2020

ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter

Thilini Wijesiriwardene

Hale Inan

Ugur Kursuncu

Manas Gaur

Valerie L. Shalin

See next page for additional authors

Publication Info

Preprint version *Proceedings of Social Informatics 2020*, 2020.

ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter

Thilini Wijesiriwardene[1*] Hale Inan[4*] Ugur Kursuncu[1] Manas Gaur[1] Valerie L. Shalin[2] Krishnaprasad Thirunarayan[3] Amit Sheth[1] I. Budak Arpinar[4]

[1] AI Institute, University of South Carolina
thilini@sc.edu, kursuncu@mailbox.sc.edu, mgaur@email.sc.edu, amit@sc.edu
[2] Department of Psychology, Wright State University
valerie.shalin@wright.edu
[3] Department of Computer Science and Engineering, Wright State University
t.k.prasad@wright.edu
[4] Department of Computer Science, University of Georgia
hale.inan25@uga.edu, budak@cs.uga.edu
[*] *Equally contributed.*

Abstract. The convenience of social media has also enabled its misuse, potentially resulting in toxic behavior. Nearly 66% of internet users have observed online harassment, and 41% claim personal experience, with 18% facing severe forms of online harassment. This toxic communication has a significant impact on the well-being of young individuals, affecting mental health and, in some cases, resulting in suicide. These communications exhibit complex linguistic and contextual characteristics, making recognition of such narratives challenging. In this paper, we provide a multimodal dataset of toxic social media interactions between confirmed high school students, called ALONE (AdoLescents ON twittEr), along with descriptive explanation. Each instance of interaction includes tweets, images, emoji and related metadata. Our observations show that individual tweets do not provide sufficient evidence for toxic behavior, and meaningful use of context in interactions can enable highlighting or exonerating tweets with purported toxicity.

Keywords: Toxicity · Harassment · Social Media · Resource · Dataset

1 Introduction

The language of social media is a socio-cultural product, reflecting issues of relevance to the sample population and evolving norms in the exchange of coarse language and acceptable sarcasm, employing toxic, questionable language, and sometimes constituting actual harassment. According to a 2017 Pew Research Center survey, 41% of U.S. adults claim to have experienced some type of online harassment, offensive name-calling, purposeful embarrassment, physical threats, harassment over a sustained period of time, sexual harassment or stalking[5].

[5] https://www.pewresearch.org/fact-tank/2017/07/11/key-takeaways-online-harassment/

Toxic behavior is prevalent among adolescents, sometimes leading to aggression [26, 27]. Adolescents exemplify a population that is particularly vulnerable to disturbing social media interactions[6] [47], and this behavior is observable in a network of high school students [5]. Further, a toxic online environment may cause mental health problems for this population[7] [2, 40, 20, 48]. While a victim may experience a negative reaction from a toxic environment of offensive language, this differs from *targeted* toxicity which is usually directed whose content collected and confirmed with a unique method towards one individual. The analysis of single tweets or individual users is potentially misleading as the context of interactions between the two people (e.g., source and target) dictates the determination of toxicity. In other words, two individuals who are friends may use coarse keywords or language that is seemingly toxic, but it may be sarcastic, exonerating them from toxicity.

In this paper, we provide a dataset and its details, specific to toxic behavior in social media communications. This dataset has two particular contributions: (i) the population is *high school students* whose content was collected and confirmed with a unique method, and (ii) it was designed based on the *interactions* between participants. The detection of true toxic behavior against a persisting background of coarse language poses a challenging task. Moreover, the scope of the original crawl has great bearing on the prevalence of toxicity features and the criteria for toxic behavior itself. To address these issues, we have assembled a social media corpus from Twitter for a sample of midwestern American High School Students. We assert a dyadic, directed interaction, between a source and a target. Existing related datasets (see Related Work section) focus mainly on the user or tweet level for the task of detecting toxic content. Such datasets fail to capture adequately the fundamental and contextual nuances in the language of these conversations. Thus, our corpus preserves and aggregates the social media interaction history between participants. This enables the determination of existing friendship and hence possible sarcasm. Because individuals can communicate with multiple partners, we have the potential of detecting unique toxic person-victim pairings that would be otherwise undetectable in the raw original crawl.

Each entry in our dataset consists of 12 fields: *Interaction Id, Count, Source User Id, Target User Id, Emoji, Emoji Keywords, Tweets, Image Keywords, created at, favorite count, in reply to screenname* and *label* where the *Tweets* field contains an aggregation of the tweets between a specific pair of source and target. For preliminary analysis, we define a single dimension of *toxic language*, pegged at one end by benign content and the other by harassment. This dimension can be partitioned into several, partially overlapping classes, determined by a decision rule. We have identified and experimented with three levels of toxic interactions between source and target: *Toxic (T), Non-Toxic (N), or Unclear*

[6] https://www.cim.co.uk/newsroom/release-half-of-teens-exposed-to-harmful-social-media/

[7] https://www.cnn.com/2016/12/14/health/teen-suicide-cyberbullying-continues-trnd/index.html

(U). However, the boundaries between levels are discretionary, accommodating construct definitions that are, at best, debatable.

We include examples across the continuum of toxic language, with sufficient context to determine the nature of toxicity. We detect true toxicity on Twitter by analyzing interactions among a collection of tweets, in contrast with prior approaches where the main focus is performing user or tweet level analysis. Further, we assert that detecting a user as a toxic person with respect to one victim does not provide evidence of being a universal toxic person because they can be friendly to a majority of others.

2 Related Work

We reviewed prior work for the variety of overlapping constructs related to toxic exchanges. The social media literature related to toxic behavior lacks crisp distinctions between: offensive language [14, 19, 37], hate speech [14, 10, 4, 50], abusive language [14, 33, 31] and cyberbullying [8, 18, 11]. For example, the following definition of offensive language substantially overlaps with the subsequent definition of hate speech. According to [14], *offensive language is profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group. Hate speech is language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of a group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.* [42] classifies swearing, aggressive comments, or mentioning the past political or ethnic conflicts in a non-constructive and harmful way as hateful: *@user_name nope you just a stupid hoe who wouldn't know their place* 😂😂 comprises both offensive and hate speech. Specifically, the challenge lies in operationalizing the contextual differences between offensiveness, hate speech and harassment. As the existing work on offensive content, harassment and hate speech fails to take into account the nature of the relationship between participants, we focus our attention on the context-aware analyses of targeted exchanges.

Offensive— [4] annotated 16K tweets from [52] with the labels, racist, sexist or neither. 3383 and 1972 tweets were sexist and racist respectively, and others were labeled as neither. In [31], their aim was to detect abusive language on online user comments posted on Yahoo. 56,280 comments were labeled as 'Abusive' and 895,546 comments as 'Clean'.

Hate Speech—[44] developed a dataset to identify the main targets of online hate speech including the nine categories such as race, behavior, physical, sexual orientation, class, gender, ethnicity, disability, religion, and other for non-classified hate targets. 178 most popular targets from Whisper and Twitter were manually labeled, unveiling new forms of online hate that can be harmful to people. [10] focused on distinguishing hate speech from other forms of offensive language. They extracted 85.4 million tweets from 33,458 users, and randomly sampled 25K tweets containing words from a hate speech lexicon. Individual tweets were labeled as hate speech, offensive or neither. [43] presented an annotated corpus of tweets classified by different levels of hate to provide an onto-

logical classification model to identify harmful speech. They randomly sampled 14,906 tweets and developed a supervised system used for detection of the class of harmful speech. In [52], tweets were sampled from the 130K tweets, and in addition to "racism", "sexism", and "neither", the label "both" was added. A character n-gram based approach provided better performance for hate speech detection. [51] examined the influence of annotators' knowledge for hate speech on classification models, labeling individual tweets. Considering only cases of full agreement among amateur annotators, they found that amateur annotators can produce relatively good annotations as compared to expert annotators.

Harassment—A number of researchers have attempted to identify dimensions or factors underpinning harassment. [29] drew on the model [6] that conceptualized aggression on four dimensions: *verbal, physical, direct-indirect,* and *active-passive*. [38] analyses the linguistics aspects of harassment based on different harassment types. Consistent with our interest in interaction history between participants, cyberbullying emphasizes the repetitiveness of aggressive acts [35]. The harasser may target a victim over a period of time, or a group of harassers may target a victim about the same demeaning characteristic or incident. Apart from repetitiveness, the difference of power between the harasser and victim suggests cyberbullying. However, this work [35] is not computationally oriented. Golbeck [16] introduced a large, human labeled corpus of online harassment data including 35,000 tweets with 5495 non-harassing and 29505 harassing examples.

In contrast to this literature, our approach to the problem is to focus on interactions between participants to capture the context of the relationship rather than solely tweets or users. As online toxic behavior is a complex issue that involves different contexts and dimensions [22, 21, 1], tweet-level or user-level approaches do not adequately capture the context with important nuances due to the fluidity in the language. Our interaction-based dataset will enable researchers to uncover critical patterns for gaining a better understanding of toxic behavior on social media. Additionally, our dataset is unique in its focus on high school student demographic.

3 Dataset

For the dataset ALONE, we retrieved 469,786 tweets from our raw Twitter data, and used a harassment lexicon provided by [39] to filter tweets that are likely to contain toxic behavior, obtaining a collection of 688 interactions with aggregated 16,901 tweets.

3.1 Data Collection

We focused on tweets as the source for our dataset because of its public access. Besides text, tweets can contain images, emoji and URLs as additional content. To create a ground truth dataset, we reviewed public lists of students, such as the list of National Merit Scholars published in newspapers, identifying 143 names of the attendees of a high school. Using the list of identified individuals, we searched Twitter for the profiles associated with these students using Twitter

APIs. Then, with the guidance of our cognitive scientist co-author, we confirmed that the users that we retrieved were high school students, through their profiles and tweets conversing on their school mascot, clubs or faculty members. The 143 user profiles with their tweets constituted the seed corpus.

Dataset Expansion: As a typical network of high school students is larger than 143 users, we expanded the network using the friend and follower relationships. We followed the following procedure:

- Collect friends and followers lists for each seed profile.
- Exclude non-student accounts: We identified the accounts following each other considering them as candidate students, and removed accounts that are not both following and being followed by the accounts in the friends and followers lists of seed accounts (not common profiles). As the adults, such as teachers, would notice any toxic behavior, such as harassment, bullying or aggression, which may have consequences, students with potentially toxic behavior would avoid following their social media accounts [46, 30] to sequester social network behavior [28, 30]. We obtained 8805 accounts that follow and are being followed by at least one seed account, as candidates for student accounts in the high school. We removed 80 accounts as they were suspended or deleted or otherwise protected by account owners.
- Retain only the peer profiles that follow and are being followed by more than 10% of the seed profiles, yielding 320 likely peers. To confirm the absence of false positives, 50 accounts out of the 320 likely peers were randomly selected and manually validated that all the 50 were confirmed student accounts. When tweets of the newly added 320 accounts were crawled, seven accounts were deleted or restricted. Hence, we removed them from the dataset, resulting in 456 accounts (143 seed and 313 added).

After we finalized the 456 accounts, tweets (up to 3200 if available) were collected for each, starting from the most recent (May 2018), along with their account metadata, using the Twitter API.

Interaction-based Dataset: As our toxic behavior construct requires interactions between participants, we pruned the tweet corpus to retain a dataset that consists of interactions. We define an interaction as a collection of tweets exchanged between the two participants (e.g., source and target) in one direction, and on Twitter, we consider mentions (including replies) and retweets as interactions. For instance, one user may mention another user in a tweet for harassing, bullying or insulting. Moreover, retweeting a harassing tweet potentially boosts popularity, which creates the role of bystander for the source, suggesting that the retweeting user (source) is actually supporting or helping the harasser (target). We have left retweet indicators (e.g., RT @username:) in the data. Further, some tweets are included in multiple interactions; hence, these communications are a part of a group communication that is not dyadic. For some instances, source

and target are the same users, and we left these conversations in the dataset as they may be likely a part of group aggression.

We aggregated tweets that qualify as interactions between users, potentially reducing the false alarm rate of an analysis solely based on the presence of characteristics of offensive language [3]. This allows for the detection of a particularly intriguing combination of positive and negative sentiment lexical items, suggestive of sarcasm, e.g., *happy birthday @user_name love you but hate your feet* 💖💗💟💕💖 and *Happy birthday ugly!!* 💗. The presence of "Happy Birthday" or positive emoji (see above) alters the interpretation of content that would otherwise be regarded as potentially suggestive of toxic behavior and the phenomenon of conflicting valence *exoneration* content, assuming that the toxic content is sarcastic, e.g., the source does not really believe the recipient has unattractive feet or is generally ugly. Moreover, contextual analysis reveals that some of these are not truly toxic. Prior tweets in an interaction provides exonerating context, by indicating the presence of friendship, thus correcting the false positive. Designing the dataset based on interactions captures the context of the relationship between the two user; thus, enabling one to employ computational techniques to retrieve meaningful information concerning true toxicity.

A portion of the tweets does not include any interaction indicator, but they refer to a person indirectly without mentioning or writing the name with malicious intent, to avoid the authority figures. This is called *Subtweeting*[8][9] [36, 13, 9]. Adolescents have specifically developed such practice due to their own privacy concerns and parental intrusion. For each user, we aggregated the tweets that do not mention the target explicitly, and indicated the target as *"None"*.

Then, a harassment lexicon [39] was utilized to filter the interactions that potentially contain toxic content. For online harassment, source and target dyads can be considered as *harasser-victim* or *bystander-harasser*. Further, as capturing context to determine the toxicity in the content is critical, an interaction should include a sufficient number of tweets. Therefore, we set an empirical threshold for one interaction as having at least three tweets, to capture context.

We have fully *de-identified* the interactions by replacing; (i) Twitter usernames and mentions in tweets with a numeric user id, (ii) URLs with the token of $< url >$, and (iii) person names with the token of $< name >$. We have also included the following metadata for each tweet in the interactions: timestamp, favorite counts, and the de-identified user id of the replied user (if the tweets is a reply). Thus, researchers will have the ability to study a variety of aspects of this problem such as time series analysis. The finalized dataset includes 688 interactions with 16,901 tweets. The fields in an instance are as follows: Interaction Id, Count, Source User Id, Target User Id, Label, Emoji, Emoji Keywords, Tweets, Image Keywords, Timestamp, in reply to and favorite count. "Count" field holds information for the number of tweets in an interaction. "Source and Target User Id" fields hold numeric identification *(after de-identification)* information. A

[8] https://www.theguardian.com/technology/blog/2014/jul/23/subtweeting-what-is-it-and-how-to-do-it-well

[9] http://bolobhi.org/abuse-subtweeting-tweet-school-cyber-bullying/

"Label" field holds the assigned label (T,N,U) for the interaction. While the "Emoji" field holds the emoji being used in the tweets, "Emoji Keywords" field provides the keywords that explain the meaning of the emoji, retrieved from EmojiNet [53]. The "Tweets" field has the tweets, and the following fields holds the metadata for each tweet: (i) Timestamp: time information of a tweet, (ii) in reply to: (non-real) user id of the target if the tweet is a reply, (iii) favorite count: number of favorites. See Table 1 for example interactions from the dataset with four fields.

Label	**Tweets**
T	if you gon say n.... this much, the LEAST you could do is hit the tanning bed < *url* > *** you're f..... the most hideous and racist piece of s... *** YOU ARE LITERALLY F..... RACIST SHUT THE F... UP *** yeah 😂😂 you're not racist at all !!!!!!!! *** are you in f..... politics no, you're like 17 s... the f... up and stop putting your f̈actsön...
T	ight f... you again *** nah f... all of you frfr bunch of f..... f..... *** f... you < *url* > you have no room to be talking s... shut your bum a.. up frfr 😂 ** you're halarious, f... you and everyone that favorited that and retweeted that
N	"Kix is the handjob of cereals"- John Doe 😂😂 < *imageurl* > *** Explain to that i..... that doing it spreads the word and the chance of someone donating XD fedora wearing as... *** get the f... off my twitter b.... BOI *** guys follow bc he's an i.... and forgot his password.
U	This tweet was dumb I agree with u this time *** hahaha I'm so dumb *** that's my mom f.... *** boob *** never seen a bigger lie on the Internet then this one right here

Table 1: Examples from the dataset with labels Toxic (T), Non-Toxic (N) and Unclear (U). The expletives were replaced with the first letter followed by as many *dots* as there are remaining letters.

Multimodality: As it will be described in Section Descriptive Statistics, different modalities of data, such as text, image, emoji, appear in Toxic and Non-Toxic interactions with different proportions. Therefore, we provided explanations of potentially valuable emoji and images. Each image name was created by combining "source user id", "target user id", and "tweet number" in an interaction that each image pertains to. For example: the image 0023.0230.5.jpg is from a tweet between "user 0023" and "user 0230" and the 5th tweet in their interaction. We processed these images utilizing a state-of-the-art image recognition tool, ResNet[10] [17], providing the objects recognized in images with their probabilities (top-5 accuracy= 0.921). We kept the top 20 (empirically set) recognized object names. For example, an image has the following set of recognized objects: "television", "cash machine", "screen", "monitor", "neck brace", "toyshop", "medicine chest", "library", "home theater", "wardrobe", "score-

[10] https://github.com/onnx/models/tree/master/vision/classification/resnet

board", "moving van","entertainment center", "barbershop", "desk", "web site". We utilized EmojiNet [11] [53] to retrieve the meanings of the emoji in the interactions, and provided in the dataset. For instance, for the emoji 😂, EmojiNet provides the following set of keywords: "face", "tear", "joy", "laugh", "happy", "cute", "funny", "joyful", "hilarious", "teary", "laughing", "person", "smiley", "lol", "emoji", "wtf", "cry", "crying", "tears", "lmao". Specifically, the significant difference in the use of image, video and emoji between the content of Toxic and Non-Toxic interactions, suggests that the contribution of multimodal elements would likely be critical.

Privacy and Ethics Disclosure: We use only public Twitter data, and our study does not involve any direct interaction with any individuals or their personally identifiable private data. This study was reviewed by the host institution's IRB and received an exemption determination. As noted above, we follow standard practices for anonymization during data collection and processing by removing any identifiable information including names, usernames, URLs. We do not provide any Twitter user or tweet id, or geolocation information. Due to privacy concerns and terms of use by Twitter, we make this dataset available upon request to the authors, and researchers will be required to sign an agreement to use it only for research purposes and without public dissemination.

Three Label	**Two Label**
0.63	0.65

Table 2: For **three** and **two** labels, agreement scores between the three annotators using **Krippendorff's alpha**.

Kappa	**A**	**B**
B	0.77	-
C	0.52	0.62

Table 3: Pairwise agreement for the **three** label scheme, agreement scores between the three annotators (A,B,C) using Cohen Kappa

3.2 Annotation

Capturing truly toxic content on social media for humans requires reliable annotation guidelines for training annotators. Our annotators have completed a rigorous training process including literature reviews and discussions on online toxic behavior and its socio-cultural context among adolescents. Three annotators labeled the interactions using three labels: *Toxic (T), Non-Toxic (N)* and *Unclear (U)*. The annotators were trained by our co-author cognitive scientist to consider the context of the interaction rather than individual tweets while determining the label of an interaction. We developed a guideline for annotators to follow that comprises intent-oriented criteria for labeling interactions as Toxic (T). That is, a tweet is toxic if the interactions contain: (i) Threat to harm a

Kappa	**A**	**B**
B	0.82	-
C	0.49	0.63

Table 4: Pairwise agreement for **two** labels, agreement scores between the three annotators (A,B,C) using Cohen Kappa

[11] http://wiki.aiisc.ai/index.php/EmojiNet

person, (ii) Effort to degrade or belittle a person, (iii) Express dislike towards a person or a group of people, (iv) Promote hate/violence/offensive language towards a person or a group of people, (v) Negatively stereotype a person or a minority, (vi) Support and defend xenophobia, sexism or racism.

Number of Tweets	**Mean**	**Min**	**Max**
Toxic	13.28	3.0	304.0
Non-Toxic	7.15	3.0	99.0

(a)

Number of Emoji	**Mean**	**Min**	**Max**
Toxic	6.72	0.0	290.0
Non-Toxic	3.51	0.0	60.0

(b)

Number of URLs	**Mean**	**Min**	**Max**
Toxic	2.70	0.0	73.0
Non-Toxic	1.63	0.0	26.0

(c)

Number of Images	**Mean**	**Min**	**Max**
Toxic	1.18	0.0	20.0
Non-Toxic	0.86	0.0	12.0

(d)

Table 5: (a)Descriptive statistics of tweets per interaction. (b) Descriptive statistics of emoji per interaction. (c) Descriptive statistics of URLs per interaction. (d) Descriptive statistics of images per interaction. There were 140 images showing Toxic Behavior and 471 images showing Non-Toxic Behavior.

If an annotator could not arrive at a conclusion after assessing the interaction following this guideline, it was labeled as *Unclear*. After the annotations were completed by *the three annotators*, the labels were finalized by majority vote. Then, agreement scores were computed utilizing Krippendorff's alpha (α) and Cohen's Kappa (κ). Note that the instances labelled Unclear (U) can be included in the training to exercise the robustness of a learned model, or they can be removed as they add noise (as per the consensus of the annotators). To accommodate both scenarios, we create two schemes: (i) three label (T, N, U), (ii) two label (T, N) removing Unclear (U) instances [15]. We perform two annotation analysis for both schemes: (i) A group-wise annotator agreement to find the robustness of the annotation by the three annotators using Krippendorff's alpha (α) [45], (ii) A pair-wise annotator agreement using Cohen's Kappa (κ) to identify the annotator with highest agreement with others. In the three-label scheme, α was computed as 0.63, and for the two label scheme, (α) was 0.65. The agreement scores reported in Table 2 imply substantial agreement[12] [7]. We also computed the agreement between annotators using κ and provided in Table 3 and Table 4, for three label and the two label, respectively. While the annotators A and B have substantial and near perfect agreement, C has moderate and substantial agreement with A and B, both for the three and two label schemes respectively [7].

[12] http://homepages.inf.ed.ac.uk/jeanc/maptask-coding-html/node23.html

3.3 Descriptive Statistics

In this section, we provide descriptive statistics of the dataset concerning the distribution of tweets, images, emoji and URLs with respect to labels. Table 6 shows the overall distribution of the instances as Toxic interactions constitute the 17.15% of the dataset, while 79.51% remains as Non-Toxic. A minority group of interactions with 3.34% comprises the Unclear instances where annotators agreed that no conclusion could be derived. While the imbalance in the dataset provides challenges in the modeling of toxic behavior, it is reflective of the nature of occurrence in real life. On the other hand, although the number of toxic interactions is smaller, they are richer in content as well as multimodal elements, compared to non-toxic interactions [23] (see Tables 5a, 5b, 5c, 5d, and 7). Prior research shows that appropriate incorporation of multimodal elements in modeling with social media data would improve performance [23, 24, 12, 32]. In Table 5a, we see mean and maximum number of tweets per interaction for Toxic ones being significantly higher than Non-toxic ones, suggesting the intensity of the toxic content. Further, according to Tables 5a, 5b, 5c, 5d, and 7, in the Toxic content, the use of multimodal elements such as image, video, and emoji, is clearly higher, suggesting that the incorporation of these different modalities in the analysis of this dataset will be critical for a reliable outcome [24, 23, 12, 32].

Toxic	**Non-Toxic**	**Unclear**
118 (17.15%)	547 (79.51%)	23 (3.34%)

Table 6: Overall distribution of the data instances over the three labels.

Type of URLs	**Number of URLs**
Image URLs	140 (43.88%)
Video URLs	44 (13.79%)
Text URLs	48 (15.04%)

Table 7: Different types of URLs in **toxic** interactions.

4 Discussion and Conclusion

We created and examined the multimodal ALONE dataset for adolescent participants utilizing a lexicon [39] that divides offensive language into different types concerning appearance, intellectual, political, race, religion, and sexual preference. Given its unique characteristics concerning (i) adolescent population and (ii) interaction-based design, this dataset is an important contribution to the research community, as ground truth to provide a better understanding of online toxic behavior as well as training machine learning models [42, 25] and performing time-series analysis. Specifically, quantitative as well as qualitative analysis of this dataset will reveal patterns with respect to social, cultural and behavioral dimensions [34, 49, 41] and shed light on etiology of toxicity in relationships. Further, researchers can develop guidelines for different kinds of toxic behavior such as harassment and hate speech, and annotate the dataset accordingly. Lastly, we reiterate that the ALONE dataset will be available upon request to the authors, and the researchers will be required to sign an agreement to use it only for research purposes and without public dissemination.

Acknowledgement

We acknowledge partial support from the National Science Foundation (NSF) award CNS-1513721: "Context-Aware Harassment Detection on Social Media". Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

1. Arpinar, I.B., Kursuncu, U., Achilov, D.: Social media analytics to identify and counter islamist extremism: Systematic detection, evaluation, and challenging of extremist narratives online. In: 2016 International Conference on Collaboration Technologies and Systems (CTS). pp. 611–612. IEEE (2016)
2. Arseneault, L., Bowes, L., Shakoor, S.: Bullying victimization in youths and mental health problems:'much ado about nothing'? Psychological medicine (2010)
3. Badjatiya, P., Gupta, M., Varma, V.: Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In: The World Wide Web Conference. pp. 49–59 (2019)
4. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: WWW (2017)
5. Brener, N.D., Simon, T.R., Krug, E.G., Lowry, R.: Recent trends in violence-related behaviors among high school students in the united states. JAMA (1999)
6. Buss, A.H.: The psychology of aggression (1961)
7. Carletta, J., Isard, A., Isard, S., Kowtko, J.C., Doherty-Sneddon, G., Anderson, A.H.: The reliability of a dialogue structure coding scheme. (1997)
8. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A.: Mean birds: Detecting aggression and bullying on twitter. In: ACM Web Science (2017)
9. Crumback, D.: Subtweets: The new online harassment (2017)
10. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: AAAI-ICWSM (2017)
11. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: AAAI-ICWSM (2011)
12. Duong, C.T., Lebret, R., Aberer, K.: Multimodal classification for analysing social media. arXiv preprint arXiv:1708.02099 (2017)
13. Edwards, A., Harris, C.J.: To tweet or 'subtweet'?: Impacts of social networking post directness and valence on interpersonal impressions. Computers in Human Behavior **63**, 304–310 (2016)
14. Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N.: Large scale crowdsourcing and characterization of twitter abusive behavior (2018)
15. Gaur, M., Alambo, A., Sain, J.P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., Sheth, A., Welton, R., Pathak, J.: Knowledge-aware assessment of severity of suicide risk for early intervention. In: The World Wide Web Conference. pp. 514–525. ACM (2019)
16. Golbeck, J., Ashktorab, Z., Banjo, R.O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A.A., Gergory, Q., Gnanasekaran, R.K., et al.: A large labeled corpus for online harassment research. In: ACM Web Science (2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

18. Hosseinmardi, H., Mattson, S.A., Rafiq, R.I., Han, R., Lv, Q., Mishra, S.: Analyzing labeled cyberbullying incidents on the instagram social network. In: SocInfo (2015)
19. Jay, T., Janschewitz, K.: The pragmatics of swearing. Journal of Politeness Research. Language, Behaviour, Culture (2008)
20. Kumpulainen, K., Räsänen, E., Puura, K.: Psychiatric disorders and the use of mental health services among children involved in bullying. Aggressive Behavior Journal (2001)
21. Kursuncu, U.: Modeling the Persona in Persuasive Discourse on Social Media Using Context-aware and Knowledge-driven Learning. Ph.D. thesis, University of Georgia (2018)
22. Kursuncu, U., Gaur, M., Castillo, C., Alambo, A., Thirunarayan, K., Shalin, V., Achilov, D., Arpinar, I.B., Sheth, A.: Modeling islamist extremist communications on social media using contextual dimensions: Religion, ideology, and hate. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–22 (2019)
23. Kursuncu, U., Gaur, M., Lokala, U., Illendula, A., Thirunarayan, K., Daniulaityte, R., Sheth, A., Arpinar, I.B.: What's ur type? contextualized classification of user types in marijuana-related communications using compositional multiview embedding. In: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). pp. 474–479. IEEE (2018)
24. Kursuncu, U., Gaur, M., Lokala, U., Thirunarayan, K., Sheth, A., Arpinar, I.B.: Predictive analysis on twitter: Techniques and applications. In: Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining, pp. 67–104. Springer (2019)
25. Kursuncu, U., Gaur, M., Sheth, A.: Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning. In: Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice. Stanford University, Palo Alto, California, USA. AAAI-MAKE (2020)
26. Liu, J., Lewis, G., Evans, L.: Understanding aggressive behaviour across the lifespan. Journal of psychiatric and mental health nursing (2013)
27. Lowry, R., Powell, K.E., Kann, L., Collins, J.L., Kolbe, L.J.: Weapon-carrying, physical fighting, and fight-related injury among us adolescents. American journal of preventive medicine (1998)
28. Mishna, F., Schwan, K.J., Lefebvre, R., Bhole, P., Johnston, D.: Students in distress: Unanticipated findings in a cyber bullying study. Children and youth services review (2014)
29. Namie, G., Namie, R.: Bully at work: What you can do to stop the hurt and reclaim your dignity on the job (2009)
30. Nilan, P., Burgess, H., Hobbs, M., Threadgold, S., Alexander, W.: Youth, social media, and cyberbullying among australian youth:"sick friends". Social Media+ Society (2015)
31. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: WWW (2016)
32. O'Halloran, K., Chua, A., Podlasov, A.: The role of images in social media analytics: A multimodal digital humanities approach. In: Visual communication (2014)
33. Papegnies, E., Labatut, V., Dufour, R., Linarès, G.: Detection of abusive messages in an on-line community. In: CORIA (2017)
34. Parent, M.C., Gobble, T.D., Rochlen, A.: Social media behavior, toxic masculinity, and depression. Psychology of Men & Masculinities **20**(3), 277 (2019)
35. Patchin, J.W., Hinduja, S.: Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth violence and juvenile justice (2006)

36. Rafla, M., Carson, N.J., DeJong, S.M.: Adolescents and the internet: what mental health clinicians need to know. Current psychiatry reports **16**(9), 472 (2014)
37. Razavi, A.H., Inkpen, D., Uritsky, S., Matwin, S.: Offensive language detection using multi-level classification. In: CCAI (2010)
38. Rezvan, M., Shekarpour, S., Alshargi, F., Thirunarayan, K., Shalin, V.L., Sheth, A.: Analyzing and learning the language for different types of harassment. Plos one **15**(3), e0227330 (2020)
39. Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V.L., Sheth, A.: A quality type-aware annotated corpus and lexicon for harassment research. In: ACM Web Science (2018)
40. Rivers, I., Poteat, V.P., Noret, N., Ashurst, N.: Observing bullying at school: The mental health implications of witness status. School Psychology Quarterly (2009)
41. Safadi, H., Li, W., Rahmati, P., Soleymani, S., Kursuncu, U., Kochut, K., Sheth, A.: Curtailing fake news propagation with psychographics. Available at SSRN 3558236 (2020)
42. Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S.g., An, J., Kwak, H., Jansen, B.J.: Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: ICWSM. pp. 330–339 (2018)
43. Sharma, S., Agrawal, S., Shrivastava, M.: Degree based classification of harmful speech using twitter data. arXiv preprint arXiv:1806.04197 (2018)
44. Silva, L., Mondal, M., Correa, D., Benevenuto, F., Weber, I.: Analyzing the targets of hate in online social media. In: AAAI-ICWSM (2016)
45. Soberón, G., Aroyo, L., Welty, C., Inel, O., Lin, H., Overmeen, M.: Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In: CrowdSem 2013 Workshop (2013)
46. Søndergaard, D.M.: Bullying and social exclusion anxiety in schools. British Journal of Sociology of Education (2012)
47. Unicef, et al.: An Everyday Lesson: End Violence in Schools (2018)
48. Viner, R.M., Aswathikutty-Gireesh, A., Stiglic, N., Hudson, L.D., Goddings, A.L., Ward, J.L., Nicholls, D.E.: Roles of cyberbullying, sleep, and physical activity in mediating the effects of social media use on mental health and wellbeing among young people in england: a secondary analysis of longitudinal data. The Lancet Child & Adolescent Health (2019)
49. Wandersman, A., Nation, M.: Urban neighborhoods and mental health: Psychological contributions to understanding toxicity, resilience, and interventions. American Psychologist **53**(6), 647 (1998)
50. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: ACL (2012)
51. Waseem, Z.: Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In: NLP-CSS (2016)
52. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: NAACL (2016)
53. Wijeratne, S., Balasuriya, L., Sheth, A., Doran, D.: Emojinet: An open service and api for emoji sense discovery. In: AAAI-ICWSM (2017)