First Steps in Quantifying Toxicity and Verbal Violence on Twitter

Joshua Guberman

Illinois Institute of Technology
Chicago, IL 60616, USA
jguberma@hawk.iit.edu

Carol Schmitz

Illinois Institute of Technology
Chicago, IL 60616, USA
cschmitz@hawk.iit.edu

Libby Hemphill, PhD

Illinois Institute of Technology
Chicago, IL 60616, USA
lhemphil@iit.edu

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
Copyright is held by the owner/author(s).
CSCW '16 Companion, February 27 - March 02, 2016, San Francisco CA, USA
ACM 978-1-4503-3950-6/16/02.
<http://dx.doi.org/10.1145/2818052.2869107>

Abstract

Online harassment is a continuing problem, endemic to many social media platforms and other means of web-based communications, and few means exist to analyze web content for instances of verbal violence and aggression. We are developing a scale of online aggression that can be applied to Twitter posts (tweets) and that is based on existing measures of trait aggression and cyberbullying. For the purpose of testing and validating our scale, we are relying on Mechanical Turk, an Amazon Web Service, through which we can enlist and pay workers to code our dataset of tweets. Preliminary results suggest that aggression in tweets is difficult for human coders to identify and that we have not reached consensus about what constitutes harassment online. We discuss our preliminary results and propose next steps such as scale modification and automated classifier development.

Author Keywords

verbal violence; online harassment; scale development; mechanical turk; Twitter

ACM Classification Keywords

H.5.3 [Group and Organization Interfaces]: Web-Based Interaction–*Evaluation/Methodology*

Introduction

Interest in online harassment is growing, due in part to the rise of #GamerGate and its related conversations. Regardless of goals or definitions of the movement, #GamerGate has indisputably directed a lot of media attention to various types of cyberbullying and harassment that occur on Twitter. We leverage data from #GamerGate to evaluate an aggression scale for tweets. In doing so, we hope to gain further insights into verbal violence on Twitter and to provide design recommendations for automated harassment detection systems to help both end users and moderators.

Existing Analyses of Harassment on Twitter

Presently, analyses of harassment and aggression associated with #GamerGate are limited to sentiment analyses [8, 2], an informal survey [1], and a three-week study of harassment on Twitter during a period of time that coincided with #GamerGate [6].

Few studies exist about harassment and aggression specific to Twitter. One pertains to machine learning and its applications for learning about bullying on Twitter. It provides insights into who talks about bullying, what types of bullying are tweeted about, reasons for tweeting about bullying, and where users posting such tweets are located [3].

In another study, Woman, Action, and the Media!’s (WAM!) analysis of reported Twitter harassment, took place during a period of time when the #GamerGate hashtag was popular. This study focuses exclusively on instances of harassment taking place on Twitter. During a three-week period, WAM! reviewed 811 reports of harassment [6]. Their findings suggest that much of the harassment on Twitter is ongoing and hard to curb,

despite in-place reporting measures. For nearly 30% of reporters in WAM!’s study, the reported harassment had been going on for more than three weeks [6]. The most frequently reported types of harassment, in descending order, were hate speech, publically sharing someone else’s private information, unspecified, and threats of violence [6].

Two studies of the sentiment of #GamerGate tweets show that most have negative sentiments [2, 8] and that #GamerGate supporters tweet more at various women in the video-game industry than they do at video-game journalists [8].

Despite the important findings of these analyses, little is known about the nature of harassing tweets or about the people who write them. The existing methods for labeling harassment require users to read through individual tweets. Our scale is designed to facilitate automatic detection and to help us understand patterns of harassment.

Scale Development

As a first step in developing our scale, we turned to existing measures of violence and aggression. There are many such measures in the psychology literature, most of which rely on self-reported responses. Initially, this posed a problem within the context of analyzing tweets: a tweet cannot answer a question about its mood or about the intentions of its author. As such, existing scales could not be directly applied to tweets. Our solution was to substitute objective language in some of the items from these existing scales. Most of the items in our scale are adapted from the Buss-Perry Aggression Questionnaire (BPAQ) [4] and the Facebook Cyber-Aggression Measure [7]. For example, the BPAQ

Participant Instructions for Tweet Coding HITs

In this HIT, you will read a tweet and rate it on a 14 item scale. You will rate each item a 3-point Likert style scale, where 1 = "Uncharacteristic [of this tweet]" and 3 = "Characteristic."

If the tweet contains a link, it is advised that you open the link into a separate tab or window of your web browser. Doing so may add context to the tweet and may influence your ratings. In the event of a dead link, please rate the tweet to the best of your ability based on the text in the tweet alone.

Click the source URL provided below the tweet to determine whether the tweet is part of a conversation. This may add additional context. Again, in the event of a dead link, please rate the tweet to the best of your ability based on the text in the tweet alone. If after doing this you are still unsure how to code an item, you may select "Not Sure."

item "I can't help getting into arguments when people disagree with me" [4] now reads as "User engages in or attempts to start arguments with people, groups, or social movements that the user disagrees with." Questions that could not be suitably reframed (i.e. "Some of my friends think I'm a hothead") were omitted. Like the BPAQ, the items of our scale fall into four categories: physical aggression, verbal aggression, anger, and hostility.

We implemented the scale on a random sample of tweets collected with TwitterGoggles [5]. After having a single coder manually code the same batch of 200 tweets twice, we edited some scale items for clarity. Items were coded on a 5-point Likert style scale, where 1 = "Very Uncharacteristic" and 5 = "Very Characteristic." We deleted items open to too much subjective interpretation, as well as those that failed code-recode tests.

To expedite the coding of our dataset and to validate our scale, we created a number of human intelligence tasks (HITs) on Amazon's Mechanical Turk (MTurk) web-service (see sidebar for instructions Turkers received). The initial batch of HITs consisted of 20 tweets randomly selected from our dataset. Each tweet was coded by two separate people. Coders earned \$1.00 for each completed HIT. There was little agreement (52%) between coders on this first batch.

In an effort to increase inter-rater reliability, we used MTurk's *qualifications* feature to run practice coding tasks.¹ In these tasks, workers were shown examples

¹ The first version of these tasks can be found at: https://iitpsychology.az1.qualtrics.com/SE/?SID=SV_2ogpIJ0Ikyghobj

of tweets we had already coded (this time on a 4-point scale), for which we provided detailed explanations of our ratings. After studying our examples, workers were asked to code several tweets themselves. We then compared workers' ratings to our own ratings. The participants who performed satisfactorily on the qualification test (agreeing with at least 70% of our own ratings) were granted a qualification that allowed them to code a second batch of 20 tweets. Though this produced more agreement (63.9% across all ratings), there was still less agreement between coders than we would have liked.

Ratings in the second round of HITs suggested that while coders agreed on the direction of the scale (i.e., whether or not something was characteristic of the tweet), they disagreed on the magnitude of the judgment. Therefore, we tested a 3-point Likert style scale on which 1 = "Uncharacteristic," 2 = "I'm Not Sure," and 3 = "Characteristic." In this round of qualification tests, we required 85% agreement to qualify. Fewer participants qualified for the subsequent coding tasks at this higher threshold, but those who did qualify agreed more often than coders in previous rounds (70% agreement across all ratings).

Discussion

In reading through the tweets in our dataset, it is clear that a fair amount of verbal violence does, in fact, occur in tweets containing #GamerGate. The amounts and kinds of violence present in our dataset remain unclear, as does our scale's ability to adequately detect aggression. Our current best measure of between-coder agreement is 70% (see sidebar on next page for more info on agreement measures and results).

Measures of Between-Coder Agreement

Agreement was calculated by dividing the total number of items in a given batch on which both coders' ratings were in agreement by the total number of rated items in a batch (14 rated items for each tweet in batch). Sample sizes in these preliminary coding batches were not large enough to allow for the calculation of more robust statistics of inter-rater reliability.

Round 1: 52% agreement between coders across all rated items.

Round 2: 63.9% agreement between coders across all rated items.

Round 3: 70% agreement between coders across all rated items.

In analyzing the results of our most recent coding HITs, several points of interest stand out. First, coders were present. For example, none of the tweets in this batch contained threats of bodily harm. Thus, on scale items related to such threats, coders were in near-perfect agreement that said items were "Uncharacteristic."

In cases for which the appropriate response *might* have been "Characteristic," coders often opted to answer "I'm Not Sure." There may be several reasons for this. First, lack of agreement among coders may indicate flaws in the scale. It is quite possible that the language in the scale is too ambiguous, and open to too much subjective interpretation. More interestingly, perhaps, the lack of agreement may indicate that people have varying ideas about what constitutes aggression, argumentativeness, and hostility in tweets.

With subsequent iterations of our scale, it will be imperative that coders agree upon working definitions of the terms found in scale items. We may benefit from incorporating concrete operational definitions of these terms into our practice/training exercises. Once coded, we hope to use coded tweets to train supervised machine learning classifiers to automatically detect harassment as it occurs. In working toward automating detection of online harassment on Twitter, we seek to reduce rates of online harassment, and to make the web a safer, more productive place.

References

1. @The_Nimbus. 2015. "Repost of results for my 50 day survey of #GamerGate tweets. INFO: <http://pastebin.com/Pg6vT3GS> pic.twitter.com/JKnSzgPhzh". Tweet. (17 June, 2015) Retrieved October 19, 2015 from

https://twitter.com/The_Nimbus_/status/611294904489246721

2. Andy Baio. 2014. 72 Hours of #Gamergate: Digging through 316,669 tweets from three days of Twitter's two-month-old trainwreck. *Medium*. Retrieved October 19, 2015 from <https://medium.com/message/72-hours-of-gamergate-e00513f7cf5d>
3. Amy Bellmore, Angela J. Calvin, Jun-Ming Xu, and Xiaojin Zhu. 2015. The five W's of "bullying" on Twitter: Who, What, Why, Where, and When. *Computers in Human Behavior* 44: 305–314. <http://doi.org/10.1016/j.chb.2014.11.052>
4. Arnold H. Buss and Mark Perry. 1992. The Aggression Questionnaire. *Journal of Personality and Social Psychology* 63, 3: 452–459. <http://doi.org/10.1037/0022-3514.63.3.452>
5. Phil Maconi, Libby Hemphill, and Sean Goggins. 2015. *TwitterGoggles*. Retrieved October 24, 2015 from <http://dx.doi.org/10.6084/m9.figshare.1352028>
6. J. Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, Reviewing, and Responding to Harassment on Twitter. *arXiv preprint arXiv:1505.03359*. Retrieved October 19, 2015 from <http://arxiv.org/abs/1505.03359>
7. Sara Pabian, Charlotte J. S. De Backer, and Heidi Vandebosch. 2015. Dark Triad personality traits and adolescent cyber-aggression. *Personality and Individual Differences* 75: 41–46. <http://doi.org/10.1016/j.paid.2014.11.015>
8. Taylor Wofford. 2014. Is GamerGate About Media Ethics or Harassing Women? Harassment, the Data Shows. *Newsweek*. Retrieved October 19, 2015 from <http://www.newsweek.com/gamergate-about-media-ethics-or-harassing-women-harassment-data-show-279736>