Workshop On Online Harassment

Jennifer Golbeck
University of Maryland
College Park, MD 20742, USA
jgolbeck@umd.edu

Allison Druin
National Park Service.
Washington DC 20024, USA
allisond@umd.edu

Joel Tetreault
Grammarly, Inc.
San Francisco, CA 94104, USA
tetreaul@gmail.com

CHI 2017, May 6–11, 2017, Denver, CO, USA.
ACM ISBN 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3027070

Abstract
Online harassment is pervasive and disruptive across social media. It ranges from mild teasing to death and rape threats accompanied by publicly posting the target's home address. This harassment seriously degrades the user experience on a platform; nearly all users are left feeling unsafe and participation decreases. In order to design methods to reduce online harassment, it is first necessary to know what it is. This requires large labeled datasets of harassing comments that can be used in training and for analysis. This workshop will be a true working event, where participants will develop codebooks and create the first steps in a public dataset that can be used to fight online harassment.

Author Keywords
Online harassment, social media

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Background
Online harassment is a major part of online life, especially for women and minorities. There is no straightforward, single solution to the problem. Ultimately, it will require combined efforts in artificial intelligence, computational linguis-

tics, sociology, communications, social network analysis, interface design, and online communities researchers to understand the full scope of the issue and create technical solutions to address it.

This issue is of particular importance to the HCI community because (1) harassment seriously impacts the way people interact with and through online environments and (2) **it will ultimately be interface design that helps address this problem**. No technologies will chase online harassers from the internet. Human moderation does not scale to even moderate usage, let alone to large scale social media sites. Automated filtering can help with scale, but unlike spam, it is unlikely that it will be desirable for all types of harassment to be blocked by default. Some of the offensive content (racist, sexist, homophobic, etc.) is desirable for some users. They may enjoy consuming it or read it to understand the proliferation of these kinds of comments online. Ultimately users will need control over the type of content they block, a way to adjust the aggressiveness of the filters, and an understanding of how the filters are working.

Before any of that work can begin, it is necessary to understand exactly what is happening. That means all interested researchers will need a large, diverse, representative corpus of online harassment data to work with.

The purpose of this workshop is to take a major step forward in building a publicly accessible online harassment corpus.

The workshop will bring together researchers who have this data, existing taxonomies or typologies of online harassment, and those who simply want to work on building this resource. Those who bring data to the table could include organizations with large volumes of blocked or flagged content, researchers who have been building their own databases of offensive or hateful content, and those who have large general databases of social media content from which they can extract harassing information.

Our coding schemes will build on existing codebooks for various types of harassment, including [5, 4, 2, 3, 1]

Our workshop goals are to:

- Develop a shared vocabulary of harassment types (hate speech, individual harassment, threats, deeply offensive language, etc.).
- Develop a codebook, with examples, that allows qualitative researchers to hand-label these messages
- Identify simple linguistic patterns, hashtags, and other features that will aid in the search for harassing messages
- Map codebooks from existing projects to our group codebook
- Create hand-coded data as a first step toward building a large community corpus of harassing content.
- Create and maintain a website which will serve as a nexus for work on developing guidelines and corpora.

We anticipate this workshop will be the inauguration of a virtual research community where volunteers can add content, reviewed and approved by other members, to an ever-growing research corpus.

Tentative Schedule

Hour 1: Overview of purpose, introductions

Hour 2: Presentation of position papers, description of available datasets

Hour 3: Codebook development - We plan to discuss ahead of time which labels to include. These may be hate speech, harassment, threats, etc. Development of a codebook will focus on creating working guidance on how to apply these pre-selected terms.

Hour 4: Lunch and informal discussion

Hour 5: (until end of day): coding messages from supplied datasets

Organizers

Jennifer Golbeck (main contact)- Associate Professor, University of Maryland College of Information Studies. Computer Scientist. Dr. Golbeck specializes in the study of social media from a computational perspective. She uses a combination of techniques to model users and online behavior, including machine learning, computational linguistics, statistical learning, as well as qualitative analysis. Her Trollbusting group has a 50,000 tweet hand-labeled corpus of harassing comments in development that will be linked in to this project as an initial contribution.

Joel Tetreault - Director of Research, Grammarly. Computer Scientist. Dr. Tetreault is a scientist in computational linguistics, focusing on applying natural language and machine learning techniques to education and social analysis. In his prior role as a Senior Scientist at Yahoo Labs, he developed an algorithm (which has gone live) to automatically detect abusive language in online user comments and is releasing an annotated dataset of abusive comments. He is also organizing a workshop on computational approaches to detecting abusive language.

Allison Druin was named Special Advisor for National Digital Strategy last year by the National Park Service. By taking a leave of absence from the University of Maryland for two years, Dr. Druin focuses on how to better leverage digital tools to excite the next generation of park visitors, to change how the national parks share their stories, and to better preserve our cultural and natural resources. Her work at the National Park Service and previously at the University of Maryland has also included developing co-mentoring programs to support the advancement of women in academia and now in the national parks. Her co-mentoring programs have come out of her HCI methods of co-design in giving all people a voice in change. She develops methods for facilitated discussion, design, and consensus-building.

Website

URL: http://social.umd.edu/woh//

The workshop website will begin as a central place for workshop details. After accepting papers, it will contain a repository for datasets, codebooks, and other resources for participants to share ahead of CHI 2017. After the workshop, we plan to convert it to the working home for our community and a central place for researchers to access and contribute to our corpus.

Pre-Workshop Plans

The proposers are all currently part of research networks and social media networks interested in online harassment. We plan to recruit participants through these and other CHI-related channels.

Participants can submit position papers. Because the focus of this workshop will be to do actual *work* during the one day meeting, we will share these selected papers along

with codebooks and datasets with participants ahead of time. We plan to establish a listserv for accepted participants (and any others who want to participate) to initiate some discussions ahead of the actual meeting. This will ensure when we all come together in Denver, we will have a group foundation that we can immediately build on.

Within that community, we plan to discuss which labels to focus on in our working codebook. There are many options - hate speech, harassment, threats, misogyny, racism, broadly offensive content, violence, extreme language, etc. Discussion ahead of time will let us decide which labels are of the greatest interest to the participants and allow us to have a more focused and efficient discussion of the codebook in the workshop itself.

Post-Workshop Plans

We plan to release the following resources to the research community:

- A codebook with different types of harassment and guidelines for coding them
- A labeled dataset of online content. This will contain content labeled in the workshop with our codebooks
- A portal with links to related datasets available for use by harassment researchers

We also anticipate a journal article detailing the dataset and codebooks, authored by active workshop participants, to be published after the workshop.

Ideally, this workshop will launch a collaborative community that can continue this work together remotely and in subsequent workshops.

Call for Participation

Format and goals of the workshop
The goals of the workshop are to:

- Develop a shared vocabulary of harassment types (hate speech, individual harassment, threats, deeply offensive language, etc.).
- Develop a codebook, with examples, that allows qualitative researchers to hand-label these messages
- Identify simple linguistic patterns, hashtags, and other features that will aid in the search for harassing messages
- Map codebooks from existing projects to our group codebook
- Actually create some hand-coded data as a first step toward building a large community corpus of harassing content

This will be a true *working* workshop. We will spend the morning reviewing available data and developing a typology of online harassment, and the afternoon actually coding data and refining a codebook. The dataset, codebook, and mappings to existing typologies will be published at the end of the workshop.

Participation
Anyone interested in the topic can participate in the workshop. If you attend, you must commit to working on developing the corpus in the session.

We will be accepting short papers (up to 4 pages) from researchers who have resources they can contribute to this effort. This includes:

- Available social media datasets that could be part of this repository
- Code books, coding schemes, taxonomies, and typologies of online harassment, generally or of a particular sub-type
- Other resources that may be useful to building a central repository of online harassment for research

Papers should describe the resources, how they were developed or collected, and where they are available. Submissions should be in standard ACM format and submitted through Easy Chair.

Deadlines: paper submissions by 21 December 2016

Note: at least one author of each accepted position paper must attend the workshop. All participants must register for both the workshop and for at least one day of the conference.

Website: http://social.umd.edu/woh

References

[1] Sofia Berne, Ann Frisén, Anja Schultze-Krumbholz, Herbert Scheithauer, Karin Naruskov, P Luik, C Katzer, Rasa Erentaite, and Rita Zukauskiene. 2013. Cyberbullying assessment instruments: A systematic review. *Aggression and violent behavior* 18, 2 (2013), 320–334.

[2] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7, 2 (2015), 223–242.

[3] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing" trolling" in a feminist forum. *The Information Society* 18, 5 (2002), 371–384.

[4] Laura Leets. 2002. Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of social issues* 58, 2 (2002), 341–361.

[5] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of NAACL-HLT*. 88–93.