## Conversational Agents to Address Abusive Online Behaviors

Beauxis-Aussalet, Emma

**Publication date**
2019
**Document Version**
Final published version

# Conversational Agents to Address Abusive Online Behaviors

**Emma Beauxis-Aussalet** [1]

## Abstract

Abusive online behaviors occur at a large scale on all social media, and have dire consequences for their victims. Although the problem is largely acknowledged, technological solutions remain limited to detecting and hiding abusive comments. More can be done to address abusers themselves. We propose to investigate the potential of conversational technologies to dialogue with abusers. In this problem description paper, we outline directions for studying the effectiveness dialogue strategies, e.g., to educate or deter abusers, or keep them busy with chatbots thus limiting the time they spend perpetuating abuses.

## 1. Online Abusive Behaviors

A variety of abusive online behaviors occur at a large scale, every day, on all social media. Among others, the scope of abuses include racism, sexism, bullying, anti-LGBT+, antisemitism, islamophobia, or body shaming. The diversity of abusive behaviors makes it difficult to establish a definition of online abuses (Tokunaga, 2010; Foody et al., 2015). We can retain this definition of cyberbullying: *"any behavior performed through electronic or digital media by individuals or groups that repeatedly communicates hostile or aggressive messages intended to inflict harm or discomfort on others"* (Tokunaga, 2010).

The motivations underlying online abuses are also diverse. Several theories can be used to investigate abusers' psychology. Among others, Mishna (2012, Chapter 3) considers *social learning theory* (e.g., by witnessing and mimicking abusive peers, abusers acquire mental models where aggression yield rewards and social status), *coercion theory* (e.g., if repeating and escalating pressure on peers is successful, abuses are deemed rewarding), *cognitive behavioral theory* (e.g., assumptions and interpretations associated with life events are the primary determinants of behaviors, and can hinder or foster abuses), and *attribution theory* (e.g., abusers perception of their peers' hostility may be exaggerated, a phenomenom called hostile attribution biases, which may leading abusers to respond aggressively).

[1]Digital Society School, Amsterdam University of Applied Science, The Netherlands. Correspondence to: <e.m.a.l.beauxis@hva.nl>.

*Appearing at the International Conference on Machine Learning AI for Social Good Workshop*, Long Beach, United States, 2019.

## 2. The Need for Addressing Online Abuses

Online abuses have dire short-term and long-term consequences for their victims, especially teenagers. For instance, online abuses were identified by the UK government as one of the main online threats to address: *"Online platforms can be a tool for abuse and bullying [...]. The impact of harmful content and activity can be particularly damaging for children, and there are growing concerns about the potential impact on their mental health and wellbeing"* (Javid & Wright, 2019). Consequences of online abuses include depression, anxiety, suicide, emotional distress, anger, panic symptoms, lower self-esteem, school truancy, or substance use (Tokunaga, 2010; Sinclair et al., 2012).

Hence the need to address online abusive behaviors is a matter of public health, as well as basic human rights (Mishna, 2012). For instance, the Universal Declaration of Human Rights states that *"Everyone [...] is entitled to realization [...] of the [...] social and cultural rights indispensable for his dignity"*. The United Nations Convention on the Rights of Child states that *"State Parties shall take all appropriate measures to protect the child from all forms of physical or mental violence"*. Online abuses are a form of mental violence that threatens the victims' dignity.

Yet *"Western society has customarily tolerated bullying behaviors"* (Mishna, 2012). Failing to address abusive online behaviors stages abuses has socially acceptable behaviors. Victims may not identify the meaning, impact and immoral nature of the abuses. As bullying may seem normal, in the context of online bullying among youth, for instance, victims are commonly being abusers at time too (Mishna, 2012) although such ambivalent bully-victim behaviors remain rare in the context of offline bullying. As Mishna (2012, Chapter 4) highlights, **"inaction is not simply a lost opportunity but also represents a stance and may lead to more harm"**. Hence it is crucial to curb the culture of impunity associated with online abuses.

## 3. The Potential of Conversational Agents

Methods to address online abusers are limited as most approaches focus on providing support to victims. However, addressing abusers themselves has potential for positive impacts on all parties: abusers (whose behaviors is a symptom of psychological issues), victims (whose risks of abuse may diminish as abusers behaviors are tackled), and society at large (to prevent abuses from becoming social norms).

Online abusers can be identified by analysing their social media messages. A variety of classifiers are available for detecting abusive messages (Rosa et al., 2019) as this field of research has been rapidly developing (e.g., workshops on Abusive Language Online are held at ACL conferences since 2017). Once abuses are detected, responses are usually limited to blocking abusers, deleting abusive messages, and possibly displaying a single message to address the abusers.

The use of conversational agents (chatbots) remains unexplored. However, chatbots have the potential to create constructive dialogues with abusers, without requiring victims or human mediators to be exposed to further aggression. For instance, we envision three opportunities:

- **Educate abusers**: chatbots may advocate for a more inclusive society, encourage abusers to develop healthier social skills, and encourage abusers to address the underlying conditions or psychological issues that lead them to perpetuate abuses.
- **Deter abusers**: chatbots may inform abusers of the legal consequences of their actions. Although legal proceedings are impractical and rarely successful (Mishna, 2012), this approach may still contribute to curbing the culture of impunity.
- **Keep abusers busy**: chatbots can aim at making conversations with abusers as long as possible. The time spent interacting with chatbots is time not spent perpetuating abuses (and time spent being exposed to educating or deterring argument).

Significant future research is required to establish efficient methods for using conversational agents to address online abusers. Such research must involve psychologists and social scientists, as well as computer scientists and experts in artificial intelligence and natural language processing. Multidisciplinary teams are required to balance the socio-technical risks. From a sociological perspective, conversational agents may have adverse effects and fuel online abuses (e.g., abusers may perpetuate more of worst abuses for the sake of triggering the chatbots, or remaining undetected). From a technical perspective, chatbots must adapt to the language of abusers to behave as seemingly real interlocutors. Chatbots may thus learn and reproduce abusive behaviors, becomign automated bullies as did Microsoft's Tay (Wolf et al., 2017).

Designing conversational agents to address abusive online behaviors is challenging and risky. However, inaction may be more risky (Section 2) and exploring the potential of conversational agents may be deemed a moral obligation.

Exploring the potential for educating, deterring or keeping abusers busy may not be succeed at first, or even after several design iterations. Yet, even is unsuccessful, the benefits of trying are two-fold: (i) communicate that abuse and impunity are not becoming our social norm, (ii) provide the first framework for conducting scientific research on conversational agents to address online abuses.

## 4. Architecture & Experimental Framework

Our first experiments will target a mainstream social network that authorises chatbots (e.g., Twitter). Three software components are necessary to implement an experimental framework for studying conversational agent to address abusive online behaviors:

- **Abuse Detector**: The automatic detection of abusive messages can be performed using a variety of classifiers (Rosa et al., 2019). As the research community is actively developing such classifiers, we will reuse the existing software without developing new techniques or datasets.
  The detection of abusive messages can be performed by either (i) targeting specific keywords (hashtags) to query social media using public API, or (ii) recruit social media users willing to experiment with chatbots, and enable the detection of abuses in all the messages they receive.
  Preferably, the Abuse Detector should be able to detect different kinds of abuses, to adapt dialogues accordingly. It should also provide a score describing the level of abuse (e.g., severe or mild aggression). Such score can serve to adapt dialogues to the level of abuse, to monitor the success of the dialogues, and to monitor the levels of abuse within communities before and after interactions with chatbots (e.g., to assess the long-term impacts of chatbots).
- **Conversational Agent**: Once abuses are detected, chatbots will lead conversation with abusers. Different levels of dialogue complexity can be envisioned, depending on how chatbots may (i) adapt to abusive language and topics of abuse to personalise the dialogues, (ii) deliver complex argumentation, e.g., using Toulmin model (Kneupper, 1978). To enable such dialogues, we may use commercial, open-source or academic software, or develop our own component.

At first, we will focus on the basic content of the argumentation within the dialogues, rather than on personalising the dialogues. Thus we will use existing software components, and implement predefined dialogue structures and messages with little to no personalisation.

- **Abuse Monitor**: To evaluate the success of the dialogues, we must monitor the frequency, type and level of abuse before, during and after abusers interact with chatbots. The Abuse Detector component can measure the frequency of abuse (i.e., by providing the numbers abusive and non-abusive message). However, classifiers are imperfect, thus we may apply error estimation and bias correction methods (Beauxis-Aussalet & Hardman, 2017). Monitoring the type and level of abuse would require specific classifiers (i.e., multiclass and probabilistic classifiers). If unavailable, we will rely solely on monitoring the frequency of abuses.
  The Abuse Monitor component will also serve to control for potential adverse effect of chatbots. If abuses are worsened by the chatbots, experiments must be discontinued.

Provided with this architecture, experiments can be conducted to test the efficiency of alternative approaches for designing dialogues (e.g., the 3 approaches in Section 2), and alternative structures and contents of dialogues for each approach. The results of the dialogues can be analysed quantitatively (i.e., by measuring the frequency and intensity of abuses) to monitor a large range of human interactions on social media. However, qualitative analysis must complement the quantitative results to ensure that the quantification of abuses is representative of actual abusive behaviors. For instance, the Abuse Detector may indicate a certain level abuse that psychologists would consider otherwise.

To compare quantitative and qualitative assessments, we will randomly sample messages detected as abusive or not, and ask human experts to classify them and grade their level manually. Comparing the manual and automated classifications can serve to validate the quantitative measurement, and obtain larger test sets for applying bias correction methods (Beauxis-Aussalet & Hardman, 2017) or deriving refined estimate of the level of abuse as suggested by Beauxis-Aussalet (2019, Section 5.7.2).

## 5. Discussion

We argue that well-designed conversational agents may lead abusers to completely or partially refrain from conducting further abuses. Such positive impact might not be achievable with all abusers, especially for the most radicalised. Yet chatbots may achieve non-negligible impact on abusers, as well as victims and society. We also argue that experimenting with conversational agents will enable novel studies of the strategies for addressing abusers themselves. Such line of work remains overlooked, as most approaches focus prevention methods, therapies for the victims, or simple deletions of abusive messages.

However promising conversational technologies might be, we acknowledge crucial socio-technical challenges and risks. Introducing chatbots may generate new kinds of abuse, or intensify them. Triggering chatbots, or avoiding to trigger chatbots, may add to abusers motivations. Such adverse effects must be monitored, and if they surpass positive impacts, chatbots must be disabled.

The variety of underlying motives for abuse makes the design of efficient automated dialogues difficult to establish. The design space is extremely large as many alternative dialogues can be envisioned, e.g., to target generic or specific types of abuse and abusers. Hence, further scientific research must be conducted to establish efficient conversational methods, and developing a first framework for studying such conversational methods is necessary.

A key challenge for establishing successful dialogues with abusers is to retain their attention, as chatbots can be easily dismissed. Strategies to retain abusers attention may involve provocation or trickery. Provoking abusers by challenging their dominating attitude (e.g., by challenging their lack of social intelligence) may incite them to engage with chatbots. It may also also point at the underlying conditions that prompt them to adopt abusive behaviors, and encourage them to develop more balance and mature mindsets. Tricking abusers to believe that chatbots are real persons, with a profile that abusers typically choose to victimise (e.g., chatbots modeled after female personas or other vulnerable social groups), may also incite abusers to engage with chatbots.

Finally, developing conversational technologies to be deployed at a large scale on social media yield ethical concerns beyond the topic of online abuses. Such technologies can be used to manipulate public opinion, (e.g., for political or commercial purposes) and abusers may develop their own chatbots to perpetuate abuses automatically, as an arms race. Such risks may be addressed with regulations that would authorise only trusted partners to massively deploy chatbots on social media. To enforce such regulations, technologies to detect and block unauthorised chatbots would be necessary. Regulating chatbots in online spaces carries complex socio-ethical concerns, but these are beyond the scope of our this paper.

## Acknowledgement

We are grateful to Caroline Sinders and Abdo Hassan for their interest in this project, and for their willingness to contribute to this endeavour with their skills and expertise.

## References


Beauxis-Aussalet, E. *Statistics and Visualizations for Assessing Class Size Uncertainty*. PhD thesis, University Utrecht, 2019.

Beauxis-Aussalet, E. and Hardman, L. Extended methods to handle classification biases. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 765–774. IEEE, 2017.

Foody, M., Samara, M., and Carlbring, P. A review of cyberbullying and suggestions for online psychological therapy. *Internet Interventions*, 2(3):235–242, 2015.

Javid, S. and Wright, J. Online harms white paper. UK Government Department for Digital, Culture, Media & Sport, Home Office, 2019. URL https://www.gov.uk/government/consultations/online-harms-white-paper.

Kneupper, C. W. Teaching argument: An introduction to the Toulmin model. *College Composition and Communication*, 29(3):237–241, 1978.

Mishna, F. *Bullying: A guide to research, intervention, and prevention*. Oxford University Press, 2012.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P., Carvalho, J., Oliveira, S., Coheur, L., Paulino, P., Simão, A. V., and Trancoso, I. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93: 333–345, 2019.

Sinclair, K. O., Bauman, S., Poteat, V. P., Koenig, B., and Russell, S. T. Cyber and bias-based harassment: Associations with academic, substance use, and mental health problems. *Journal of Adolescent Health*, 50(5):521–523, 2012.

Tokunaga, R. S. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior*, 26(3):277–287, 2010.

Wolf, M. J., Miller, K., and Grodzinsky, F. S. Why we should have seen that coming: comments on microsoft's tay experiment, and wider implications. *ACM SIGCAS Computers and Society*, 47(3):54–64, 2017.