Challenges in Modifying Existing Scales for Detecting Harassment in Individual Tweets

Joshua Guberman
Illinois Institute of Technology
jguberma@hawk.iit.edu

Libby Hemphill
Illinois Institute of Technology
libby.hemphill@iit.edu

Abstract

In an effort to create new sociotechnical tools to combat online harassment, we developed a scale to detect and measure verbal violence within individual tweets. Unfortunately, we found that the scale, based on scales effective at detecting harassment offline, was unreliable for tweets. Here, we begin with information about the development and validation of our scale, then discuss the scale's shortcomings for detecting harassment in tweets, and explore what we can learn from this scale's failures. We explore how rarity, context, and individual coder's differences create challenges for detecting verbal violence in individual tweets. We also examine differences in on- and offline harassment that limit the utility of existing harassment measures for online contexts. We close with a discussion of potential avenues for future work in automated harassment detection.

1. Introduction

Online harassment is a continuing problem, endemic to many social media platforms and forms of online computer-mediated communications. A remarkable 40% of all adults and 32% of teenagers connected to the internet have experienced at least one type of online harassment [18,27]. For some people, the experience of online harassment is ongoing, lasting for weeks on end [30]. Though individuals who witness online harassment may be apathetic toward it, those on the receiving end are often extremely perturbed by the experience [18].

Online harassment can have severely deleterious effects on individuals. Among youths, cyberbullying is associated with school violence, suicidal ideation, offline victimization, substance abuse, as well as other negative effects [28]. Studies of the effects of online harassment on adult victims are not as numerous as those focusing on teens and adolescents, but mainstream media outlets frequently relay the stories of adult victims. As of late, in part due to an online movement often characterized as a concerted effort to harass women [see 28] media and academic attention to online harassment have increased. A number of recently publicized cases of harassment have been so extreme that women have fled their homes for fear of their safety [23,31,34]. In one particularly distressing case, sustained online harassment may have been a factor in a young woman's suicide [13].

Within this paper, we report on the rationale for and the process of developing a scale for measuring verbal violence in individual tweets so that we may eventually automatically detect malicious content. We detail the problems we encountered that stemmed from faulty assumptions and methodological problems, we explain why it was difficult to reach appropriate levels of agreement on what constituted verbal violence, and we consider the utility of using Amazon's Mechanical Turk (MTurk) for scale validation purposes. We share this experience with the hope that our mistakes may help to steer others in the right direction. We close with a discussion of avenues for future work in verbal violence detection and measurement.

2. Background

Gender-based online harassment is not a new occurrence and has been observed and recorded since the early days of online computer-mediated communications [25]. Popular platforms such as Twitter that permit users to easily obfuscate their real identities may beget such harassment [5,46]. That harassing and abusive messages sent over the internet can reach so many people in such a short time makes managing such harassment an onerous task [11]. Manual reporting and commercial content moderation (CCM) [see 34] are currently the most common approaches to combatting harassment [14]. In this model, it is either incumbent upon the victim or a third-party observer of harassment to report and/or manage the harassing content. Automated detection efforts have so far had limited success [see, e.g., ,15,35,41,45], but machine learning approaches hold promise [35,39,40].

URI: http://hdl.handle.net/10125/41422
ISBN: 978-0-9981331-0-2

Under the existing manual moderation model, the burden too-often falls on underpaid CCM workers [36] to act as gatekeepers, screening content (and seeing everything) so that the end-user does not have to [37]. Harassment continues despite these in-place reporting measures [30], which can exact a heavy toll on the CCM workers tasked with evaluating content. These workers are subjected to the worst that humanity has to offer within online environments. They report a multitude of emotional disturbances including developing existential dread, intrusive imagery, and desensitization to [12,36,37]. Mental health services are often not provided to these workers [12]. We hoped to devise more effective means of content moderation that insulates both the would-be victims as well as the content moderators from the effects of verbal violence.

By leveraging data generated by Twitter users to increase our understanding of and ability to detect toxicity and verbal violence as they occur on Twitter, we hoped to develop sociotechnical tools for combatting online harassment. Our initial approach involved hand-coding content with the end-goal of using human-labeled data to train machine-learning classifiers to automate the detection and management of malicious Tweets. Although, for a number of reasons, our attempt was ultimately unsuccessful, we gained a deeper understanding of just how complex online harassment is as well as how difficult it is to detect and manage.

3. Methods

3.1. Developing a scale

We identified individual tweets as an ideal place at which to detect verbal violence. Individual tweets constitute a discrete unit of analysis and the ability to manage content at this level would be both useful and computationally inexpensive. Despite the existence of several decades' worth of extant literature on online harassment, we were unable to find any metrics of harassment or cyber-aggression that could be readily applied to individual tweets. As such, we endeavored to create our scale.

There are existing approaches for detecting harassment in online communities (such as Slashdot, MySpace [45], and "social news" sites [41]), and for user models on Twitter [16]. Our approach of coding tweets for content, then using MTurk to code more broadly, is in line with these other harassment detection efforts. Literature shows that profanity is often used in bullying attempts [12,41,45]. The automatic detection and blocking of profanity would be relatively easy to accomplish. However, that approach would yield many false positives. Hence, we sought a solution that would be sensitive to the context in which such keywords occur. To accomplish this more complex task, we decided to begin with human coders instead of automated methods. Our plan was to use human coders, able to capture these subtleties, in order to generate training data for automated classifiers that could aid moderators and users in the detection of violent content. Our scale was designed to detect harassment [see ,35 for a report on binary [present/absent] harassment detection] *and* to indicate the specific type of harassment occurring.

Although there are studies that focus on harassment occurring [30] and being discussed [4] on Twitter, none offered insights into how we might achieve context-aware harassment detection. As such, we turned to psychological literature regarding online harassment cyberbullying, aggression (both online and in-person), and the measurement thereof.

A large body of scholarly work focuses on the differences in individual traits and situational factors that predispose individuals to perpetrate aggressions [see 6,31]. The same trait differences that correlate with aggression also relate to the perpetration of online harassment [33]. In addition to the shared personality correlates of offline-aggression and online harassment, the dynamics of cyberbullying bear a close resemblance to those of face-to-face bullying. In both cases, the perpetrator intends to harm its victim [20]. Thus, online harassment constitutes a human aggression [1].

Because online harassment is a manifestation of aggression, we felt justified in modeling our scale on existing measures of face-to-face aggression. We chose to model our scale items primarily on the Buss-Perry Aggression Questionnaire (BPAQ) due to its high reliability and internal consistency [10] and because it has been used to check the criterion validity of existing measures of cyber-aggression [11]. Several additional items pertaining to doxing, rumoring, and the sending of threatening reactions were adapted from the Cyber Victim Bullying Scale (CVBS) [11] and the Facebook Aggression Measure [33]. We objectified the language used in the original scale items for use on tweets. For example, we changed the BPAQ item "I can't help getting into arguments when people disagree with me" to read as "User shares personal opinions about people, groups, or institutions that the user disagrees with." This allowed for human coders to read a given tweet and respond to each of the scale items on a Likert-style scale where lower numbers indicate that the item is uncharacteristic of the tweet and higher numbers indicate that the item is characteristic of the tweet. We omitted from our scale items from the BPAQ that could not be objectified in a way that would allow for coding by a third party. These included most of the items

related to physical aggression (e.g. "Once in a while, I can't control the urge to strike another person"), as well as some items related to hostility (e.g. "I wonder why sometimes I feel so bitter about things") [10]. We also created an item related to *ad hominin* attacks based on the number of such tweets we saw when reading through our dataset. Our item pool initially consisted of 18 items and was eventually reduced to 14 items in the final revision of the scale (see Table 1 for the final scale items and their levels of agreement between coders).

It should be noted here that the BPAQ measures *trait* aggression. In this approach, we had hoped that the types of aggression observed in tweets would correlate with the behaviors captured by the BPAQ. As we will explain in this paper's discussion, this approach was unsuccessful.

3.2. Collecting data

Using TwitterGoggles [29], we collected millions of tweets containing several hashtags, including but not limited to #GamerGate and #NotYourShield. Because of the media coverage of harassment coming from both sides of the #GamerGate controversy, we believed that #GamerGate tweets would provide an ideal population of tweets for the development and testing of new means to study and detect toxicity on Twitter.

3.3. Training human coders

We turned to MTurk to expedite the coding of our dataset. Prior to hiring Turkers to participate in our study, we received approval to do so from the Illinois Institute of Technology's institutional review board.

To gain eligibility for our tweet-coding tasks, Turkers were required to complete an online training program using Qualtrics and disseminated via MTurk. The program consisted of three components. First, Turkers were shown a mockup of the coding form for a tweet already rated by the authors. Detailed explanations of the authors' rationale for each rating were provided. When ready, Turkers proceeded to the next page where they were given a blank coding form and asked to rate the tweet from the previous page. Turkers were not permitted to go back to the previous section to check the authors' ratings. Those who agreed with the authors' ratings on at least 12 out of the 14 scale items were permitted to move on to the final component. The Turkers who performed satisfactorily were then given a new tweet and another blank coding form. This tweet had already been coded by the authors, but Turkers were not permitted to know the authors' ratings. Those whose ratings were in agreement with our own on at least 12 out of the 14 items "passed" the training program and were granted an MTurk qualification that allowed them to work on subsequent tweet-coding tasks. Instituting this program increased between-coder reliability (see [22]). All participating Turkers were compensated $2.50 through MTurk for attempting the training program, which took an average of ~14 minutes to complete.

3.4. Human coding

Once we reached reasonable reliability between coders on our modified scale, we proceeded to have human coders, recruited and trained through MTurk, rate 900 tweets (see [14]) from our #GamerGate dataset. Each tweet was coded only once. A total of six human coders participated in the coding process (see Table 2 for information about the coders). The coders made a total of 10,771 "Uncharacteristic" ratings, 1,535 "Characteristic" ratings, and 294 "I'm not sure" ratings. Turkers were paid $0.75 per tweet, each of which took ~1 minute to code. We also provided coders a large text box in which to enter comments on their ratings, and we provide many examples of those comments here.

Each coded tweet was given a composite aggression score, accounting for each scale item that was coded as being "characteristic" of the tweet. Possible scores ranged from 0-14, with higher scores indicating higher levels of aggression present in a tweet. Aggression scores for the coded tweets ranged from 0-9 (M = 1.7, SD = 2.24). Using these 900 coded tweets as training data, we hope to build machine-learning classifiers for the scale items.

4. Explaining disagreements among coders

The third column of Table 1 shows the degree to which coders agreed with one another on a practice round of coding tasks consisting of 20 tweets. Each of the 20 tweets was coded by two independent coders. As you can see, between-coder agreement varied greatly by item. However, the average agreement score across all 14 items reached 70%, which we felt was suitable for a first pass at coding the dataset. We identified four primary mechanisms for explaining the disagreement between coders we witnessed: rare events, insufficient context, questions of audience, and individuals' perceptions.

4.1. Rare events

The two items with the highest level of between-coder agreement are items 1 and 2 (see Table 1). This

does not reflect the ease with which coders were able to apply this item to tweets, but rather the fact that almost no tweets in our dataset appear to be characteristic of these two scale items. In our batch of 900 coded tweets, only two tweets were coded as being characteristic of either of these two items. To our surprise, both tweets were coded as being characteristic of *both* items 1 and 2. Though both items are similar in that they relate to threats of physical violence, one requires the threat to not be made as a means to protect one's rights, while the other requires that the threat is *not* made for the aforementioned reason.

Together, the practice-coding agreement levels and the coding of these two items in the batch of 900 tell us several things. First, the proportion of #GamerGate tweets containing threats of physical violence appears to be quite low. Second, it tells us that humans are reliably able to agree on the *absence* of violent threats in tweets but not the *presence* of violent threats. Given the rarity of some kinds of harassment (e.g., threats of physical violence), the agreement levels may overestimate actual agreement because chance agreement is so likely for uncharacteristic tweets. Even when coders do detect violent threats and code tweets accordingly, they are unable to discriminate between the motives for the physical threats. This lack of discrimination may be a function of the 140-character limit imposed on tweets.

Similar to items 1 and 2, item 9 which related to the public disclosure of private information (i.e., doxing) was rare within our coded sample (N = 2). However, we suspect based on comments provided to us by our coders that the actual rate of occurrence may be slightly higher. One of our coders wrote,

> It looks like this user may have shared personally identifiable information and embarrassing images of someone else, but that info wasn't included in this particular conversation so I didn't rate those sections as characteristic.

Thus, it is important to note that the absence of information, such as pictures, that was originally included in a tweet but is now missing, may have caused the misclassification of tweets on some items.

4.2. Insufficient context

Other sources of disagreement are likely related to the lack of specific context human coders have access to when reading a 140-character string of text which may or may not include links to other text or images, or to information about the author. We provided coders with both the text from a tweet and the URL to view the tweet online. We asked that coders follow the provided links whenever possible to gain additional context (i.e., to see if a tweet is part of a thread to determine if it's argumentative), but we have no way to know if or how often coders actually followed the links. We do know that some coders followed the links, as we received a number of comments from coders relating to dead-links making tweets hard to code. Coders specifically referred to difficulties relating to lack of context 42 times, and to dead links/missing content 107 times.

Another potential source of context (broadly) for the tweets in our sample is knowledge about #GamerGate. We did not ask our coders to rate their level of familiarity with #GamerGate, as we were concerned that seeing references to #GamerGate before coding would prime individuals with strong opinions to code differently. However, our attempt to avoid priming effects may have introduced more variance into our rating dataset. For instance, ratings for the scale items for which between-coder agreement was less than 70% may have been influenced by the coders' knowledge of #GamerGate. This conclusion is supported by a number of comments provided by Coder 2, for example:

> Rated 13 & 14 as characteristic because both of this user's tweets in the conversation seem to indicate that the user thinks GamerGate is being misrepresented as a group that dislikes games.

> If it's not clear, I rated #14 characteristic because the user is defending their (and other pro-Gamergate individuals') stance as being for ethical journalism instead of against women in gaming.

Clearly, her ratings are influenced by her understanding of the differences between the two main sides involved in the #GamerGate controversy. The effects of the lack of prior understanding of the topic can best be shown by explaining the coding of a tweet that requires prior knowledge. Take the following tweet text, for instance:

> New to #GamerGate? We love inclusivity & diversity. Notice how our opponents are all left wing authoritarians, telling you what to think? [32]

If you are familiar with #GamerGate, you realize that the statement about the nature of gamergaters (love inclusivity and diversity) is likely in response to comments by the media and by other Twitter users suggesting otherwise. If so, this tweet may be

characteristic of items 11-14, depending on how one interprets the items. It may also be considered characteristic of item 5, if one considers the act of providing contrary information to be equivalent to engaging in an argument. Or, the tweet is a sarcastic response mocking gamergaters and could be coded characteristic of items #5 and #11. Without an understanding of the issues surrounding #GamerGate, however, it is unlikely that a coder would rate any of these items as being characteristic of this particular tweet.

4.3. Promotion and audience

Sarcasm is just one challenge to interpreting the text of a tweet. Coders commented on a number of Twitter conventions that figured into their decisions about what codes to assign. For instance, they disagreed whether posting a link necessarily implies support for the content at the link. Coders commented:

> Linked article suggests unfair treatment of twitter poster's group. Linking of article is tacit defense of tweeter's group and image.

and

> The tweet itself may not be inflammatory or contain any opinions, but the link itself does. Since this person is trying to spread the link, then regardless of whether they actually wrote up the content in that link I think this tweet counts as inflammatory and retaliatory content.

While another said,

> The links themselves are definitely pro-GG and share some opinions of the opposing side, but because this user seems to just be posting these links to be "informative" and doesn't directly share any opinions of his own I didn't rate this tweet as being inflammatory in any way.

It was difficult for coders to decide whether sarcastic or informative tweets constituted attempts to start an argument. The conventions around link sharing in Twitter are developing, and these comments highlight the challenge in detecting whether posting a link is supportive. Some of that detection boils down to context as mentioned earlier, but we saw other Twitter-related disagreements that indicate something unique about Twitter (and it's #GamerGate discussions specifically) are at play here: publicness.

As one coder points out, tweets are public even when they contain @mentions or @replies:

> Just wanted to make a point about the "starting an argument" question. In this instance the tweeter is responding directly to someone he sides with. However, I'm taking the tweet to be "public" and therefore readable by, and somewhat directed at, people not necessarily in agreement with his comments…I'm assuming that because tweets are public, they are de facto made to the broader population, especially when they include a hashtag, and not just the individual person they might be addressed to or responding to.

The public nature of tweets complicates the question of audience, and for topics such as harassment, the audience is of particular importance. Twitter accounts and users can conflate individuals and groups when assigning authorship to tweets, further complicating the notion of audience. For instance, accounts for companies, celebrities, and politicians are at once individual and institutional. What it means for an institution to be the target of harassment was an issue our coders faced:

> If it was directed at an individual it wouldn't be characteristic. But since it's directed at a company, whose reputation is partly tied to their business, I chose not sure.

Whether or not a post or link constitutes promotion and how to judge whether a target or author is a group or individuals are problems unique to the online context of harassment.

4.4. The eyes of the beholders

Another source of between-coder differences in ratings may result from individual differences. A one-way analysis of variance (ANOVA) shows that there is a small but significant difference between the overall aggression scores of tweets coded by women versus those coded by men. Women find, on average, tweets to be more aggressive ($F(1,898) = 10.286$, $p = .001$). This difference is compatible with findings that women are better able than men to detect more subtle forms of aggression (i.e., microaggressions), possibly because women are, unfortunately, more likely to have personally experienced certain types of microaggressions [2]. Women are also more likely to accurately (based on legal definitions) perceive a wider range of potentially ambiguous behaviors as harassment [38].

However, this observed difference in our data is not necessarily a function of gender. It may be the case that women were simply given more aggressive (according to our scale) tweets to code than men were. It is difficult to make definitive conclusions given the small number of coders we employed and the number of tweets they coded.

Other cognitive factors related to individual differences in the perception of aggression are likely at play as well. Perceptions of external stimuli appear to be influenced by individuals' attributions of intent. Hostile attributional bias refers to a tendency for some individuals to interpret ambiguous stimuli as being intentionally aggressive and is found in both children and adults [17,19]. Additionally, people with angry or anxious dispositions are more likely to interpret ambiguous prose as being negative [43]. It is possible that our coders fell within one of these populations. However, given the rarity of such events, it seems more reasonable to infer that coders may underestimate the harassment that occurs rather than overestimate it.

We initially thought it possible that different users found various tweets "funny" rather than "malicious" or "violent," but existing research suggests that people generally agree when sexual humor is offensive [24]. We do not know, however, how people decide whether other types of humor are offensive rather than funny. In the #GamerGate dataset, posts about where a person lives, how many friends a person has, and whether a person does drugs or engages in other illegal activities are also mentioned (in addition to sexual content). Our coders were not sure what to do with these kinds of tweets, as evidenced by this comment:

> The tweet is an allusion to speculation that […] a gaming journalist […] was using cocaine at a press conference (I believe E3). I consider this in the grey area of posting a potentially reputation-damaging rumor, because while that rumor could certainly be reputation-damaging, the post seems to be mostly made in jest.

This comment also indicates that tweets made in jest are not real in their consequences–the potential to damage one's reputation is mitigated by the jest here.

5. Discussion

We found that individual tweets were not reliably categorized by multiple coders, at least not using existing measures of harassment. While the "rare events" problem could potentially be solved with more data, the lack of context and the variation in individual perceptions of malicious content pose potentially insurmountable challenges for this "individual tweets" approach to manual harassment coding. Without reliable labeled data, it will be difficult to construct supervised learning classifiers using this approach.

Given the rarity of harassment relative to all kinds of posts on Twitter, 900 tweets were likely not enough data to train an automated classifier effectively. However, 900 tweets were enough to reveal some patterns in disagreement between coders, as we have described above. Coding more tweets could potentially increase our ability to detect harassment, but it is not clear, given all the kinds of disagreement we documented, that the marginal benefits of doing so are worth the costs (in either computation or Turker time).

5.1 Labelling users versus labelling content

Existing tools take the "individual user" approach to content control. For instance, Twitter currently provides a blocking[1] feature that allows a user to prevent others from following them and a muting[2] feature that prevents another users' content from appearing. Both of those features operate at the user level. Rather, we were trying to label content, so that new tools would allow users to avoid certain types of posts instead of avoiding certain users altogether. This content approach would be useful in a number of scenarios including

- *doxing* – if I mute an account who doxes me, I won't know it happened
- *disagreement* – I may be willing to engage in arguments as long as I'm not being physically threatened.

The "individual tweets" approach to detecting verbal violence assumes that an individual utterance can be violent (or at least exhibit violent characteristics) without labeling the speaker "violent." The BPAQ [10], on which our scale is based, is a measure of trait aggression and considers aggression as a personality trait assumed to be correlated with acts of aggression. Our results indicate that violent traits in content are not readily analogous to violent traits in people. We used this approach in order to avoid labeling individual users as "violent," but it was challenging for coders to detect violence in the absence of information about the users and their other behaviors and opinions.

Prior research on cyberbullying has also taken a user approach, labeling users as bullies and even using content from multiple platforms to build user models [15,16]. Labeling users also risks a "whack-a-mole"

[1] https://support.twitter.com/articles/117063

[2] https://support.twitter.com/articles/20171399

problem in which individual accounts are abandoned as soon as they are labeled "bullies," and the offending user just opens new accounts to continue the behaviors. Labeling content has the potential to enable us to build tools that let users set individualized thresholds for particular types of tweets without encouraging throwaway account creation.

5.2. Translating existing measures of computer-mediated communication

A number of items on existing harassment measures were poor fits for user-generated content. This finding bodes poorly for the method of adapting existing measures of aggression, cyber-aggression, and cyberbullying for Twitter. Rather than using items from existing scales, it may be beneficial to create items based on the types of harassment actually observed in the data.

For example, item #10 (attacking credibility in order to undermine) was created based on our observations of the data and was a better fit than many of the items adapted from other measures. In contrast, the distinction between items 1 and 2 did not translate from offline to Twitter. We suggest that future attempts at measuring verbal violence on Twitter take a *"bottom up"* or grounded approach in which coders first identify the kinds of harassment occurring and then build a model. Further, validity is always a concern when using or developing measures of personality [26]. Avoiding adapting personality measures in favor of a grounded approach reduces the possibility of having a highly reliable but invalid measure.

6. Conclusion and future work

We have discussed how rarity, context, audience, and individual differences create challenges for detecting verbal violence in individual tweets. We have also identified differences in how on- and offline harassment unfold, thus limiting the utility of adapting existing harassment measures for online contexts. We are still committed to combatting harassment, though, and think that identifying when and how it occurs remain important first steps in that battle. We now turn to promising avenues for future research.

First, we could return to the "individual users" approach to detecting harassment. By rating tweets from a single user we could determine whether the *user* is aggressive by using existing measures. These results could be cross-validated with the BPAQ by rating a user's tweets, and then having the same user complete the BPAQ. This approach would at least measure whether a user's content matches their personality. One study shows that at least one "real-life" personality trait often thought to be associated with aggression, narcissism (see [3,9,30,35]), persists in online environments; this is reflected in how people scoring high in narcissism conduct themselves on Facebook [8]. Even among the authors of this paper, however, there is disagreement about the utility of this approach given that people may behave differently in different online communities where norms of behavior are different [6,7].

Situational differences are a challenge for all psychological measures, though, and second, we suggest future work consider the social situation in which users operate. For instance, we could use tweets' context such as the volume and velocity of tweets, the number of accounts involved in a discussion, and the number of similar tweets sent to multiple people simultaneously to detect harassment. Each of these represents a way in which harassment online manifests differently from harassment offline. Online harassment, especially under the #GamerGate tag, often involves many people targeting a single individual instead of one person harassing one other person (i.e., dogpiling) and floods of tweets [48].

Lastly, we could examine various groups or conversations of tweets instead of focusing on individual utterances. A coding scheme like the Perpetrator-Act-Target (PAT) scheme [47], first developed for detecting violence on television, could potentially be applied to conversations. The PAT coding scheme takes a holistic approach to coding for violence, measuring violence at three separate levels: (1) the individual act (with a focus on the perpetrator, act, and target), (2) the scene in which an act(s) occurs, and (3) the complete program that the various scenes comprise. To apply a similar hierarchal scheme here, coders could simultaneously label the individual tweet within a conversation *and* the conversation as a whole. This would give us a two levels of detection – the individual tweet level and the conversation/thread level – while still avoiding labeling individual users/accounts.

These areas of future work represent different approaches to improving our harassment response tools. The first improves on the user-labeling tools, potentially leading to new automated blocking or muting functions. The second leverages the unique features of online harassment to afford system-level tools that detect a situation in which harassment is likely to occur. The third considers the conversational context of the tweet to both improve coding and add a level of analysis. Approaches that combine information about users, situations, and conversations will likely be more effective. A combination of these approaches in which we attend to both users and their situation will

likely be most useful and emphasizes both the technical and social aspects of the response to harassment.

7. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1525662.

8. References

[1] Anderson, C.A. and Bushman, B.J. Human Aggression. *Annual Review of Psychology 53*, 1 (2002), 27.

[2] Basford, T.E., Offermann, L.R., and Behrend, T.S. Do You See What I See? Perceptions of Gender Microaggressions in the Workplace. *Psychology of Women Quarterly 38*, 3 (2014), 340–349.

[3] Baumeister, R.F., Bushman, B.J., and Campbell, W.K. Self-Esteem, Narcissism, and Aggression Does Violence Result From Low Self-Esteem or From Threatened Egotism? *Current Directions in Psychological Science 9*, 1 (2000), 26–29.

[4] Bellmore, A., Calvin, A.J., Xu, J.-M., and Zhu, X. The Five W's Of "Bullying" on Twitter: Who, What, Why, Where, and When. *Computers in Human Behavior 44*, (2015), 305–314.

[5] Bishop, J. The Effect of De-Individuation of the Internet Troller on Criminal Procedure Implementation: An Interview with a Hater. *International Journal of Cyber Criminology 7*, 1 (2013), 28.

[6] Boyd, D. *It's complicated: the social lives of networked teens*. Yale University Press, New Haven, 2014.

[7] Bruckman, A. Finding One's Own Space in Cyberspace. *Technology Review 99*, 1 (1996), 48–54.

[8] Buffardi, L.E. and Campbell, W.K. Narcissism and Social Networking Web Sites. *Personality and Social Psychology Bulletin 34*, 10 (2008), 1303–1314.

[9] Bushman, B.J., Baumeister, R.F., Thomaes, S., Ryu, E., Begeer, S., and West, S.G. Looking Again, and Harder, for a Link Between Low Self-Esteem and Aggression. *Journal of Personality 77*, 2 (2009), 427–446.

[10] Buss, A.H. and Perry, M. The Aggression Questionnaire. *Journal of Personality and Social Psychology 63*, 3 (1992), 452–459.

[11] Çetin, B., Yaman, E., and Peker, A. Cyber Victim and Bullying Scale: A Study of Validity and Reliability. *Computers & Education 57*, 4 (2011), 2261–2271.

[12] Chen, Y., Zhou, Y., Zhu, S., and Xu, H. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, (2012), 71–80.

[13] Cohen, N. and Spoto, M. Transgender N.J. Game Developer Jumps from GWB After Online Bullying. *NJ.com*, 2015. http://www.nj.com/monmouth/index.ssf/2015/04/post_18.html.

[14] Crawford, K. and Gillespie, T. What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society*, (2014), 1461444814543163.

[15] Dadvar, M. and de Jong, F. Cyberbullying Detection: A Step Toward a Safer Internet Yard. *Proceedings of the 21st International Conference on World Wide Web*, ACM (2012), 121–126.

[16] Dadvar, M., Ordelman, R., Jong, F. de, and Trieschnigg, D. Towards User Modelling in the Combat against Cyberbullying. In G. Bouma, A. Ittoo, E. Métais and H. Wortmann, eds., *Natural Language Processing and Information Systems*. Springer Berlin Heidelberg, 2012, 277–283.

[17] De Castro, B.O., Veerman, J.W., Koops, W., Bosch, J.D., and Monshouwer, H.J. Hostile Attribution of Intent and Aggressive Behavior: A Meta-Analysis. *Child Development 73*, 3 (2002), 916–934.

[18] Duggan, M., Rainie, L., Smith, A., Funk, C., Lenhart, A., and Madden, M. *Online Harassment*. Pew Research Center, 2014.

[19] Epps, J. and Kendall, P.C. Hostile attributional bias in adults. *Cognitive Therapy & Research 19*, 2 (1995), 159–178.

[20] Grigg, D.W. Cyber-Aggression: Definition and Concept of Cyberbullying. *Australian Journal of Guidance & Counselling 20*, 2 (2010), 143–156.

[21] Guberman, J. and Hemphill, L. Descriptors and Measurements of Verbal Violence in Tweets. 2016. https://figshare.com/articles/Descriptors_and_measurements_of_verbal_violence_in_tweets/3179368.

[22] Guberman, J., Schmitz, C., and Hemphill, L. Quantifying Toxicity and Verbal Violence on Twitter. *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, ACM (2016), 277–280.

[23] Hart, A. Yet Another Game Developer Flees Her Home After Death Threats. *The Huffington Post*, 2014. http://www.huffingtonpost.com/2014/10/11/game-developer-death-threats_n_5970966.html.

[24] Hemmasi, M., Lee Graf, A., and Russ, G.S. Gender-Related Jokes in the Workplace: Sexual Humor or Sexual

Harassment?1. *Journal of Applied Social Psychology 24*, 12 (1994), 1114–1128.

[25] Herring, S.C. The Rhetorical Dynamics of Gender Harassment On-Line. *The Information Society 15*, (1999), 151–167.

[26] Kerlinger, F. Objective Tests and Scales. In *Foundations of Behavioral Research Second Edition*. Holt, 1973.

[27] Lenhart, A. Cyberbullying. *Pew Research Center's Internet & American Life Project*, 2007. http://www.pewinternet.org/2007/06/27/cyberbullying/.

[28] Lenhart, A. Cyberbullying 2010: What the Research Tells Us. *Pew Research Center: Internet, Science & Tech*, 2010. http://www.pewinternet.org/2010/05/06/cyberbullying-2010-what-the-research-tells-us/.

[29] Maconi, P., Hemphill, L., and Goggins, S. *TwitterGoggles*. 2015.

[30] Matias, J.N., Johnson, A., Boesel, W.E., Keegan, B., Friedman, J., and DeTar, C. *Reporting, Reviewing, and Responding to Harassment on Twitter*. Women, Action, and the Media, 2015.

[31] McDonald, S.N. Gaming Vlogger Anita Sarkeesian Is Forced from Home After Receiving Harrowing Death Threats. *The Washington Post*, 2014. http://www.washingtonpost.com/news/morning-mix/wp/2014/08/29/gaming-vlogger-anita-sarkeesian-is-forced-from-home-after-receiving-harrowing-death-threats/.

[32] @Nephanor. New to #GamerGate? We love inclusivity & diversity. Notice how our opponents are all left wing authoritarians, telling you what to think? *@Nephanor*, 2014. https://twitter.com/Nephanor/status/527694154671734784.

[33] Pabian, S., De Backer, C.J.S., and Vandebosch, H. Dark Triad Personality Traits and Adolescent Cyber-Aggression. *Personality and Individual Differences 75*, (2015), 41–46.

[34] Parkin, S. Zoe Quinn's Depression Quest. *The New Yorker*, 2014. http://www.newyorker.com/tech/elements/zoe-quinns-depression-quest.

[35] Reynolds, K., Kontostathis, A., and Edwards, L. Using Machine Learning to Detect Cyberbullying. *10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*, IEEE (2011), 241–244.

[36] Roberts, S. Social Media's Dirty Work: Contextualizing the Facebook Screening Controversy. *Sarah T. Roberts | The Illusion of Volition*, 2012. https://illusionofvolition.com/2012/02/26/social-medias-dirty-work-contextualizing-the-facebook-screening-controversy/.

[37] Roberts, S. Commercial Content Moderation: Digital Laborers' Dirty Work. *Media Studies Publications*, (2016).

[38] Rotundo, M., Nguyen, D.-H., and Sackett, P.R. A meta-analytic review of gender differences in perceptions of sexual harassment. *Journal of Applied Psychology 86*, 5 (2001), 914–922.

[39] Smets, K., Goethals, B., and Verdonk, B. Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. *AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, (2008), 43–48.

[40] Sood, S., Antin, J., and Churchill, E. Profanity Use in Online Communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 1481–1490.

[41] Sood, S.O., Churchill, E.F., and Antin, J. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology 63*, 2 (2012), 270–285.

[42] Stucke, T.S. and Sporer, S.L. When a Grandiose Self-Image Is Threatened: Narcissism and Self-Concept Clarity as Predictors of Negative Emotions and Aggression Following Ego-Threat. *Journal of Personality 70*, 4 (2002), 509–532.

[43] Wenzel, A. and Lystad, C. Interpretation biases in angry and anxious individuals. *Behaviour Research and Therapy 43*, 8 (2005), 1045–1054.

[44] Wofford, T. Is GamerGate About Media Ethics or Harassing Women? Harassment, the Data Shows. *Newsweek*, 2014. http://www.newsweek.com/gamergate-about-media-ethics-or-harassing-women-harassment-data-show-279736.

[45] Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., and Edwards, L. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB 2*, (2009), 1–7.

[46] Zhong, C.-B., Bohns, V.K., and Gino, F. Good Lamps Are the Best Police: Darkness Increases Dishonesty and Self-Interested Behavior. *Psychological Science 21*, 3 (2010), 311–314.

[47] National Television Violence Study. In *National Television Violence Study*. Sage Publications, Thousand Oaks, CA, 1998, 384.

[48] Is GamerGate About Media Ethics or Harassing Women? Harassment, the Data Shows. *Newsweek*. http://www.newsweek.com/gamergate-about-media-ethics-or-harassing-women-harassment-data-show-27973

Table 1. Our scale items and between-coder agreement for each item

#	Item	Agreement*
1	User threatens physical violence as a means of protecting the user's rights.	94%
2	User threatens other people or groups of people with physical harm and/or sexual violence.	94%
3	User openly expresses disagreement.	56%
4	User shares personal opinions of people, groups, or institutions that the user disfavors.	56%
5	User engages in or attempts to start arguments with people, groups, or social movements that the user disagrees with.	63%
6	User tweets potentially reputation damaging rumors about something else	94%
7	User tweets non-physical threats or threatening reactions to or about someone.	94%
8	User shares potentially embarrassing photos or videos of someone else.	88%
9	User shares someone else's personally identifiable information.	94%
10	User attacks the credibility of another person or group of people in an attempt to invalidate the other party's stance or argument.	63%
11	User writes retaliatory comments in response to another person or group's words or actions	69%
12	User expresses feelings that user or a group that user belongs to is being treated unfairly	44%
13	User expresses feelings of being misrepresented and/or under-represented by other people, groups of people, the media, etcetera.	44%
14	User defends user's self or user's image, or the image of a group that the user belongs to or associates with.	31%

* Agreement percentages are indicative of the overall level of between-coder agreement on all 14 ratings across 20 tweets. Kappa statistics are not provided, as they do not provide useful information given the low number of coders per tweet.

Table 2. Information about our human coders

Coder ID	Age	Gender	Tweets Coded (N)
1	24	Man	274
2	23	Woman	306
3	37	Man	123
4	32	Woman	13
5*	27	Man	183
6	36	Man	1

* Coder 5 reported multiple ages and genders but most frequently identified as a 27-year-old man