Investigating the impact of interventionist causal approach on the study of verbal aggression and discrimination online

Patrycja Tempska and Rafal Urbaniak

Contents

1 Section Abstract

This article describes an experimental intervention study conducted in a naturalistic, digital setting (Q&A forum on Reddit), utilizing a collective intelligence approach to content moderation and reduction of the level of verbal aggression among a selected group of Reddit users who regularly attack other community members. Collective Intelligence in this sense means exploring the collaboration between human and machine intelligence to develop solutions to social challenges. Artificial Intelligence was used to detect verbal aggression (personal attacks) and notify human volunteers about attacks. Volunteers after receiving notifications employed interventions based on norm or empathy promotion. We find that only those who were sanctioned with norms-inducing interventions had their personal attacks' level significantly decreased.

2 Introduction

Although much effort has been made in order to tackle the problem of verbal aggression and harassment online, looking at various reports and surveys (Laub, 2019; Sorrentino, Baldry, Farrington, & Blaya, 2019; Vogels, 2021), it remains a common hindrance for people engaging with social media in their everyday lives. The problem got exacerbated in the midst of the COVID19 pandemic, during which the majority of our social life moved to cyberspace. During this shift, there was an increase in cyberbullying attitudes and perpetration (Barlett, Simmers, Roth, & Gentile, 2021), 90% increase in public reports of illegal online content (Grant, 2021), including 114% increase in non-consensual sharing of intimate images, 30% increase in cyberbullying, as well as 40% of increase in adults reporting online harassment. According to a report conducted by company L1ght (L1ght, 2020), hate speech

directed towards China and the Chinese went up by 900% on Twitter. Gaming platforms were in the spotlight as well, with a 40% increase in toxicity on Discord [1].

But alongside the growing need for even more efficient and proactive moderation, the capacity to execute it did not go hand in hand, forcing companies and policymakers to rethink the current model of moderation processes and workforce. Due to the COVID19 restrictions including social distancing, a lot of those serving the role of moderators had to be sent home (Bhattacharya, 2021) without the ability to work remotely because of the constraints affiliated with restrictive non-disclosure agreements (NDA) among others. Curtailing the moderators' workforce was accompanied by more agency given to algorithms and AI-based moderation. Those changes, as argued by Gerrard (2020), can be seen as a serious red flag in terms of safety for all users on online platforms.

3 Automated versus Human-based Moderation

The hindrances and threats that go along with the Artificial Intelligence-based methods for moderation have been widely debated, with the most critical discussions revolving around technology performance (MacAvaney et al., 2019; Schmidt & Wiegand, 2017). State-of-the-art solutions are mostly governed by statistical methods including deep learning and machine learning (Jordan & Mitchell, 2015; LeCun, Bengio, & Hinton, 2015; Sejnowski, 2020). Their performance is inherently tied to the amount of data being fed to the system and the quality of its annotation. At different stages of the process, from data gathering and preparation, annotation to the training or algorithms themselves, biases seem to be omnipresent Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2021).

This deserves a developed paragraph with examples

Users' of online services are also creative in their strategies to circumvent automated content moderation systems and as shown by Gröndahl, Pajola, Juuti, Conti, & Asokan (2018), current techniques are vulnerable to the most common evasion attacks like word changes (inseting typos and leetspeak), word-boundary changes (inserting or removing whitespace), or word appending (appending common or non-hateful words like "love").

This deserves a developed paragraph with examples

Lack of generalisability of the models—the ability to perform well on datasets coming from sources other than the one used for training—is a serious shortcoming as well (Rosa et al., 2019; Swamy, Jamatia, & Gambäck, 2019; Yin & Zubiaga, 2021).

This deserves a developed paragraph with examples

As shown by Wu, Ribeiro, Heer, & Weld (2019), Lipton & Steinhardt (2019), and Musgrave, Belongie, & Lim (2020) in practice, the development of such models often lacks thorough error analysis and legitimate experimental methodology, which can result in non-reproducibility. This is also connected with a potential lack of thorough understanding of the limitations of the models and spurious conclusions being announced to a wider public. Specifically, Lipton & Steinhardt (2019) distinguishes four dysfunctional patterns occurring in the current research paradigm in the industry and academia alike.

1. First, the inability to draw a clear distinction between speculation and explanation, with the first one often being disguised as the second. For instance, in a paper from 2017 [@steinhardt2017certified], Jacob Steinhardt, the author of @lipton2019troubling, admitted to stating that "the high dimensionality and abundance of irrelevant features... give the attacker more room to construct attacks" - although no experiments were conducted to measure what is the effect of dimensionality of the neural network on its attackability.
2. Second, the inability of successful identification of the sources of performance improvement (whether it was problem formulation, optimization of the heuristics, data-preprocessing, hyperparameter tuning, or perhaps yet another aspect). As was shown by Gabor Melis, Chris Dyer, and Phil Blunsom, some improvements in language modeling which originally were ascribed to complex innovations in the architecture of the network, stem from hyperparameter tuning [@melis2017state]. As mentioned by @lipton2019troubling, there is a tremendous value coming from the thorough understanding of a particular method, and a variety of techniques are vital in the process (like ablation, robustness checks, qualitative error analysis) for the benefit of the whole community.
3. Third, "mathiness"—the use of obscure language and often covering weak argumentation with the alluring but often apparent depth of technical jargon. Again Jacob Steinhardt admitted infusing his 2015 paper co-authored with Percy Liang [@steinhardt2015learning] with an irrelevant theorem to amplify the empirical results. They discussed "staged strong Doeblin chains" which

[1] In both reports, the increase is reported as a relative change between the year 2019 and 2020, with no absolute indicators.

actually had limited pertinence to the learning algorithm—the main subject of a paper.

4. Last but not least—misuse of language. This includes suggestive definitions without proper explanation of what they mean in the context (e.g. inflating good performance in simple NLP tasks to human-level understanding), overloading the papers with technical terminology, or suitcase words (words that can encompass a variety of meanings, e.g. consciousness)

Give an example

Yet another obstacle in the process is the lack of gold standard in dataset creation and taxonomies of abusive language being used for instance in the process of annotating different datasets. Frequently people obtain data from various sources and do not follow any universally used instructions when it comes to annotation, leading to discrepancies between various datasets being tagged within one domain (e.g. hate speech). Lack of expert annotators and proper annotation criteria and instructions are also widespread, with the common practice of hiring untrained workers from Mechanical Turk or other crowdsourcing platforms.

Although there are some initiatives developed in response, most notably, functional tests for Hate Speech Detection Models created by Röttger et al. (2020), or the Online Safety Data Initiative (OSDI) (*About the Online Safety Data Initiative*, n.d.), focused on projects related to improving access to data, standardizing the description of online harms, as well as creating tools and benchmarks for evaluation of technologies focused on safety, much effort must be made before wider adoption of such solutions comes into force.

At the same time, only automated methods can scan through the massive amount of content being generated every day on different platforms. On Facebook, there are more than 3B comments and likes daily ("Facebook," 2012), 500M tweets are sent daily on Twitter (*10 Twitter Statistics Every Marketer Should Know in 2021 [Infographic]*, 2021), and over 2B comments made by users of Reddit in 2020 ("Reddit in 2020," 2020) which is almost 3M comments made daily. With this amount of content, it's either impossible or extremely costly to scale the moderation workforce. One can also have doubts about the ethical aspects of hiring workers who are often unaware of how this kind of task will affect their well-being. Being submersed in the cyber-Augean stables takes a toll on many—as examined by Roberts (2014) & Roberts (2016). Workers hired for such tasks are often low-status and low-wage, isolated and asked to keep what they've seen in secret under restrictive NDAs. This in turn makes the research in the area extremely difficult, since moderators are not allowed to talk about their work conditions or any other related subject. Those who decided to break the NDA are risking a penalty. Screening through the reported user-generated content is connected with exposure to violent and deeply disturbing materials, with child pornography, murders, or suicides as examples of the most extreme cases. This can lead to serious psychological damage, such as depression, or PTSD (Roberts, 2014). Although there are certain initiatives being developed or introduced to reduce the emotional impact of the moderation, like stylistic alterations to a content (applying grayscale or blurring to images) (Karunakaran & Ramakrishan, 2019), workplace wellness programs, clinical support, or psychological training (Steiger, Bharucha, Venkatagiri, Riedl, & Lease, 2021), none of the methods can eliminate the psychological distress completely. Some of the employees filed a lawsuit against Facebook and as a result, the company agreed to pay $52M in compensation for mental health issues developed during the job (Newton, 2020). Also as described by Parks (2019), the work is often performed under time pressure, reviewing 25K pieces of content per day. Spending on average three to five seconds on each image reported for moderation might not lead to the most thoughtful decisions and as shown Stepanikova (2012), high time pressure can amplify human biases. Taking into consideration that Facebook employs 15K moderators (Koetsier, 2020) and most likely more are needed to keep up with the growing amount of content, with the parallel considerations about the negative effects of content moderation on mental health, a collaboration between humans and machines in this area seems inevitable.

4 Pro-active and reactive moderation

There are different approaches when it comes to the moderation of online content. One can follow the workflow of reactive moderation, which happens once the content is published. Harmful messages can be either reported by the users of the platforms or automated methods and then sent for review to moderators. A set of actions can be then taken depending on the platform and their community guidelines—on the content or user level. A harmful message can be deleted, made invisible to other users, or certain profanities can be altered with special signs to censor them. Depending on the type and amount of infraction, a particular user can be warned, muted, shadowbanned, or banned from

further participation in the community for a period of time. The weakness of the reactive method is that the damage is done. Whoever is the recipient of the abusive message has the chance to see it and potentially suffer Wachs et al. (2019). Yet another weakness connected with relying solely on human reports is the content that is harmful but unreported by a receipted or any bystander. Although the exact scale of unreported content is not known, various self-report studies show that a lot of children, teens, or even adults do not report cyberbullying or harassment online ("Free to Play?" 2020; French, 2021; "In," 2017).

Yet another type of moderation can be distinguished as pro-active or pre-moderation. In pre-moderation, automated methods are either based on Artificial Intelligence or other less sophisticated tools (e.g. blacklists) and can screen the content before it gets published. If a type of harmful message gets detected, it can be removed before reaching the recipient. Due to the aforementioned dubious performance of state-of-the-art statistical methods, particularly low precision, they are rarely used autonomously.

Pro-active moderation can be utilized using AI or other methods to promote socially positive engagement. Instead of or in the combination with punitive solutions such as privileges restriction, one can induce empathy or community norms with counter-speech. Counter speech as described by Dangerous Speech Project ("Counterspeech Dangerous Speech Project," 2017) is "any direct response to hateful or harmful speech which seeks to undermine it." As examined by Munger (2017), counter-speech can be effective in the reduction of racist tweets (although only in the condition in which a white male with high followers was approaching another white male). In a study conducted by Bilewicz et al. (2021), a bot disguised as a Reddit user, equipped with normative and empathetic interventions, significantly decreased the amount of personal attacks generated on Reddit. In yet another study by Miškolci, Kováčová, & Rigová (2020), this technique was not effective in changing the behavior of the users (counter-speech here aimed at reducing the prejudice against Roma minority in Slovakia), but encouraged bystanders to express pro-Roma comments on specific Facebook posts. Counter-speech also has been shown to have the potential to increase civility online in studies conducted by Friess, Ziegele, & Heinbach (2021), Molina & Jennings (2018), Han, Brazeal, & Pennington (2018). ...

5 Collective Intelligence Approach to Counter-speech

Traditionally collective intelligence has been defined as "a group or a team's combined capacity and capability to perform a wide variety of tasks and solve diverse problems" ("Collective Intelligence," n.d.). In our paper and in the theoretical underpinnings of the experiment itself, we will be relying on a collective intelligence scope proposed by Nesta, an innovation foundation (https://www.nesta.org.uk), which focuses on a collaboration between human and machine intelligence to develop innovative solutions to social challenges. (Gdzieś stopka(?): Samurai Labs has been one of the 15 recipients of the second round of collective intelligence grants awarded by Nesta: https://www.nesta.org.uk/project-updates/second-round-collective-intelligence-grants/)

The main objective of the experiment was to test whether the level of verbal aggression (personal attacks) of a group of users' regularly attacking others on Reddit can be significantly decreased by community-driven, counter-speech interventions conducted by volunteers in partnership with Artificial Intelligence. Instead of using negative motivation system, the assumption was to test a positive one - convincing verbally violent users to refrain from using cyberviolence based on peer-pressure regulation and experiential learning of a positive set of norms and empathy. Algorithms developed for the detection of personal attacks were used to monitor the activity of experimental groups and notify volunteers about all attacks generated by its' members. Volunteers, after receiving a notification on Slack, could then react with a proper intervention. Such an approach served as a distributed bottom-up voluntary model of moderation based on collective intelligence—utilizing human + machine intelligence.

In the end, what we were able to compare was the following: the effectiveness of the existing Reddit moderation system (predominantly grounded in a punitive authoritarian paradigm) versus the existing moderation system combined with collective intelligence—Artificial Intelligence supported with a crowd of volunteers—who introduced the element of positive peer-pressure.

Empathetic interventions In the first treatment condition, volunteers were encouraged to send empathy-inducing messages focusing either on the target of verbal aggression (e.g. „Hey such words might hurt the other person"), stressing the common humanity aspect that we call share („We are all humans of flesh and blood"), or even infusing the intervention with the emphatic response to the

attacker (Hey I understand your strong emotions...”). At the core of the empathetic interventions, we put forth the notion that goes back to David Hume and Adam Smith. The first one conceived empathy (at the time referred to as sympathy) as mirroring the emotional state of another person. In the academic psychological literature, similar phenomenon was distinguished and coined in the term emotional contagion (https://www.sciencedirect.com/topics/psychology/emotional-contagion): “the process in which an observed behavioral change in one individual leads to the reflexive production of the same behavior by other individuals in close proximity, with the likely outcome of converging emotionally.” For Adam Smith, sympathy consisted of visualizing how the sympathetic person would feel in the particular circumstances of the other — thus here the process was based not so much on mirroring, but rather projecting my imagination of what it is like to be that person in a certain moment). Without further delving into the differentiation between those two, sympathy in both accounts is crucial in the constitution of human beings as social and moral creatures (https://plato.stanford.edu/entries/empathy/). It enables the emotional connection to others and concern for their well-being. In the psychological literature, various kinds of empathetic responses were distinguished - the aforementioned emotional contagion, affective/proper empathy, sympathy, personal distress or cognitive empathy ((https://plato.stanford.edu/entries/empathy/). During the experiment, we couldn't observe or measure whether the empathetic response indeed was evoked. Also, interventions that stated that “such words might hurt the other person” have the underlying assumption that the receiver of the attack might be hurt, but in reality that might not even be the case. Intuitively, and following the argumentation of Thomas Nagel in “What is it like to be a bat?” one can only imagine what it is like - but for me - to be a bat or to be a receiver of the attack. To each observed or imagined experience there is an array of subjective quality to actually experiencing it and in this way receiving a particular message might be met with a unique reception by each conscious being. Also, full epistemic access to the mind and body of another is impossible. But even though such access is impossible and putting aside the broad spectrum of phenomena related to empathetic response, our goal was by utilizing interventions referring to empathy in various forms, changing the behavior of the attacker - convincing him to refrain from using verbal aggression towards others. Whether the message indeed gave rise to empathy was not of importance - one can imagine a hypothetical scenario of a successful intervention in which the attacker changed his behavior long-term and stopped attacking others, hopefully as a result of interventions, but the empathetic response was not even evoked. The behavioral change could be induced by other motivations - e.g. unwillingness from the exclusion from the community. Thus the goal is not to evoke empathy, but through empathetic interventions - substantially limiting the use of verbal aggression.

Normative interventions Normative interventions were grounded in broadly defined norms and normative theories expressed via various means - either specific community guidelines imposed by Reddit or through ethics - deontology, virtue ethics. Whatever the strategy was used, it was supposed to express the unacceptability of verbal aggression in a direct or non-direct way.

A more direct approach referred to deontology in which „actions are good or bad according to a clear set of rules” (https://ethics.org.au/ethics-explainer-deontology/). Such interventions could state that „we have a duty to respect each other during the discussions“, „we have a moral obligation to act in a civil manner while participating in online communities,” „we shouldn't use ad personam in here”. The basic assumption behind those interventions is that there is something we ought to do or are morally required to do - and such actions are the right actions. There are also actions we ought not to do - and such actions are morally wrongful. This line of thinking can be traced back to Immanuel Kant who thought that all universal moral obligations can stem from categorical imperative: “act only in accordance with that maxim through which you can at the same time will that it become a universal law” (Groundwork of the Metaphysic of Morals).

Yet other interventions were encouraged to be expressed in the light of virtue ethics and stressed the importance of adhering to or practicing certain virtues. Such a message was assumed to motivate the moral agent in the process of positive reinforcement and create a feeling of the desirability of certain behaviors, e.g. „capacity to be respectful in a heated discussion is a virtue and requires hard work”. The core assumption here is that the virtue itself is not genetically inherited but is rather a potential or a disposition of a character that can be practice and mastered like a practical skill (even though as mentioned by Natasza Szutta in [cite her book about virtue ethics] virtues and practical skills share some fundamental differences). NS distinguished two types of virtues - affective and cognitive. Affective virtues encompass the emotions and feelings that play an important and positive role in morality and can act as a support in the course of becoming a virtuous man. A cognitive virtue

relates to the intellectual aspect in which one knows how to act in certain situations and understands the rules of morally rightful actions (e.g. what is kindness and how a kind man acts). Here, just as in the case of empathetic interventions, any attempts to measure whether the target of the intervention in the case of a positive outcome (behavioral change) indeed acted in a virtuous way. A virtue of kindness may manifest itself in particular behaviors but as such cannot be identified with the virtuous deed, as highlighted by Natasza Szutta. Although again, the goal of the experiment itself was to change the behavior of the attackers, and measuring whether particular messages contributed to more flourishing individuals in terms of virtue development and character creation lies beyond the scope of this work. Interventions could affect people either way - stimulate their moral reflections or limit their attacks due to consequentialist way of thinking and fear of potential outcomes (e.g. being excluded from the community).

Additionally, Reddit created their social etiquette called "Rediquette" (https://www.redditthelp.com/hc/en-us/articles/205926439) which is "an informal expression of the values of many redditors, as written by redditors themselves." All users are encouraged to abide by it and moderators of communities (called subreddits) existing as a part of the platform have the authority to exclude its members based either on the basis of breaking the rules of the reddiquette or any other local rules imposed by specific subreddit. Volunteers were encouraged to refer to those norms as well, citing specific points, for instance, "Hey there, have you read the reddiquette? It says remember the human." or "Let's recall the reddiquette and adhere to the same standards of behavior online that you follow in real life."

As we have seen, those two categories of interventions - normative and empathetic - encompass a whole variety of categories within. One might think that it would be useful to construct a more diverse set of treatment conditions in which we test each of the ethics or empathy differentiation separately - e.g. one for virtue ethics, deontology, perhaps utilitarian approach, empathy toward the receiver of the attack, empathy towards the attacker, so on and so forth. Additionally, different types of interventions could be tested and tailored depending on the spectrum on which one can be found in the foundations of moral reasoning, as proposed by Jonathan Haidt: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation, and Liberty/Oppression (J. Haidt, The Righteous Mind). Albeit for the sake of the sanity of our volunteers, we decided to create only two but broad categories and let them decide based on the context of the conversation, how they want to respond to the attacker.

Cyberviolence was defined in this experiment as a personal attack - any kind of verbal harassment, insult, or threat directed against the interlocutor in a text-based conversation online. Those were detected using Samurai Labs' cyberviolence detection system.

The following hypotheses were formulated:

H1: If a group of human volunteers notified by an AI-based cyberviolence detection system about cyberviolence generated by the treatment group users (cyberviolence will be defined as a personal attack, harassment, or a threat targeted against an interlocutor) responds with counter-speech interventions, this will result in a decreased cyberviolence level for the whole group after the intervention period.

H2: If two groups receive different types of interventions (empathy-based or normative), then the decrease in cyberviolence will be larger in the case of normative interventions in comparison to the empathy-based ones.

6 Experimental Design and data collection

This was a 6-months field experiment in a digital setting conducted on a popular Q&A and news forum, Reddit (www.reddit.com). We formed treatment and control groups based on three main criteria:

1. During the intervention period, we have expected to have 20 active volunteers at any given time, each willing to conduct 10 interventions daily. Thus, we needed approximately 200 attacks daily generated by the treatment groups.
2. The identification of users who regularly attack others was necessary to measure the effect of interventions at the end.
3. The identification of users who were active during the whole preliminary monitoring period was necessary to minimize the risk of attrition during the study.

User identification process:

1. First, we obtained 1 week of real-time (coming from 15-22nd of February 2020), unmoderated data from Reddit. The content was downloaded from the data stream provided by pushshift.io.

2. Samurai Labs Artificial Intelligence for personal attacks detection was applied to identify users who attacked others at least once within the aforementioned timeframe. This resulted in the identification of 93966 users.
3. We removed all accounts which we suspected not to be run by humans (AutoModerator and all users which had "bot" in the username string). This resulted in the removal of 388 users, thus 93578 were left on our list.
4. Next, we removed users who generated only 1 personal attack during the week (leaving only those who attacked at least twice). As mentioned, the group of those regularly attacking others was crucial to measure the effects of the interventions. This step resulted in the removal of users below the third quartile (Q3). 20124 users were left in our group.
5. Moving forward, we removed users who generated less than 14 comments in this week. We cared about most active users, and 2 comments per day per person on average seemed reasonable (not sure yet how to justify this - 14 comments is below 1st quartile (Q1:28, Q2: 63, Q3:126, mean=103)). This resulted in the removal of 2192 users, so 17932 were left.
6. We discarded users whose personal attacks to all comments ratio was below 2%. This means the inclusion in the sample of users above the 1st quartile. 4422 users were removed, leaving us with a group of 13510.
7. The next step of the process begun on March 9th, 2020, and lasted until May 5th, 2020 (9 weeks). During this period we have monitored the activity of the identified group of 13510 users and applied further selection criteria to make sure we select those who were regularly active and attacked other users.
8. The period of monitoring was divided into weeks. We have discarded those weeks during which technical difficulties occurred with the pushshift.io (resulting in missing data). Thus, we have taken into consideration only 6 full weeks for the period.
9. Users who generated at least 1 attack during 5 out of 6 weeks were identified. First, we planned to restrict the list to only those users, who generate at least 1 attack during each week (6/6) but such restrictive criterion led to only 255 users left, which was not enough for the study. The less restrictive criterion (at least 1 attack generated during 5/6 weeks) resulted in 694 people.
10. Next, we calculated the daily average number of personal attacks generated by the group who met the above criteria (which resulted in 357 attacks per day, 1.94 attacks daily per person on average).
11. Knowing that we need around 200 attacks/daily per treatment group (just enough for volunteers to keep up according to our assumption), we have randomly selected 195 users per each treatment group (normative and empathetic). The rest was delegated as a control group (304 users).

The duration of the experiment, 6 months, was divided into three 2-months periods. The first two months served as a monitoring period to properly select groups and establish baselines. The next 2 months served as treatment period, during which groups received counter-speech comments from volunteers, in response to personal attacks detected by the Artificial Intelligence-based system. The last 2-months served as the post-treatment monitoring period to gather the data needed to evaluate the effectiveness of interventions.

7 Results

References

10 Twitter Statistics Every Marketer Should Know in 2021 [Infographic]. (2021). Retrieved from https://www.oberlo.com/blog/twitter-statistics

About the online safety data initiative. (n.d.).

Barlett, C. P., Simmers, M. M., Roth, B., & Gentile, D. (2021). Comparing cyberbullying prevalence and process before and during the COVID-19 pandemic. *The Journal of Social Psychology*, 1–11.

Bhattacharya, A. (2021). How Covid-19 lockdowns weakened Facebook's content moderation algorithms. Retrieved from https://qz.com/india/1976450/facebook-covid-19-lockdowns-hurt-content-moderation-algorithms/

Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social

network environment. *Aggressive Behavior*, *47*(3), 260–266.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. *International Conference on Social Informatics*, 405–415. Springer.

Collective Intelligence. (n.d.). Retrieved from https://www.oxford-review.com/oxford-review-encyclopaedia-terms/collective-intelligence/

Counterspeech Dangerous Speech Project. (2017). Retrieved from https://dangerousspeech.org/counterspeech/

Facebook: 3.2 Billion Likes & Comments Every Day. (2012). Retrieved from https://martech.org/facebook-3-2-billion-likes-comments-every-day/

Free to Play? Hate, Harassment and Positive Social Experience in Online Games 2020. (2020). Retrieved from https://www.adl.org/free-to-play-2020

French, C. (2021). As the pandemic forces us online, LGBTQ2S+ teens deal with cyberbullying. Retrieved from https://www.ctvnews.ca/canada/as-the-pandemic-forces-us-online-lgbtq2s-teens-deal-with-cyberbullying-1.5430945

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication*, 1–23.

Gerrard, Y. (2020). <? covid19?> The COVID-19 mental health content moderation conundrum. *Social Media+ Society*, *6*(3), 2056305120948186.

Geva, M., Goldberg, Y., & Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. *arXiv Preprint arXiv:1908.07898*.

Grant, J. (2021). Australia's eSafety commissioner targets abuse online as covid-19 supercharges cyberbullying | the strategist. Retrieved from https://www.aspistrategist.org.au/australias-esafety-commissioner-targets-abuse-online-as-covid-19-supercharges-cyberbullying/

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is" love" evading hate speech detection. *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2–12.

Han, S.-H., Brazeal, L. M., & Pennington, N. (2018). Is civility contagious? Examining the impact of modeling in online political discussions. *Social Media+ Society*, *4*(3), 2056305118793404.

Hoff, D. L., & Mitchell, S. N. (2009). Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration*.

In:Game Abuse. (2017). Retrieved from https://www.ditchthelabel.org/research-papers/ingame-abuse/

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Karunakaran, S., & Ramakrishan, R. (2019). Testing stylistic interventions to reduce emotional impact of content moderation workers. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 50–58.

Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2016). *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis.

Koetsier, J. (2020). Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day. Retrieved from https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/

L1ght. (2020). L1ght releases groundbreaking report on corona-related hate speech and online toxicity. Retrieved from https://l1ght.com/l1ght-releases-groundbreaking-report-on-corona-related-hate-speech-and-online-toxicity/

Laub, Z. (2019). Hate Speech on Social Media: Global Comparisons. Retrieved from https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine learning scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue*, *17*(1), 45–77.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech

detection: Challenges and solutions. *PloS One*, *14*(8), e0221152.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on facebook: The case of the roma minority in slovakia. *Social Science Computer Review*, *38*(2), 128–146.

Molina, R. G., & Jennings, F. J. (2018). The role of civility and metacommunication in facebook discussions. *Communication Studies*, *69*(1), 42–66.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649.

Musgrave, K., Belongie, S., & Lim, S.-N. (2020). A metric learning reality check. *European Conference on Computer Vision*, 681–699. Springer.

Newton, C. (2020). Facebook will pay $52 million in settlement with moderators who developed PTSD on the job. Retrieved from https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

Parks, L. (2019). Dirty data: Content moderation, regulatory outsourcing, and the cleaners. *Film Quarterly*, *73*(1), 11–18.

Reddit in 2020. (2020). Retrieved from https://www.reddit.com/r/blog/comments/k967mm/reddit_in_2020/

Roberts, S. T. (2014). *Behind the screen: The hidden digital labor of commercial content moderation*. University of Illinois at Urbana-Champaign.

Roberts, S. T. (2016). *Commercial content moderation: Digital laborers' dirty work*.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., ... Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, *93*, 333–345.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2020). Hatecheck: Functional tests for hate speech detection models. *arXiv Preprint arXiv:2012.15606*.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, *117*(48), 30033–30038.

Sorrentino, A., Baldry, A. C., Farrington, D. P., & Blaya, C. (2019). Epidemiology of cyberbullying across europe: Differences between countries and genders. *Educational Sciences: Theory & Practice*, *19*(2).

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., & Lease, M. (2021). The psychological well-being of content moderators. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI*, *21*.

Stepanikova, I. (2012). Racial-ethnic biases, time pressure, and medical decisions. *Journal of Health and Social Behavior*, *53*(3), 329–343.

Swamy, S. D., Jamatia, A., & Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 940–950.

Vogels, E. A. (2021). The State of Online Harassment. Retrieved from https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Wachs, S., Wright, M. F., Sittichai, R., Singh, R., Biswal, R., Kim, E., ... others. (2019). Associations between witnessing and perpetrating online hate in eight countries: The buffering effects of problem-focused coping. *International Journal of Environmental Research and Public Health*, *16*(20), 3992.

Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. S. (2019). Errudite: Scalable, reproducible, and testable error analysis. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 747–763.

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, *7*, e598.