Online Abuse and Human Rights: WOAH Satellite Session at RightsCon 2020

Vinodkumar Prabhakaran
Google Brain

Zeerak Waseem
University of Sheffield

Seyi Akiwowo
Glitch!

Bertie Vidgen
The Alan Turing Institute

Abstract

In 2020 The Workshop on Online Abuse and Harms (WOAH) held a satellite panel at RightsCons 2020, an international human rights conference. Our aim was to bridge the gap between human rights scholarship and Natural Language Processing (NLP) research communities in tackling online abuse. We report on the discussions that took place, and present an analysis of four key issues which emerged: Problems in tackling online abuse, Solutions, Meta concerns and the Ecosystem of content moderation and research. We argue there is a pressing need for NLP research communities to engage with human rights perspectives, and identify four key ways in which NLP research into online abuse could immediately be enhanced to create better and more ethical solutions.

1 Introduction

The Workshop on Online Abuse and Harms (WOAH[1]), previously known as the Abusive Language Workshop (ALW), is a leading publication venue for cutting-edge computational research on detecting, analysing and tackling online abuse, primarily using Natural Language Processing (NLP) techniques. Over the past three iterations, WOAH has built an interdisciplinary community of computer scientists, social scientists, critical theorists, legal scholars and more.

Continuing this tradition of embracing interdisciplinarity, for WOAH 2020 we organized a satellite panel at RightsCon 2020, one of the leading international human rights conferences. It brings together business leaders, technologists, academics, journalists and government representatives to discuss pressing issues at the intersection of human rights and technology. [2]

As an area directly concerned with protecting under-represented, vulnerable and marginalised communities, we anticipated that NLP research into online abuse would benefit from engaging directly with human rights scholarship. Human rights offers a powerful way of motivating work in this area, understanding what is at stake with online hate (and efforts to counter it) and bridging the gap between engineering/computational work and the groups that are affected by abuse. A global human rights framework could provide a much needed value system for work on tackling online abuse, furnishing NLP researchers with much-needed frameworks, theories and concepts. Yet, to date, research on online abuse published at WOAH/ALW and other leading NLP venues has largely lacked an explicit connection to human rights or an engagement with human rights scholarship.

Similarly, we believe the human rights scholarship and activism could also benefit from a shared understanding of the problem space which technologists working on NLP-based solutions for tackling online abuse operate in – exploring both the opportunities and limitations of such approaches. More collaboration and dialogue could also foster more useful and critical discussions of how engineering solutions are designed and implemented, helping to ensure that NLP tools have a positive impact on society.

In the satellite session at RightsCon, we brought together human rights experts with computer scientists in an effort to formulate a rights respecting approach to tackling online abuse.[3] Our objective was to bridge the gap between these communities as a way to drive new initiatives and outlooks, ultimately leading to better and more responsible ways of moderating online content. The intended outcomes of the panel were:

[1] www.workshoponline abuse.com

[2] https://www.rightscon.org

[3] Please note that the views contained in this report do not necessarily represent the views of all panellists.

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 1–6
Online, November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

1. To start a dialogue between computer scientists (primarily NLP experts) and the human rights community working on tackling online abuse.

2. To establish a shared understanding of challenges and possible solutions.

3. To invigorate a global human rights based critique of computer science research practices in this domain.

The purpose this report is to summarize the panel and reflect on the outcomes. A one-hour discussion session has been allocated during the WOAH 2020 program to reflect on the report and the RightsCon panel.

2 RightsCon Session Overview

Title Tackling Online Abuse: Bridging the Gap between Human Rights and Computer Science

Organizers Vinodkumar Prabhakaran, Zeerak Waseem, Bertie Vidgen, Seyi Akiwowo (moderator)

Moderator Seyi Akiwowo (SA)

Panelists Maria Y Rodriguez (MR), Caroline Sinders (CS), Cristian Danescu-Niculescu-Mizil (CDNM)

Participants The session had 113 RightsCon attendees (including organizers and panelists).

Session structure Initially, three questions were posed to the panelists (shared in advance and decided by the Organizers). Each panelist was requested to answer at least one of the three questions in their 10-15 minute opening statement:

1. What are the primary challenges in tackling online abuse?

2. What are the blind spots of algorithmic means of tackling online abuse?

3. What are the barriers that divide computational research and human rights scholarship?

We then opened the floor for questions from attendees. They were curated and given to the moderator in real time. This lasted for 15 minutes, after which the panel concluded.

3 Analysis and Synthesis

We transcribed the panel using automated software and then manually checked the manuscript. We used an inductive approach of qualitative text analysis to analyze the transcript in order to identify recurring topics (Thomas, 2006). We found over a couple of dozen recurring topics that were brought up by the panellists at least twice during the discussion. These recurring topics were then analysed and iteratively refined and merged together to identify four high-level themes: Problems, Solutions, Meta considerations, and Ecosystem.

3.1 Problems in tackling online abuse

Abuse is contextual Several of the participants brought up the contextual nature of online abuse and hate, and how this makes it more challenging to detect and intervene. As CS put it: ‘what sounds like harassment to one person may not be harassment at all. Harassment is very contextual. Sometimes it’s not, but for the most part it is extremely contextual. Actions are contextual... and how do you code context?’. Panellists also brought up how geographic and temporal contexts are important in understanding whether or not something is abusive. CS pointed out that ‘who is speaking and the power that they have dramatically changes what they say and how it is interpreted [...] the context of who says what should never be forgotten’. She argued that the language used by Donald Trump has different implications based on whether he uttered them as the President or as a presidential candidate. CS also brought forth the example of Pepe the frog: ‘Pepe started off being a stoner comic that was really beloved in California and LA and turned into a mean representative of the alt-right and now is being used in a proactive and protest-supporting way in Hong Kong. Language differs and changes radically and really quickly. It also differs from country to country and place to place.’

Abuse has different modes CS discussed the different ways in which online abuse happens. These can each inflict different harms, exhibit different dynamics and may require different strategies to mitigate them:

- Whether the person being harassed knows the harasser or not;

- Whether the harassment is one to one, many to one, or many to many;

- Whether the harassment is coordinated or uncoordinated;
- Whether the harassment is happening across many platforms or just one;
- Whether the harassment is also happening offline

Online and offline harms are closely connected Online harms can be associated with harm in the offline world. MR used the example of Kanye West's announcement to run for US President as an example of online abuse and harassment which could result in offline harms (in this case, how it may influence African American voters to not vote as a result of the rhetoric used in the campaign). SA summarized this as a "continuum of violence" and argued it is important to understand that what happens offline will likely be displayed online. Equally, online discussions can seep into and shape the offline world. Panellists also pointed out that not all forms of abuse that happen online are immediately obvious and that it requires understanding of, and reflection on, their offline impact. CS pointed out that 'If someone is doxing you they may not say "I am going to dox you."'

3.2 Solutions for tackling online abuse

A range of interventions are available Panellists discussed the various modes of intervention to tackle online abuse that are currently available. They also discussed where NLP may help, noting that any solution needs to be scalable to match the volume and variety of content shared online. Much NLP research is focused on finding and classifying offensive or toxic language, which is then either directly censored or flagged for human moderators to review. However, CDNM argued that this is already too late in a sense since the abuse (and associated harm) may have already been inflicted. CDNM described their work on detecting conversations that may turn toxic ahead of time as a potential alternative. One of the attendees also brought up the question of counter narratives as a means of addressing the harms inflicted by online abuse. CDNM pointed out that while banning users is a popular approach it may not always be appropriate since some offenders are 'regular people' who happened to misbehave only once or a few times. Banning is an important option for repeat and serious offenders but may be overly censorious in some contexts.

Tech solutions are often flawed Myriad issues are associated with various tech-based interventions. Both CS and CDNM pointed out the moral and ethical issues associated with bot-driven counter narrative strategies, as well as the challenges of deploying such interventions at scale. All panellists also pointed to the various biases that may be encoded into the automated detection systems. For instance, MR recalled Brandon Stewart's and Justin Grimmer's framing that 'mathematical models of language are wrong. They just are. There is no way to be so reductionist as to capture the complexity of human interaction in, you know, binary. However, they're useful; they can give us some really good information to develop.' (Grimmer and Stewart, 2013) CS and MR also brought up the issue that most NLP models being based on language data that is not representative of the communities they are deployed onto.

Content moderation should be a question for society Platforms which choose to not moderate online abuse, or do so only very minimally, are not just being 'light touch' – they are making a decision that reflects a set of values and norms. The effects of this can be highly pernicious. CDNM brought up that not moderating popular online spaces such as Facebook and Twitter may result in some groups losing their social voice if they do not feel safe to communicate. MR brought up the historical example of the printing press and how it resulted in the rise of cults and certain political movements, in addition to increases in positive things like literacy and civic engagement. She argued that if we consider social media to be the new public commons then moderation is not only a question of infrastructure and technical feasibility – the key issue is what civil society agrees upon for it. Debating, arguing and contesting our expectations for these public spaces is key to figuring out what sort of moderation is needed.

The past is not a good way of understanding the future MR argued that many researchers implicitly assume that historical data can be used to train new models and then applied to future data. For instance, in content moderation many systems are trained on old annotated datasets, which are often years out of date. This approach works well in some fields like demography (which studies population dynamics and migration patterns), but is often inappropriate when researching how individuals communicate online and interact within groups as: 'we can in no way assume that [...] the future will look like the past because the present is actually where all of this is occurring.'

Tech solutions raise surveillance/privacy concerns and tradeoffs Tech based interventions that review content on a continuous basis (so called 'proactive monitoring') raise concerns about privacy as they involve constant evaluation and classification of users' messages. CDNM raised the issue that such technology can be put to dual uses. MR identified that interventions of all kinds require surveillance when they are implemented and scaled. Therefore there will always be a need to balance intervening with minimizing losses to privacy.

3.3 Meta considerations

Who does the research? Panellists brought forth the capability gap within the community of researchers and developers working on tech-based solutions for online abuse – where there is a lack of engagement with groups actually affected by online abuse. MR pointed out the importance of lived experiences for understanding the harms that online abuse causes to communities. MR urged researchers to ensure that communities affected by abuse are represented within their teams. This would help to ensure the nuances of the problems are better understood and accounted for in interventions. It could also avoid any unethical or inappropriate uses of technology.

NLP is not always the solution Panellists repeatedly pointed out the over-reliance on AI/ML/NLP above potentially more efficient and simple solutions. MR highlighted that the main focus of many computational researchers is to improve model performance, instead of understanding how abuse occurs, why, and the nuances associated with it. CS pointed to non-tech solutions such as intervening in a conversation to engage with the offender, and CDNM raised the example of Wikipedia discussion forums, where a fixed wording message was shown to experienced editors when engaging with newcomers. It had positively affected them to help newcomers to fit in (Halfaker et al., 2011). A socio-technical approach to tackling abuse could help address these issues.

Vocabularies needs to be bridged Panellists agreed that there is a need for more effective interdisciplinary conversations, such as this panel and the associated report. At present there is a lack of shared understanding about the problem posed by online abuse, and a common vocabulary to talk about issues does not yet exist. More discussion and collaboration requires some shared terminology and willingness to understand cross-disciplinary scholarship.

3.4 The Ecosystem

More transparency about moderation processes is needed Participants discussed how there is a lack of transparency regarding the processes and pipelines that platforms follow in handling online abuse. This is a core challenge which limits how much civil society and independent researchers can meaningfully engage with their work. CS pointed out that there is not enough clarity in how online abuse (including hate speech and harassment) is handled by different platforms compared with misinformation/disinformation, or what mechanisms platforms use to adjudicate reports from users. More information from platforms would be useful for civil society, such as knowing the aggregate number of user reports, what responses were taken, what information was available to the content moderators, how much time was given to each one, and so on. SA also pointed to the issue that user-reporting often ends up being free labor for platforms to improve their models but they are often not transparent about it. Often, user reports are provided by activist groups who are in effect giving free labour to highly profitable companies.

More education is needed Panellists identified the need for better education of users, civil society and governments about the harms caused by online abuse. SA argued that there is a huge deficit in how people understand digital security, safety, self defense, and agency online. CDNM also pointed to the need to educate the public about what can and cannot be done technically to ensure user's safety. Efforts should be made to be more open about the flawed nature of tech-based interventions for online abuse.

Platforms should be more accountable Panellists identified the deficit in accountability when it comes to how tech platforms deal with online abuse. SA drew parallels with how civic bodies hold extensive case reviews when severe incidents like homicide happen so as to understand what happened, and to identify any failures. She questioned why such case reviews do not happen when grave injustices occur on tech platforms. SA argued there is a need, and growing opportunity, for academics, policy makers and civil society to work together to bring tech accountability into the mainstream.

4 Discussion

Our RightsCons session is just the starting point for a wider set of conversations which need to take place between technologists, engineers and civil society. Numerous shortcomings in how platforms moderate content and NLP research is conducted have been identified throughout the panel. There is a pressing need for them to be addressed if we are to build systems which meaningfully tackle online abuse. A human rights based approach is a promising way of bridging the historical and disciplinary divides between these groups but it may not be the only way. We strongly encourage that, aside from anything else, a wider range of perspectives are adopted.

From our analysis we identify four main areas which could immediately be tackled by the NLP community to create better and more ethical solutions for online abuse.

1. **Problems**: the NLP community tends to work on datasets which do not account for context, often focusing on just language by itself. In particular, less attention is paid to the particular social and political setting in which abuse is sent, and the role of the speaker and audience. The dynamics of how abuse spreads, such as the role of networks and 'trigger events' are often not considered in NLP work. Particularly important focuses for future work include (1) how the abuse is sent and whether it is one to one or many to one, (2) whether it is coordinated or not and (3) the connections between online and offline abuse.

2. **Solutions**: the NLP community primarily focuses on "find offensive or toxic language", with much less focus on other strategies such as counter narratives or preemptive interventions. Questions around moral, ethical and bias issues are relatively new to the field, although progress has been made. For instance, the theme of WOAH 2020 is social bias and unfairness. The NLP community rarely talks about the potential misuse of the systems we develop, such as invading users' privacy through surveillance. It also gives insufficient focus to whether systems can actually be used in the real-world, focusing too much on maximizing performance through more sophisticated engineering solutions. Efforts to integrate these concerns directly into NLP research, or at least demonstrating awareness they exist, would be a substantial help.

3. **Meta considerations**: NLP is not the only answer. There is a concerning tendency to rely too heavily on tech to 'solve' the problem of online abuse, even though it is a field that itself has numerous representation problems. The field should aim to improve representation and to work more closely with the groups affected by online abuse, as well as other end-users.

4. **Ecosystem**: Many NLP solutions are hard to explain and their limitations are not well understood by the general public. NLP researchers could better advance efforts to improve transparency, accountability, and literacy. Areas include: (1) working on more interpretable models, (2) unearthing the implicit assumptions being made in NLP work flows and (3) explaining their work in non-technologist language to help it reach a wider audience.

Tackling online abuse is an important task; one that is too important to let disciplinary divides hold back. NLP and other computational approaches have the potential to make an incredibly positive impact on this problem – and are already being used in many settings. But NLP researchers need to move beyond seeing online abuse detection as solely an engineering problem, instead recognizing the social and ethical impetuses that motivate it. Human rights frameworks offer a powerful starting point for achieving this. Ultimately, it will require far more meaningful dialogue between the engineers and technologists who have largely been responsible for building systems and the groups who are directly affected by them. This report is intended as one step towards achieving this goal.

Acknowledgments

The authors would like to thank all participants and attendees at the panel for their contributions and feedback.

References

Justin Grimmer and Brandon M Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 1(1):1–31.

Aaron Halfaker, Bryan Song, D. Alex Stuart, Aniket Kittur, and John Riedl. 2011. Nice: Social translucence through ui intervention. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, page 101–104, New York, NY, USA. Association for Computing Machinery.

David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2):237–246.