The comparison results reveal that except for the racial type with 4% decrease of accuracy, all other type-specific classifiers gained in performance being higher or comparable to the type-unaware classifier. Note that the accuracy of our classifier will improve on a generic tweet corpus because our current corpus has been crawled 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.

Comparison to the state-of-the-art. To verify the effectiveness of the type-oriented classifiers, we ran them on the harassing tweets of Golbeck corpus which is a publicly available state-of-the-art harassment-related corpus [8]. The corpus contains 20,428 annotated tweets of which only 5,277 are labeled as harassing. It does not distinguishes the nature of harassment. In [35], we annotated the harassing tweets of Golbeck corpus with respect to our types using human judges. The proportion of harassing tweets per type is represented in the first column of Table 7. The proportion of harassing tweets after running our type-aware classifiers is represented in the second column of Table 7. Note that the racial type is dominant. The difference between the two proportions for the appearance-related and intellectual types is small. However, for sexual type, we observe a significant increase in the proportion. To make sense of the errors, we looked at a couple of tweets classified as sexual. E.g., for `@usr you deserved to be raped by a thousand Muslims in your cunt asshole`, 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 fucking feminist cunts like you. # feminazi # womenagainstfeminism` was classified as sexual. This analysis shows that these 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 Jews f`, (ii) ambiguous `MarshaBlackburn You're a whore 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 Jews rule`, (iv) poorly captured through annotation, e.g., the tweet `TrueNugget FeministPeriod OregonState Man college is becoming more and more a mistake.` in Golbeck corpus and our classifier misses them as they are weak cases of harassment.

Category	Features	SVM			KNN			GBM		
		P	R	F	P	R	F	P	R	F
Appearance	T	0.82	0.82	0.82	0.79	0.83	0.81	**0.85**	**0.89**	**0.86**
	W	0.62	0.63	0.62	0.71	0.62	0.66	**0.72**	**0.67**	**0.69**
	P	0.41	0.44	0.42	**0.49**	**0.53**	**0.51**	0.47	0.44	0.45
	TL	0.87	0.88	0.87	0.87	0.86	0.86	**0.89**	**0.92**	**0.90**
	WL	0.71	0.75	0.73	**0.84**	**0.79**	**0.82**	0.81	0.77	0.79
	PL	0.42	0.45	0.43	**0.52**	**0.56**	**0.54**	0.51	0.57	0.53
	TLW	**0.91**	**0.93**	**0.92**	0.88	0.86	0.87	0.90	0.93	0.91
Intellectual	T	0.80	0.86	0.82	0.61	0.55	0.58	**0.92**	**0.86**	**0.88**
	W	0.65	0.71	0.68	0.72	0.69	0.70	**0.83**	**0.85**	**0.83**
	P	0.39	0.40	0.39	0.29	0.33	0.31	**0.41**	**0.44**	**0.42**
	TL	0.87	0.92	0.89	0.66	0.60	0.63	**0.94**	**0.89**	**0.91**
	WL	0.72	0.84	0.77	0.80	0.79	0.79	**0.90**	**0.92**	**0.91**
	PL	0.54	0.57	0.55	0.41	0.49	0.45	**0.65**	**0.66**	**0.65**
	TLW	**0.91**	**0.94**	**0.93**	0.84	0.82	0.83	0.80	0.77	0.78
Political	T	0.87	0.85	0.86	0.80	0.79	0.75	**0.91**	**0.90**	**0.90**
	W	0.74	0.74	0.74	0.65	0.62	0.63	**0.79**	**0.81**	**0.80**
	P	**0.37**	**0.40**	**0.38**	0.31	0.34	0.32	0.36	0.34	0.35
	TL	0.91	0.90	0.90	0.88	0.86	0.87	**0.94**	**0.94**	**0.94**
	WL	0.83	0.82	0.82	0.72	0.70	0.71	**0.89**	**0.90**	**0.89**
	PL	**0.51**	**0.57**	**0.54**	0.47	0.49	0.48	0.52	0.50	0.51
	TLW	**0.95**	**0.94**	**0.94**	0.91	0.90	0.90	0.81	0.80	0.80
Racial	T	0.60	0.60	0.60	0.79	0.20	0.31	**0.77**	**0.69**	**0.72**
	W	**0.62**	**0.53**	**0.57**	0.59	0.51	0.54	0.57	0.55	0.56
	P	0.34	0.36	0.35	0.30	0.33	0.31	**0.42**	**0.44**	**0.43**
	TL	0.72	0.75	0.73	0.89	0.44	0.58	**0.87**	**0.84**	**0.84**
	WL	**0.74**	**0.68**	**0.71**	0.72	0.69	0.70	0.71	0.68	0.69
	PL	0.52	0.55	0.53	0.42	0.48	0.45	**0.56**	**0.61**	**0.58**
	TLW	0.83	0.81	0.82	0.84	0.81	0.83	**0.89**	**0.86**	**0.88**
Sexual	T	**0.95**	**0.92**	**0.93**	0.86	0.65	0.74	0.95	0.91	0.93
	W	0.41	0.36	0.38	0.61	0.67	0.64	**0.67**	**0.65**	**0.66**
	P	0.29	0.33	0.31	0.34	0.25	0.29	**0.42**	**0.45**	**0.43**
	TL	0.96	0.94	0.94	0.91	0.82	0.86	**0.97**	**0.95**	**0.96**
	WL	0.52	0.49	0.50	0.78	0.76	0.77	**0.85**	**0.81**	**0.83**
	PL	0.41	0.47	0.44	0.49	0.53	0.51	**0.69**	**0.66**	**0.67**
	TLW	0.62	0.58	0.60	0.57	0.51	0.53	**0.72**	**0.70**	**0.71**
Combined	T	0.84	0.81	0.82	0.76	0.71	0.73	**0.88**	**0.86**	**0.87**
	W	0.59	0.62	0.61	0.51	0.54	0.52	**0.67**	**0.70**	**0.68**
	P	0.32	0.35	0.33	0.26	0.31	0.28	**0.49**	**0.53**	**0.51**
	TL	0.90	0.87	0.88	0.82	0.85	0.83	**0.91**	**0.92**	**0.91**
	WL	0.71	0.79	0.74	0.58	0.59	0.58	**0.79**	**0.82**	**0.80**
	PL	0.51	0.57	0.57	0.41	0.47	0.44	**0.62**	**0.68**	**0.65**
	TLW	0.92	0.91	0.91	0.88	0.86	0.87	**0.94**	**0.91**	**0.92**

Table 5. The accuracy results of the employed classifiers per context (SVM stands for Support Vector Machine, KNN= K-Nearest Neighbor, GBM= Gradient Boosting Machine, NB= Naive Bayes). Also T stands for tf-idf vector, W for word2vec vector, L for LIWC vector and P for paragrapg2vec vector.

F	SVM			KNN			GBM			NB		
	P	R	F	P	R	F	P	R	F	P	R	F
T	0.61	0.56	0.59	0.52	0.55	0.54	**0.66**	**0.69**	**0.68**	0.29	0.33	0.31
W	0.37	0.31	0.34	0.27	0.24	0.26	**0.43**	**0.39**	**0.42**	0.19	0.16	0.18
P	0.11	0.13	0.12	0.09	0.12	0.11	**0.29**	**0.34**	**0.32**	0.06	0.04	0.05
TL	0.67	0.63	0.65	0.61	0.64	0.63	**0.69**	**0.72**	**0.71**	0.34	0.36	0.35
WL	0.49	0.52	0.51	0.36	0.38	0.37	**0.54**	**0.59**	**0.57**	0.27	0.36	0.32
PL	0.29	0.33	0.31	0.19	0.24	0.22	**0.40**	**0.44**	**0.42**	0.19	0.17	0.18
TLW	0.69	0.66	0.68	0.64	0.61	0.63	**0.71**	**0.67**	**0.69**	0.57	0.52	0.55

Table 6. The accuracy results of multi-class classifiers for various feature settings. Also T stands for tf-idf vector, W for word2vec vector, L for LIWC vector and P for paragrapg2vec vector and F for feature.

Category	Annotation Rate	Classifier Rate
Appearance-related	2.7%	1.5%
Intellectual	7.2%	8.0%
Political	3.0%	10.5 %
Racial	78.6%	56.5 %
Sexual	7.2%	23.0 %

Table 7. Comparison of the performance of our classifiers on the Golbeck corpus.

Ethics

We anonymized data. The study was approved by the Wright State University Institutional Review Board entitling ”Student Use of Social Media” with the attached number IRB #: 06251.

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 on a type-aware tweets corpus to analyze, learn, and understand harassing language for each particular 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, the “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. 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 tough 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.

Acknowledgment

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] Herz M, Molnár P. The content and context of hate speech: rethinking regulation and responses. Cambridge University Press; 2012.

[2] 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.

[3] Mangaonkar A, Hayrapetian A, Raje R. Collaborative detection of cyberbullying behavior in Twitter data. In: Electro/Information Technology (EIT), 2015 IEEE International Conference on. IEEE; 2015. p. 611–616.

[4] ElSherief M, Kulkarni V, Nguyen D, Wang WY, Belding-Royer EM. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. CoRR. 2018;abs/1804.04257.

[5] Talukder S, Carbunar B. AbuSniff: Automatic Detection and Defenses Against Abusive Facebook Friends. CoRR. 2018;abs/1804.10159.

[6] Davidson T, Warmsley D, Macy MW, Weber I. Automated Hate Speech Detection and the Problem of Offensive Language. CoRR. 2017;abs/1703.04009.

[7] Chen Y, Zhou Y, Zhu S, Xu H. Detecting offensive language in social media to protect adolescent online safety. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE; 2012. p. 71–80.

[8] Wang W, Chen L, Thirunarayan K, Sheth AP. Harnessing twitter" big data" for automatic emotion identification. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE; 2012. p. 587–592.

[9] Antigoni-Maria, Constantinos, Despoina, Ilias, Jeremy, Gianluca, et al.. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior - Dataset; 2018.

[10] Chatzakou D, Kourtellis N, Blackburn J, Cristofaro ED, Stringhini G, Vakali A. Mean Birds: Detecting Aggression and Bullying on Twitter. CoRR. 2017;abs/1702.06877.

[11] Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishr S. Prediction of cyberbullying incidents on the instagram social network. arXiv preprint arXiv:150806257. 2015;.

[12] Hosseinmardi H, Rafiq RI, Han R, Lv Q, Mishra S. 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; 2016. p. 186–192.

[13] ;. Available from: `https://www.joshuafriedmanesq.com/sexual-harassment.html`.

[14] ;. Available from: `https://www.eeoc.gov/laws/types/sexual_harassment.cfm`.

[15] ;. Available from: `https://www.joshuafriedmanesq.com/racial-slurs-and-racial-harassment.html`.

[16] 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.

[17] ;. Available from: `http://www.corrections.com/news/article/26649-ranking-bully-types-the-points-system`.

[18] ;. Available from: `http://www.brighthub.com/office/career-planning/articles/89787.aspx`.

[19] ;. Available from: `https://www.performanceicreate.com/political-discrimination-harassment/`.

[20] ;. Available from: `https://www.businessinsider.com/roseanne-barr-tweet-racist-valerie-jarrett-clintons2018-5August`.

[21] 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.

[22] Raisi E, Huang B. Cyberbullying detection with weakly supervised machine learning. In: Proceedings of the IEEE/ACM International Conference on Social Networks Analysis and Mining; 2017.

[23] Raisi E, Huang B. Cyberbullying identification using participant-vocabulary consistency. arXiv preprint arXiv:160608084. 2016;.

[24] Dybala MPP, Masui TMF, Rzepka R, Araki K. Machine Learning and Affect Analysis against Cyber-Bullying. 2010;.

[25] Nandhini BS, Sheeba J. Online social network bullying detection using intelligence techniques. Procedia Computer Science. 2015;45:485–492.

[26] 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.

[27] 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. Citeseer; 2005. p. 321–327.

[28] 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.

[29] 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.

[30] Younus A, Qureshi MA, Griffith J, O'Riordan C, Pasi G. A Study into the Correlation between Narcissism and Facebook Communication Patterns. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on. vol. 1. IEEE; 2015. p. 511–514.

[31] Edupuganti V. Harassment Detection on Twitter using Conversations. 2017;.

[32] Kandakatla R. Identifying Offensive Videos on YouTube. 2016;.

[33] Sanchez H, Kumar S. Twitter bullying detection. ser NSDI. 2011;12:15–15.

[34] 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.

[35] Rezvan M, Shekarpour S, Thirunarayan K, Shalin V, Sheth A. A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research. In: Proceedings of the 2018 ACM on Web Science Conference. ACM; 2018.

[36] Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. 2015;.

[37] Coe R. It's the effect size, stupid: What effect size is and why it is important. 2002;.

[38] Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological reviews. 2007;82(4):591–605.

[39] 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.

[40] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013;.

[41] Le Q, Mikolov T. Distributed representations of sentences and documents. In: ICML-14; 2014.

[42] 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. 2017;.

[43] Ayyadevara VK. Gradient Boosting Machine. In: Pro Machine Learning Algorithms. Springer; 2018. p. 117–134.

[44] De Boom C, Van Canneyt S, Bohez S, Demeester T, Dhoedt B. Learning semantic similarity for very short texts. In: Data Mining Workshop (ICDMW), 2015 IEEE International Conference on. IEEE; 2015. p. 1229–1234.