Journal of Artificial Intelligence Research 71 (2021) 431-478 Submitted 12/2020; published 07/2021

Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective

Svetlana Kiritchenko SVETLANA.KIRITCHENKO@NRC-CNRC.GC.CA
Isar Nejadgholi ISAR.NEJADGHOLI@NRC-CNRC.GC.CA
Kathleen C. Fraser KATHLEEN.FRASER@NRC-CNRC.GC.CA
National Research Council Canada
1200 Montreal Rd., Ottawa, ON, Canada

Abstract

The pervasiveness of abusive content on the internet can lead to severe psychological and physical harm. Significant effort in Natural Language Processing (NLP) research has been devoted to addressing this problem through abusive content detection and related sub-areas, such as the detection of hate speech, toxicity, cyberbullying, etc. Although current technologies achieve high classification performance in research studies, it has been observed that the real-life application of this technology can cause unintended harms, such as the silencing of under-represented groups. We review a large body of NLP research on automatic abuse detection with a new focus on ethical challenges, organized around eight established ethical principles: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. In many cases, these principles relate not only to situational ethical codes, which may be context-dependent, but are in fact connected to universal human rights, such as the right to privacy, freedom from discrimination, and freedom of expression. We highlight the need to examine the broad social impacts of this technology, and to bring ethical and human rights considerations to every stage of the application life-cycle, from task formulation and dataset design, to model training and evaluation, to application deployment. Guided by these principles, we identify several opportunities for rights-respecting, socio-technical solutions to detect and confront online abuse, including ‘nudging’, ‘quarantining’, value sensitive design, counter-narratives, style transfer, and AI-driven public education applications.

1. Introduction

With the increased use of social media, especially among young people, serious concerns about safety and inclusion in online communications have been raised. According to a 2017 survey conducted by Pew Research Center, more than 40% of U.S. adults have been personally subjected to online harassment and 18% have been the target of severe behaviors such as physical threats and sexual harassment (Duggan, 2017). In a 2019 study, 36.5% of U.S. high school students said that they have been cyberbullied during their lifetime (Hinduja & Patchin, 2020). Similar or even more disturbing statistics have been collected world-wide over the past 10 years. Often, the victims of online abuse are from the most vulnerable parts of society: ethnic minorities, the LGBTQ community, or people with disabilities, for example. Exposure to toxic and hateful comments online can lead to psychological trauma, radicalization, and even self-harm and suicide (Van Geel et al., 2014; Mishra et al., 2019). In response, many social media platforms strive to monitor online content and quickly remove abusive posts, but the sheer volume of posts poses significant problems. Furthermore, human content moderators report high rates of burn-out, depression, and PTSD as a result of viewing such toxic content, day in and day out (Arsht & Etcovitch, 2018). Automatic detection of abusive content can provide assistance and partially alleviate the burden of manual inspection.

A wealth of research in Natural Language Processing (NLP) has been devoted to the problem of automatic abusive content detection. Here, we use the term *abusive* broadly, defining it as any language that could offend, demean, or marginalize another person, covering the full range of inappropriate content from profanities and obscene expressions to threats and severe insults. Abusive language detection has been studied under a plethora of names, such as detection of *flaming* (e.g., Spertus, 1997), *cyberbullying* (e.g., Dadvar et al., 2013), *online harassment* (e.g., Golbeck et al., 2017), *hate speech* (e.g., Djuric et al., 2015; Davidson et al., 2017), *toxicity* (e.g., Dixon et al., 2018; Aroyo et al., 2019), *microaggression* (e.g., Breitfeller et al., 2019; Ali et al., 2020), *stereotyping* (e.g., Nadeem et al., 2020; Fraser et al., 2021), *unhealthy conversations* (e.g., Price et al., 2020), and others. While these sub-areas of the general space of abusive language tackle similar problems, they differ in their focus and scope. Recent surveys by Schmidt and Wiegand (2017), Fortuna and Nunes (2018), Mishra et al. (2019), Vidgen et al. (2019), Vidgen and Derczynski (2020), and Salawu et al. (2020) summarize the advancements in these areas focusing mostly on the technical issues and the variety of data collection and machine learning approaches proposed for the tasks.

In contrast, here we examine the task of automated abusive language detection from the *ethical* viewpoint, bringing together both technical and social issues under a single ethical and human rights framework. We begin by gathering all the related sub-fields, and briefly survey the past work with a focus on the different task formulations, common data collection and annotation techniques, algorithms, and applications. Works addressing online abuse detection in any form, from hate speech and aggressive language to more subtle offenses such as microaggressions and stereotyping, are included in this survey. We focus here specifically on detecting abusive language in *individual textual utterances*, although recent work has also begun to tackle multimedia data (e.g., Kiela et al., 2020) and to incorporate the broader context of conversations (Pavlopoulos et al., 2020; Vidgen et al., 2021). We then discuss in detail the challenges that the field faces from the ethical perspective, using the Harvard 'Principled Artificial Intelligence' framework as a scaffold (Fjeld et al., 2020). These challenges include fairness and mitigation of unintended biases, transparency, explainability, privacy, safety, and security. We discuss the trade-off between the right to free speech and the right to human dignity, and our professional responsibility to promote and protect all human rights in our work as AI researchers and practitioners.

We then turn to the future: how can the field progress in the most responsible and ethical manner? We identify several directions for future work. Our findings from the literature emphasize that ethical considerations must be addressed throughout the entire development pipeline, from the task formulation, data collection, and annotation, through to model training and evaluation, and finally in deployment. In addition to compiling recommendations for each stage in the pipeline, we also review information from related literature in the social sciences, and suggest how we might integrate that work into our technical solutions.

Enumerating these ethical dilemmas is not simply an academic exercise; inattention to these issues can lead to human and economic harms in the real world. In Table 1 we present recent examples from the popular press describing negative outcomes related to automated content moderation on the web. Each of the ethical themes in the table will be discussed in detail in Section 3. Through a better understanding of the ethical landscape, we hope to inspire new and creative solutions to effectively confront online abuse.

2. Overview of the Common Practices

We start by summarizing the common practices in defining the task, collecting and annotating data, and training a predictive model. Further, we discuss some current applications of abusive language detection technology.

2.1 Task Formulation

The abusive language detection task has typically been formulated as a supervised classification problem across various definitions and aspects of abusive language. In addition to the main task of determining whether a text is abusive or not, several other dimensions have been explored, including categories of abuse, implicit versus explicit abuse, target of abuse, legality of abuse, and the implied stereotypes in abusive language (Waseem et al., 2017; Fišer et al., 2017; Poletto et al., 2017; Vidgen et al., 2019; Niemann et al., 2019; Zufall et al., 2020; Sap et al., 2020). Banko et al. (2020) assessed the most common definitions of abuse used in the domain of online content moderation technologies across industry, government policies, online communities, and the health sector. They unified these definitions under the umbrella term *online harm* and recommended 1) using objective criteria and fine-grained classes, and 2) considering the target of abuse and the potential downstream actions to create high-quality definitions.

Here, we look at two main dimensions often considered when formulating the task of abusive language detection: expression of abuse and target of abuse. Also, in Table 2, we show examples of utterances from existing datasets along with labels that describe the expression and the target of abuse.

Expression of abuse: Multiple terms and definitions have been used to describe abusive content depending on how the abuse is expressed (e.g. hate speech, insult, physical threat, stereotyping). Focusing on slightly different aspects of abuse, these categories have obscure boundaries, and are often challenging for humans and machines to tell apart (Poletto et al., 2017; Founta et al., 2018). Even the definitions of a single category (e.g., hate speech) can vary among researchers, and result in incompatible datasets (Fortuna et al., 2020). For example, some messages labeled as hate speech in the dataset by Waseem and Hovy (2016) would not meet the requirements for this category in the works by Nobata et al. (2016) and Davidson et al. (2017). Van Aken et al. (2018) questioned 10–15% of manually obtained labels on two widely used datasets, Kaggle Toxicity by Jigsaw and Google[1] and the one by Davidson et al. (2017). In some cases, more accurate formulations of abusive behaviour are only possible if additional information, such as attributes of the author and of the recipient,

[1] `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`

Ethics theme	Real-life example
Promotion of human values	Abusive language detection is deployed to protect people; however, in some cases it can actually silence marginalized voices. For example, Black activists have reported that Facebook deletes posts in which they discuss their own experiences of racism.[1]
Fairness & non-discrimination	In 2019 it was discovered that posts written by Black writers are 1.5 times more likely to be marked as offensive by some of the leading toxicity detectors, and posts written in African American English even more so, leading to the suppression of Black voices as well as a negative user experience.[2]
Transparency & explainability	In fall 2020, small business owners noticed that their seemingly innocuous ads were being automatically removed by Facebook's content moderation algorithm, leading to lost revenue. Because the business owners didn't understand why the ads were removed, they were frustrated and, importantly, did not know how to avoid the same thing happening again in the future.[3]
Privacy	While data privacy is a serious concern to computer scientists, maintaining user personal privacy is also critical. In the infamous GamerGate scandal, video game developer Zoe Quinn was subjected to harassment and threats on Twitter, and had to leave her home when her address was posted on the site.[4]
Safety & security	In 2018, it was discovered that simply adding positive words, such as *love*, to otherwise offensive posts was enough to fool the Perspective API toxicity detector. Systems must be secure against such simplistic, as well as more sophisticated, attacks.[5]
Accountability	Online platforms are accountable not only to their own terms of service, but to the expectations of their users and advertisers. In 2017, the proliferation of hateful and offensive content on YouTube led to major advertisers withdrawing their spots; in response, YouTube was forced to improve their approach to content moderation.[6]
Human control of technology	When decisions are made automatically, it is essential for users to be able to appeal for a human review. Activists and journalists in the Middle East claim that their Facebook accounts have been removed by artificial intelligence algorithms that misinterpret their content as promoting terrorism. Facebook acknowledged that the decisions must be reviewed by a human moderator with "regional and language expertise."[7]
Professional responsibility	AI researchers and engineers have a professional responsibility not to build technology to deliberately harm society or human well-being. For example, software similar to what is proposed for abusive language detection could be manipulated by governments for censorship and surveillance instead, such as that reported by Hong Kong activists in 2020.[8]

Table 1: Examples from recent news stories illustrating the real-life importance of each ethical theme as they relate to automated abusive language detection.

[1] https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/

[2] https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter

[3] https://fortune.com/2020/11/28/facebooks-ai-is-mistakenly-banning-some-small-business-ads/

[4] https://www.nytimes.com/interactive/2019/08/15/opinion/gamergate-zoe-quinn.html

[5] https://thenextweb.com/artificial-intelligence/2018/09/11/googles-hate-speech-ai-easily-fooled/

[6] https://www.nytimes.com/2017/04/17/arts/youtube-broadcasters-algorithm-ads.html

[7] https://www.nbcnews.com/tech/tech-news/facebook-doesn-t-care-activists-say-accounts-removed-despite-zuckerberg-n1231110

[8] https://www.nytimes.com/2020/08/25/technology/hong-kong-national-security-law.html

Utterance	**Possible Categories**	**Target of Abuse**	**Type of Abuse**
The China Pneumonia is getting out of control as mainlanders are receiving misinformation and not paying adequate attention to have this virus contained.	non-abusive, entity directed criticism	-	-
This movie was a f*cking piece of sh*t.	obscene	-	explicit
White people have been fighting eachother for millennia +now we think importing millions of 3rd world migrants isn't going to cause issues.	offensive, racist, insult	immigrants	implicit
I swear bitches wouldn't have anything to worry about but they don't know how to shut the f*ck up.	offensive, toxic, sexist, hateful	women	explicit
Shove it up your f*cking *ss and burn in hell.	attack, threat, abusive	recipient	explicit

Table 2: Examples of utterances annotated for abuse-related categories. The utterances are taken from the datasets created by Sap et al. (2020), Wulczyn et al. (2017), Vidgen et al. (2020a).

and the context of the conversation, is available. However, due to the complexities of data collection and annotation, as well as privacy concerns, most of the previous research has considered only a limited number of coarse-grained labels to characterize the expression of abuse.

Another practical distinction that ties to expression of abuse is whether the abusive language is *explicit* or *implicit* (Waseem et al., 2017). Explicit abuse is relatively easy to recognize as it contains explicit obscene expressions and slurs. Implicit abuse, on the other hand, may not be immediately apparent as it can be obscured by the use of sarcasm, humor, stereotypes, ambiguous words, and lack of explicit profanity. Collecting data using known abusive words and expressions would fail to assemble representative sets of texts with implicit abuse. Thus, in many existing datasets implicit abuse can be found in only a small proportion of instances (Wiegand et al., 2019). Furthermore, implicit abuse presents additional challenges to human annotators as sometimes specific background knowledge and experience are required in order to understand the hidden meaning behind implicit statements (Sap et al., 2020; Breitfeller et al., 2019; Field & Tsvetkov, 2020). To deal effectively with this class of abuse, annotated datasets focusing on implicitly abusive language are needed so that automatic detection systems are exposed to a wide variety of such examples through their training data (Wiegand et al., 2021).

Target of abuse: Abusive speech can be directed towards a particular person, entity, or group, or contain undirected profanities and indecent language (Zampieri et al., 2019a). While obscene language, in general, can be disturbing to some audiences, abuse targeting specific individuals or groups is often perceived as potentially more harmful and more concerning for society at large. Therefore, the majority of research on abusive language detection has been devoted to targeted abuse. Waseem et al. (2017) differentiated two target types: an individual and a generalized group. They argued that the distinction between an attack directed towards an individual or a generalized group is important from both the sociological and the linguistic points of view. Thus, this distinction may call for different handling of the two types of abusive language when manually annotating abusive speech and when building automatic classification systems. For example, in research on cyberbullying, where abusive language is directed towards specific individuals, more consensus in task definition and annotation instructions can be found, and higher inter-annotator agreement rates are often observed (Dadvar et al., 2013). A third target type—entity or concept—can also be considered (Zampieri et al., 2019a; Vidgen et al., 2019). Acceptable criticism of an entity (e.g., a country), a concept (e.g., religion), an organization, or an event, can be semantically similar to abusive language. However, there is often a thin line between criticizing a concept and attacking people associated with the concept (e.g., anti-Islamic discourse can induce hatred towards Muslims).

Each of the three target types can be further divided into subtypes (Vidgen et al., 2019; Sap et al., 2020). For example, a group of persons can be targeted because of their ethnic, religious, or political identity. Addressing one target subtype at a time (e.g., online abuse based on race) may simplify the task at all stages, from data collection and annotation to building an automatic detection system.

2.2 Language Resources

Lexicons and annotated corpora are critical resources for the automatic detection of abusive language. Generally, it is laborious and costly to create such resources due to the sparse nature of online abusive content and the ambiguities in the definitions of abusive behaviour.

2.2.1 Lexicons

Several lexicons of abusive expressions have been built manually, automatically, or semi-automatically (Razavi et al., 2010; Gitari et al., 2015; Wiegand et al., 2018a). Lexicons have been used to improve the detection of abusive utterances, usually in combination with other features. For example, Wiegand et al. (2018a) created a large lexicon and demonstrated its effectiveness in the cross-domain detection of abusive micro-posts. However, lexicons can quickly become out-of-date as users coin new hateful expressions to evade filters, and are not resilient against spelling errors and typos. Furthermore, offensive texts can contain no words or expressions commonly considered abusive in isolation.

In the current landscape of the field, lexicons are mainly used to search for examples of abusive content through querying social media APIs. Abusive content is relatively infrequent, and random sampling results in datasets extremely skewed towards benign samples (Founta et al., 2018; Schmidt & Wiegand, 2017). Several works designed specific search strategies, such as snowballing (Hosseinmardi et al., 2015), crowd-sourcing (Breitfeller et al., 2019), and characterizing hateful users (Ribeiro et al., 2018), to boost the number of abusive examples in datasets. However, most existing sampling strategies mainly rely on using known abusive/profane lexicons to find abusive content. For example, Waseem and Hovy (2016) focused on sexism and racism and collected tweets matching query words that are likely to occur in these cases. Davidson et al. (2017) used a lexicon of words and phrases identified by users as related to hate speech, and Vidgen et al. (2020a) used a list of keywords to collect hashtags associated with anti-Asian prejudice.

Although keyword search is simple and efficient, Wiegand et al. (2019) demonstrated that the choice of search terms for querying social media can lead to topic bias in trained classifiers. Poletto et al. (2020) expanded the analysis of keyword-based data collection strategies in a systematic review of hate speech lexicons in multiple languages and highlighted the need for a unified taxonomy for harmful content search.

2.2.2 Annotated Datasets

A number of datasets manually annotated for abusive language have been made available, each covering a limited range of harmful behaviours. Some of these datasets were released as part of shared tasks that attracted numerous participants (Vu et al., 2020; Basile et al., 2019; Zampieri et al., 2019b). Data can be collected from a single platform, such as Yahoo! (Djuric et al., 2015), Wikipedia (Wulczyn et al., 2017), Facebook (Kumar et al., 2018), Twitter (Waseem & Hovy, 2016; Davidson et al., 2017; Founta et al., 2018), or from multiple discussion forums (Van Bruwaene et al., 2020). Sigurbergsson and Derczynski (2020) demonstrated that although language and user behaviour vary between platforms, sharing information across languages and platforms improves the performance of automatic systems.

As mentioned in the discussion on task formulation, in addition to the inherent subjectivity of language, differing understandings of what to consider abusive language can result in vague and even contradictory category definitions. The lack of clear, intuitive definitions and comprehensive instructions for human annotators can lead to low inter-annotator agreement even within datasets. For example, Poletto et al. (2017) asked human annotators to label abusive tweets with one or more of the following five categories: hate speech, aggressiveness, offensiveness, irony/sarcasm, and stereotypes. They found low inter-annotator agreement rates, especially for aggressiveness and offensiveness, even though detailed annotation guidelines were provided. Similarly, Founta et al. (2018) found low agreement among annotators and high correlations among closely related categories.

Comprehensive annotation guidelines are crucial for obtaining reliable annotations. To ensure the clarity of the annotation process, many researchers have released the guidelines provided to the annotators as well as the annotation schema and the important examples (Waseem & Hovy, 2016; Nobata et al., 2016; Wulczyn et al., 2017; Davidson et al., 2017; Founta et al., 2018; Zampieri et al., 2019a). Combining the annotations from multiple annotators can also minimize the effects of subjectivity. The proportion of majority votes per instance represents the level of agreement and can serve as a rough estimate for severity of abuse. However, most often, the votes are aggregated into a single label. Wiegand et al. (2019) and Davidson et al. (2017) used majority voting whereas Gao and Huang (2017) annotated a statement as hate speech if at least one annotator labeled it as hateful. Golbeck et al. (2017) collected judgements from two trained annotators, and a third annotator was employed only if the first two disagreed. Additionally, Waseem et al. (2017) and Nobata et al. (2016) observed that expert annotators reach higher inter-rater agreements and produce better quality annotations compared to crowd-sourced workers.

For more detail on language resources for detecting online abuse we refer the reader to recent surveys by Poletto et al. (2020) and Vidgen and Derczynski (2020). Poletto et al.

(2020) conducted a systematic review of text collections annotated for hate speech. They compared the corpora along four dimensions: type of behaviour, data source, annotation framework, and language. Vidgen and Derczynski (2020) enumerated the sources of inconsistencies in creating abusive language datasets, highlighted the barriers to data sharing and the lack of infrastructure needed for open-source research, and recommended a set of best practices for data sampling and annotation to improve the dataset creation procedures. They also built a repository of corpora annotated for hate speech, online abuse, and offensive language.[2]

2.3 Algorithms

Equivalent to the various annotation schemes deployed to annotate datasets, the task of abusive language detection has been formulated as a supervised binary (with only two classes, e.g., Wulczyn et al., 2017), multi-class (with more than two classes, e.g., Subramani et al., 2019), multi-label (with instances belonging to one or more classes, e.g., Ibrohim and Budi, 2019), or multi-task (with multiple learning objectives solved simultaneously, e.g., Abu Farha and Magdy, 2020) classification problem. Fortuna and Nunes (2018) conducted a systematic review on automatic detection of hate speech in text and enumerated dictionary-based, rule-based, and feature-based techniques, as well as early deep learning models applied to this task. Since then, deep learning models, such as convolutional neural networks (CNN) (Gambäck & Sikdar, 2017), recurrent neural networks (RNN) (Zhang et al., 2018b), and transformers (Alonso et al., 2020), have been applied to build automatic abuse detection systems, and high performances have been achieved as these algorithms improved. Naseem et al. (2020) showed that besides the training algorithms, preprocessing methods significantly impact the performance of the trained classifiers, which is often overlooked. They developed an intelligent tweet processing method that minimizes information loss at the preprocessing stage. Ayo et al. (2020) focused specifically on hate speech classification of Twitter data and surveyed the machine learning approaches used for this task. Aluru et al. (2020) reviewed the deep learning algorithms applied to multilingual hate speech detection.

Since 2018, pretrained language models have become a ubiquitous language resource for training NLP classifiers. Salminen et al. (2020) used BERT (Devlin et al., 2019) to generate text representations and showed that these representations are robust across social media platforms. Mozafari et al. (2019) fine-tuned BERT and examined various prediction layers in the fine-tuning step. They showed that a CNN-based fine-tuning strategy is more effective than using RNN-based or nonlinear output layers. Wiedemann et al. (2020) evaluated and compared various transformer-based masked language models fine-tuned to detect offensive language and its sub-categories. They concluded that an ensemble based on the ALBERT (Lan et al., 2020) model achieved the best overall performance, and RoBERTa (Liu et al., 2019) achieved the best results among individual language models.

Despite the high performances of the state-of-the-art deep learning models in cross-validation settings, Arango et al. (2019) showed that these models are not robust when it comes to cross-dataset generalization. Risch and Krestel (2020) proposed an ensemble of multiple fine-tuned BERT models to address the problem of high variance in the output

[2] `https://hatespeechdata.com`

of models fine-tuned on a small dataset. Miok et al. (2020) proposed a Bayesian method within the attention layers of the transformer models to provide reliability estimates for the decisions made by the multi-lingual fine-tuned classifiers. They showed that this method of transformer layer calibration not only improved the performance of the classifiers, but also reduced the workload of human moderators by providing reliability estimates.

Multi-task learning has been another approach deployed to improve detection of offensive language. For example, Safi Samghabadi et al. (2020) proposed an end-to-end neural model using attention on top of BERT that incorporated a multi-task learning paradigm to learn a multi-class "Aggression Identification" task and a binary "Misogynistic Aggression Identification" task simultaneously. Waseem et al. (2018) demonstrated that learning to detect hate speech alongside an auxiliary task improved robustness across datasets originating from different distributions and labeled under differing annotation guidelines. Ousidhoum et al. (2019) showed that multi-lingual multi-task learning can improve the performance on tasks for which the amount of annotated data is limited.

Another line of research has put effort towards addressing the problem of small offensive language datasets through data augmentation or modified learning algorithms. Guzman-Silverio et al. (2020) explained that when the size of the dataset is smaller than 10K instances, different initialization random seeds for the fine-tuning of the final layer lead to substantially different models. They customized different data augmentation methods, originally developed for English, to augment a Spanish dataset and showed the effectiveness of their methods. Rizos et al. (2019) and Wullach et al. (2020) used deep generative language models to produce realistic hate and non-hate utterances and demonstrated that training with the augmented dataset improved performances across different hate speech datasets. Zero- and few-shot learning techniques are other approaches that have been shown to be effective in dealing with low resources in hate speech detection (Stappen et al., 2020).

2.4 Applications

Traditionally, harmful content has been detected by human moderators or flagged by users (Gillespie, 2018). As the amount of user-generated content grew dramatically in recent years, multiple stakeholders are starting to adopt automatic content moderation. Gorwa et al. (2020) investigated how major platforms use automated tools to manage copyright infringement, terrorism and toxic speech. They identified key political and ethical issues around relying on these systems in terms of transparency, fairness and depoliticisation. Here, we review the main applications of the technologies developed for automatic detection of abuse.

Each social media platform develops their own technologies and policies around content moderation, often questioned by public and lawmakers (Isaac & Browning, 2020). For example, at Facebook, human content moderators are employed to review the content that is flagged by automatic systems (Koetsier, 2020). On Reddit, each community sets its own rules and policies, and moderation relies on volunteer moderators who might choose to benefit from automated technologies (Basu, 2019). With the advent of COVID-19 in 2020, social media platforms were forced to rely more heavily on fully automated content moderation, which proved to be too erroneous and highlighted the importance of keeping human moderators in the loop (Scott & Kayali, 2020).

Other platforms might leverage ready-to-use APIs for content moderation. Perspective API, developed by Jigsaw, is a widely-used and commercially-deployed toxicity detector that can support human moderators and provide feedback to users while they type.[3] For example, in 2018, the New York Times announced that they use this system as part of their moderation workflow (Adams, 2018). This system generates a probability of toxicity for each queried sentence and leaves it to the users to decide how to use this toxicity score. Moreover, many research works have adopted the use of Perspective API for studying the patterns of offensive language in online platforms (Sap et al., 2019; Ziems et al., 2020).

Apart from social media and news platforms, other stakeholders have been using automatic detection of harmful content. HaterNet is an intelligent system deployed by the Spanish National Office Against Hate Crimes of the Spanish State Secretariat for Security that identifies and monitors the evolution of hate speech in Twitter (Pereira-Kohatsu et al., 2019). Smart policing is another area that could potentially benefit from automatic detection of threats to people or nationalities (Afzal & Panagiotopoulos, 2020). As another example, PeaceTech combats hate speech by identifying inflammatory lexicons on social media and offering alternative words and phrases as a key resource for local activists and organizations.[4] Raufi and Xhaferri (2018) envisioned a lightweight classification system for hate speech detection in the Albanian language for mobile applications that users can directly manage.

Automatic detection of harmful content has been deployed to increase the safety of virtual assistants and chatbots, through prevention of hate speech generation. As Gehman et al. (2020) demonstrated, automatic generative models can produce toxic text even from seemingly innocuous prompts. They showed that even with controllable generative algorithms, the produced text can still be unsafe and harmful. Xu et al. (2020) investigated the safety of open-domain chatbots. They designed a pipeline with human and trained models in the loop to detect and mitigate the risk of unsafe utterance generation, avoid sensitive topics and reduce gender bias in the generated text. As a broader preventative approach to increasing the safety of online conversations, Haapoja et al. (2020) suggested that deploying a hate-speech detection algorithm can be understood as an effort to not only detect but also preempt unwanted behavior. They uncovered strategies planned by multiple stakeholders to resist the model. They illustrated that when a model is deployed, while "gaming the system" is an important part of the interactions between human and the algorithm, sometimes humans play against each other, rather than against the technology. However, the practical and technical implications of this approach have not been studied.

Automatic detection of abusive language can be potentially used to identify illegal online behaviour. Some types of abusive statements, such as hate speech and defamatory allegations, are not only morally unacceptable, but also illegal in several countries. Social media platforms are obligated to quickly remove such statements from public view. Accordingly, Fišer et al. (2017) proposed to classify online discussions along the legal dimension into three categories: (1) legally punishable (hate speech, threats, and defamatory statements), (2) inappropriate (insults, offensive speech, obscenity, profanity, and vulgarity), and (3) acceptable.[5] To automatically determine if a statement is illegal, the corresponding laws

[3] `https://www.perspectiveapi.com/#/home`

[4] `https://www.peacetechlab.org/hate-speech`

[5] Fišer et al. (2017) put hate speech in a separate category to match the Slovene legal framework.

need to be translated into manageable NLP tasks (Zufall et al., 2020). For example, the European Union Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law defines punishable hate speech as "publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin."[6] This definition would translate into two NLP tasks: (1) target detection (whether the target is a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin), and (2) abusive act detection (whether the text incites violence or hatred). However, each jurisdiction has its own definition of online abuse that is considered illegal, and those definitions can also evolve over time. Thus, while an international social media platform or application may be interested in applying global standards to keep their audience and advertisers engaged with the platform, they must still address a wide range of abusive language as required to be in compliance with local law. Therefore, the NLP research community should focus on the broader problem of abusive language detection while still designing solutions that can be transparent and easily adaptable to a specific set of requirements.

3. Current Ethics-Related Challenges

We now review current technological and sociological challenges in the field of automatic online abuse detection with respect to eight common ethical and human rights principles. These eight principles emerged as core thematic trends outlined in many ethical AI frameworks and guidelines as summarized in the recent study by the Berkman Klein Center for Internet & Society at Harvard University (Fjeld et al., 2020). The study analysed 36 prominent AI principles documents from governments and intergovernmental organizations, the private sector, professional associations, advocacy groups, and multi-stakeholder initiatives, representing Latin America, East and South Asia, the Middle East, North America, and Europe. These include "Draft Ethics Guidelines for Trustworthy AI" by the European High Level Expert Group on AI, "White Paper on AI Standardization" by the Standards Administration of China, "Toronto Declaration on Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems" by Amnesty International, "Ethically Aligned Design" by IEEE, "Microsoft AI Principles" by Microsoft, and "AI at Google: Our Principles" by Google, among others. Despite varying cultural contexts and objectives, there seems to be a convergence towards the eight main principles: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. Moreover, the more recent documents tend to include all eight of the principles. Thus, these principles can be viewed as the "normative core" of ethical AI. Table 3 summarizes the ethical challenges in addressing online abuse, which we discuss in detail in what follows.

3.1 Promotion of Human Values

The principle of the promotion of human values is largely congruous with fundamental human rights, and includes the following three main concerns: supporting and promoting

[6] `https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:32008F0913`

Ethics Theme	Online Abuse Specific Issues	Related AI Challenges
Promotion of human values	Finding balance over two conflicting human rights, freedom of speech and respect for equality and dignity	Overcoming ambiguous and non-realistic task formulations; designing alternative applications to ensure safe communication environments for all
Fairness & non-discrimination	Striving for equal system performance on texts that are about or written by different demographics	Collecting representative datasets; identifying, quantifying and mitigating potentially unfair system outputs; optimizing measures of fairness besides overall accuracy
Transparency & explainability	Moving away from making critical decisions using black box models; providing developers with tools to inspect systems' behavior and identify risks; providing lay users with explanations on automated decisions	Producing and maintaining high-quality documentation (data sheets and model cards); designing and using interpretability tools to detect biases in models; providing accessible explanations to users
Privacy	Ensuring data privacy, personal privacy, and users' right to control their own data	De-identifying personal data; applying privacy-preserving computation (e.g., federated learning); allowing users to remove their data from training corpora
Safety & security	Considering consequences of false positive and false negative decisions; building systems that do not heavily rely on keywords, are not easy to deceive, and are robust against poisoning and adversarial attacks	Measuring and minimising the risk of false decisions; identifying system vulnerabilities, including susceptibility to spurious correlations; improving the out-of-distribution robustness; testing systems in real-world scenarios
Accountability	Auditing systems and assessing their impact on individuals, society and environment; ensuring the ability to appeal; setting legal responsibilities	Auditing design decisions internally throughout all stages of application development and deployment; designing and employing interpretability and explainability tools
Human control of technology	Moving away from fully automated moderation due to inaccurate systems; enabling users to appeal automated decisions and request human review	Enabling human-in-the-loop technologies; providing rationale to users to enable appeals
Professional responsibility	Building accurate systems; considering potential long-term effects; refusing to work on harmful applications; engaging all stakeholders; upholding scientific integrity	Evaluating system performance in various settings; involving stakeholders in the design process; raising public awareness for long-term possible harms of technology (e.g., censorship)

Table 3: Ethics and human rights related issues in online abuse detection, and the associated NLP/AI challenges. Each theme is discussed in detail in Section 3.

human values and human flourishing, benefiting society, and ensuring broad access to technology (Fjeld et al., 2020). Online abusive content detection brings forward two conflicting human values: freedom of speech and respect for equality and dignity (Maitra & McGowan, 2012; Waldron, 2012; Gagliardone et al., 2015). Freedom of speech is a fundamental human right stated in the Universal Declaration of Human Rights and recognized in the International Human Rights Law. Many national legislatures protect freedom of speech, yet they recognize the need for restrictions in certain cases, particularly when it conflicts with other rights and freedoms, for example in cases of defamation or hate speech. Abusive content can inflict significant psychological harm to its victims and even lead to physical violence (Gagliardone et al., 2015; Gelber & McNamara, 2016). Members of marginalized groups can internalize the continuous message of their inferiority (Matsuda, 2018). Furthermore, negative stereotyping and dehumanizing language can lead to reduced pro-social behavior and increased anti-social behavior towards the victims, in extreme cases leading to atrocities and exploitation (Haslam & Loughnan, 2014).

This conflict can be viewed as two sides of the same coin: protection of equality and dignity is necessary to ensure that everybody has the right to free speech, and that the voices of minority groups and individuals are not silenced through threats and offensive behavior (Delgado & Stefancic, 1997; Citron & Norton, 2011; Shepherd et al., 2015). As Sen (2009) points out, justice needs to be thought as degrees of fairness to all participants. This should equally apply to offline and online spaces. The idea that the Internet is a place where any speech, no matter how offensive, is welcomed, clearly comes from the position of privilege. The ethical responsibility of society is to ensure safe environments where everybody can be heard.

Existing laws and government regulations as well as public movements put pressure on social media platforms, such as Twitter and Facebook, to provide intervention mechanisms in the form of filtering or simple appeal procedures. However, the sheer amount of online content prevents such companies to effectively deal with abusive messages. Further, content moderation infrastructures are governed by the powerful majority and can therefore reproduce the structural problems of colonialism, patriarchy and race (Thylstrup & Waseem, 2020). Social media platforms are often viewed as mere facilitators of the speech of others; in fact, they can be active political players and can influence individual and public opinion forming through their use of data and algorithms for content curation (Helberger, 2020). They can sell their power to persuade to advertisers, political parties, or governments, or use it themselves to influence public opinion on various issues, such as copyright law.[7] Giving the mostly unchecked power of content regulation to social media corporations may present even more danger to freedom of speech than any form of government intervention (Carlson, 2017).

Automation of content moderation can also reinforce social hierarchies and amplify social inequalities by limiting access to technology to certain groups, as researchers and companies implement abusive language detection algorithms for some languages and not others. In NLP research generally, the vast majority of studies focus on a small number of highly-resourced languages, leading to disparities in access to language technologies (Joshi et al., 2020). These disparities can lead to real-world harms. In 2019, Time magazine reported

[7] `https://www.bbc.com/news/world-australia-56163550`

that Facebook's hate speech detection algorithms worked for only 40 of the world's languages; many of those languages that were not included are spoken in developing countries where extremism and incitements to violence spread on social media can have devastating impacts.[8]

Early work on abusive language detection focused almost exclusively on English. In recent years, researchers have begun branching out to a growing number of languages (e.g., Mubarak et al., 2017; Wiegand et al., 2018b; Fersini et al., 2018; Bosco et al., 2018; Zampieri et al., 2020). However, in addition to linguistic differences across languages, it is important to note that notions of what is 'offensive' may be culturally-specific, presenting further challenges to creating datasets in multiple languages and applying knowledge transfer and multi-lingual approaches.

Online hate and abuse did not emerge from the online spaces. Rather, it reflects and possibly exaggerates marginalization and othering of minority groups happening offline. In other words, it is not so much a technological, but rather a cultural problem (Phillips, 2015). Anonymity, length limits, diminished feedback, minimal social clues, and excess attention, all contribute to a higher likelihood of heated conversations and offensive behavior in online communications than in face-to-face encounters (Friedman & Currall, 2003). But while technological, corporate policy, and legal interventions are necessary today, they can be made more effective in the long run if combined with a cultural shift.

3.2 Fairness and Non-Discrimination

The concept of fairness is one of the most fundamental moral values accepted across different cultures. However, as essential as it is, the interpretation of equality among individuals and groups can be subject to various ethical, social or religious views. Algorithmic decision-making can perpetuate social biases by discriminating against individuals because of their membership in certain social groups. As Ishida (2020) explains, what makes discrimination morally wrong is the harm to the discriminatees, who will be worse off than they would be were it not for the discrimination in question. The main objective of algorithmic fairness is to design systems whose outputs are equally accurate for all subsets of the population (Canetti et al., 2019), even though improvement of algorithmic fairness might come at a cost of lower overall accuracy on a particular test set (Martinez et al., 2019). The fairness and non-discrimination theme is strongly connected to promotion of human values, as fairness is one of the shared moral values across cultures. Also, assessment of algorithmic fairness supports human control of technology as users can appeal or opt out of the automatic decision making, if the process is not fair to them. It is also an efficient way of holding the designers and developers of automatic systems accountable, and therefore is connected to the accountability theme.

In the context of online abuse detection, several fairness issues have been raised. For example, Dixon et al. (2018) found higher rates of false positive errors when texts mention certain demographic groups. Sap et al. (2019) observed similar results when utterances include markers of African American English. Also, Blodgett et al. (2020) analyzed the concept of bias in NLP systems and described some of the unfair decisions that such systems might make while allocating resources to people or representing them in society. NLP

[8] `https://time.com/5739688/facebook-hate-speech-languages/`

researchers often tackle the fairness of automated systems by diagnosing and mitigating various biases in the system development pipeline. Shah et al. (2020) enumerated the potential origins of bias in NLP systems and provided a conceptual framework for measuring and mitigating this bias. We use this framework to review the current literature on bias in abusive language detection.

3.2.1 Semantic Bias

Embedding models are one of the sources of bias in natural language processing systems. An active line of work aims to quantify bias and stereotypes in language models as representations of text. Early works focused on gender and racial bias and introduced association tests for measuring bias in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Manzini et al., 2019). For contextualized word embeddings, May et al. (2019) and Kurita et al. (2019) used pre-defined sentence templates, whereas Nadeem et al. (2020) and Nangia et al. (2020) collected crowd-sourced sentences to measure stereotypical biases hidden in language models. Bartl et al. (2020) presented a template-based corpus to measure gender bias with respect to professions and showed that language models encode not only biases found in real-world data but also those based on stereotypes. They also showed that the techniques used to measure and mitigate bias that work for English language models might not be applicable to other languages. Besides encoding social biases, language models are also prone to generating racist, sexist, or otherwise toxic language, which hinders their safe deployment (Gehman et al., 2020). However, it is not entirely clear how the bias and toxicity present in language models impact the output of the trained classifiers. Jin et al. (2021) examined the bias in language models for the case where a hate speech classifier was trained via transfer learning, and demonstrated that upstream bias mitigation of language models is transferable to downstream tasks when models are trained through fine-tuning. They concluded that upstream bias mitigation is not as effective as direct bias mitigation on the downstream task, but the former is more efficient and accessible.

3.2.2 Selection Bias

Swamy et al. (2019) revealed that the dominance of benign examples in abusive language datasets, which is a common practice to emulate reality, might have a detrimental effect on the generalizability of classifiers. Also, sampling techniques deployed to boost the number of abusive examples may result in a skewed distribution of concepts and entities related to targeted identity groups. These unintended entity misrepresentations often translate into biased abuse detection systems.

Dixon et al. (2018) and Davidson et al. (2019) focused on the skewed representation of vocabulary related to racial demographics in the abusive part of the datasets and showed that adding counter-examples (benign sentences with the same vocabulary) would mitigate the bias to some extent. Park et al. (2018) measured gender bias in models trained on different abusive language datasets and suggested various mitigation techniques, such as debiasing an embedding model, proper augmentation of training datasets, and fine-tuning with additional data. Nejadgholi and Kiritchenko (2020) explored multiple types of selection bias and demonstrated that the ratio of offensive versus normal examples leads to a trade-off between False Positive and False Negative error rates. They concluded that this ratio is more important than the size of the training dataset for training effective classifiers. They also showed that the source of the data and the collection method can lead to topic bias and suggested that this bias can be mitigated through topic modeling.

Selection bias is one of the main challenges that limits generalization across datasets. Wiegand et al. (2019) showed that depending on the sampling method and the source platform, some datasets are mostly comprised of explicitly abusive texts while others mainly contain sub-types of implicit abusive language such as stereotypes. The study demonstrated that models trained on datasets with explicit abuse and less biased sampling perform well on other datasets with similar characteristics, whereas datasets with implicit abuse and biased sampling contain specific features usually not generalizable to other datasets. Razo and Kübler (2020) reproduced the results shown by Wiegand et al. (2019) across multiple datasets and showed that for generalizability, the differences in the textual source of datasets are more important than the sampling methods. Nejadgholi and Kiritchenko (2020) demonstrated the negative impact of platform-specific topics on the generalizability and showed that removing over-represented benign topics can improve the generalization across datasets.

Contrastive analysis of collected datasets is an effective way to mitigate the selection bias. Ousidhoum et al. (2020) conducted a comparative study on multilingual hate speech datasets to examine selection bias independent of the labeling schema. They proposed two metrics to evaluate this type of bias using the semantic similarity of topics included in datasets and the lexicon frequently used to search for hateful examples on social media.

3.2.3 Label Bias

Besides skewed data representations resulting from data sampling, the bias in annotations is another barrier to building fair and robust systems. NLP researchers have investigated two types of label bias in existing datasets: annotator bias and task formulation bias. As explained in Section 2.2, these biases originate from the subjectivity and ambiguity of the definitions of abusive behavior. A common practice to handle this subjectivity is labeling an instance through majority voting; however, this can serve to amplify the opinions of the majority and suppress minority voices (Blodgett et al., 2020).

Tversky and Kahneman (1974) were the first psychologists that showed how humans employ heuristics to make judgements under uncertainty. These heuristics are formed based on complex factors and lead to systematic personal biases, which are reflected in the annotations. Wilhelm and Joeckel (2019) studied the influence of social media users' characteristics on the evaluation of hate comments, focusing on abuse aimed towards women and sexual minorities. Their results indicate that moral judgments can be gendered. Breitfeller et al. (2019) used the degree of discrepancies in annotations between male and female annotators to surface nuanced microaggressions. Annotators' knowledge of different aspects of hateful behaviour can significantly impact the performance of trained classification models (Waseem, 2016). Similarly, annotators' insensitivity or unawareness of dialect can lead to biased annotations and amplify harms against racial minorities (Waseem et al., 2018; Sap et al., 2019). In such cases, re-annotating data while accounting for speaker identity and dialect may be a more effective strategy than employing automatic model debiasing techniques (Zhou et al., 2021). In another work on annotation bias, Al Kuwatly et al. (2020)

showed that the annotator's demographic features, such as first language, age and education, significantly impact the quality of annotations. Wich et al. (2020) identified annotator groups by using annotation behaviour characteristics, highlighting the significance of the annotator's behaviour in the quality of acquired annotations.

Furthermore, the ambiguities of task formulation create a specific type of label bias in this domain. In practical applications, the definitions of abusive language heavily rely on community norms and context and, therefore, are imprecise, application-dependent, and constantly evolving (Chandrasekharan et al., 2018). To make the task more tractable and focused, previous research has mostly concentrated on specific types of online abuse (e.g., hate speech, sexism, personal attacks), and the scope of studied abusive behaviors has been limited (Jurgens et al., 2019).

Several research groups studied the task formulation bias in a cross-dataset evaluation setting. Swamy et al. (2019) used a hierarchical annotation model to reveal overlaps and redundancies in the existing datasets. They pointed out that the current datasets are not representative of all facets of the included labels. Nejadgholi and Kiritchenko (2020) demonstrated that the definition of the Toxic class in the Wikipedia Detox dataset (Wulczyn et al., 2017) is very similar to the definition of the Abusive class in the dataset by Founta et al. (2018), but does not generalize well to other labels such as Sexism and Racism in the dataset by Waseem (2016). Cecillon et al. (2020) showed that a classifier trained for detecting the Toxicity class performs reasonably well when detecting Severe Toxicity and Personal Attack labels, revealing that the trained classifiers mostly learn the general definition of abuse. Fortuna et al. (2020) demonstrated that many different definitions are being used for equivalent concepts, which makes most publicly available datasets incompatible.

3.2.4 Bias in System Output

Even though the developers of datasets and models are cognizant of the risk of various biases, quantifying the extent of this risk is a challenging task. Dixon et al. (2018) proposed a way of measuring bias in trained models by building a synthetic dataset and using an evaluation metric that computes error disparity across identity groups. The Kaggle competition on the Unintended Bias in Toxicity Classification introduced a set of metrics that measure unintended bias for identity references across multiple dimensions. Huang et al. (2020) measured the differences in true positive/negative and false positive/negative rates for each demographic factor for classifiers trained on a multilingual hate speech corpus. Gencoglu (2020) also used error disparity across groups as a measure of fairness and defined fairness constraints to guide the training of a neural model. This definition of fairness was also integrated in the TensorFlow library as a regularization framework (Prost, 2020).

Other definitions and frameworks of fairness have been used for the evaluation of automatic abuse detection systems to encourage the development of systems that are optimized not only for the overall performance but also for fair outputs across different target groups (Borkan et al., 2019; Garg et al., 2019). For example, Davani et al. (2020) compared the average rates of true positives and true negatives for detecting hate speech targeted at different groups. By swapping the target token in utterances, they generated counterfactual examples and computed a relevant fairness metric, referred to as Counterfactual Token Fairness (CTF). They applied logit pairing to improve CTF and assured robust accuracy for similar sentences targeting different demographics. In another work, Dinan et al. (2020) decomposed gender bias in text along several pragmatic and semantic dimensions and proposed classifiers for controlling gender bias.

Ultimately, it is important to note that bias is an inevitable phenomenon in any statistical model. Fairness research focuses on identifying and mitigating biases that can potentially be harmful to certain sub-populations. While it is impossible to completely eliminate all biases, AI researchers and developers can work towards quantifying unfairness in system outputs for various demographics, identifying the origins of different types of biases, and designing techniques to minimize the harmful outcomes.

3.3 Transparency and Explainability

One of the greatest challenges in AI governance is the complexity and opacity of the current technology. Understanding the technology is a critical step for the realization of other principles, such as accountability, human control of technology, safety and security, and fairness and non-discrimination. Transparent machine learning intends to shed light on the process of creating an automatic system and make it understandable by different stakeholders. As Weller (2019) clarified, transparency can refer to various practices depending on who the audience and the beneficiaries of the explanations are. Interpretability and explainability, often used interchangeably, are the two terms closely tied to the transparency of machine learning models. Interpretability mostly describes the methods that explain the underlying dynamics of opaque algorithms, such as deep neural networks. On the other hand, explainability usually refers to a set of post hoc added explanations for an existing model, understandable by lay users (Marcinkevičs & Vogt, 2020).

3.3.1 Transparency

As Felzmann et al. (2020) explained, transparency refers to multiple normative concepts that should be considered in the ecosystem of automated decision making. However, translating these concepts to a set of practical steps is a challenging task. In the field of NLP research, Mitchell et al. (2019) introduced the concept of model cards as a means to address transparency of deep learning models. They urged the developers of models to report the details of the data on which the models were trained and clarify the scope of use, including the applications where the employment of the model is not recommended. As an example, they presented a model card for the Perspective API system. IBM has proposed a similar concept, called FactSheets, for AI service providers to document the purpose, performance, safety, security, and provenance information on their products (Arnold et al., 2019). Further, data statements (Bender & Friedman, 2018) and datasheets for datasets (Gebru et al., 2018) were designed to standardize the process of documenting datasets. Bender (2019) explained how transparent documentation can help in mitigating the ethical risks.

Generally, the practical definition of transparency depends heavily on the circumstances of real-world deployment and is an essential criterion to earn the trust of the users. On the technology development side, transparency is tied to explainability and high-quality documentation, but may cause harm if not compatible with the principles of privacy (Weller, 2019).

3.3.2 Interpretability and Explainability

As the impact of AI becomes more significant in our daily lives, developers of automatic systems are expected to earn the trust of users by providing explanations for automatically-made decisions. Traditional lexicon- and feature-based models are interpretable to some extent as they use features understandable by humans. In feature-based systems, bag-of-words and character n-grams have been most frequently used, but some other explainable features, such as the ones derived from sentiment analysis, tone analysis, subjectivity, and topic modelling, have also been employed (Fortuna & Nunes, 2018). However, the accuracy of lexicon- and feature-based systems is often significantly lower than the accuracy of deep learning models (Dixon et al., 2018; Gunasekara & Nejadgholi, 2018; Founta et al., 2019).

Neural networks, on the other hand, are effectively black boxes. Recent research has leveraged the LIME (locally interpretable model-agnostic explanations) algorithm in an attempt to interpret a model's representation of abusive statements (Srivastava & Khurana, 2019; Mahajan et al., 2020). LIME's explanations consist of words highly-weighted by the model, but no further information is provided on why a text is classified as abusive (Ribeiro et al., 2016). Wang (2018) used partial occlusion to discover the keywords that are most predictive of hate speech, revealing some of the peculiarities and biases of this problem space.

Similarly, attention mechanisms embedded in deep learning architectures were used to identify the abusive parts of a text (Chakrabarty et al., 2019). However, it is not clear if such mechanisms provide meaningful explanations of a model's decisions (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Grimsley et al., 2020).

Output probability (or confidence) scores produced by classifiers have been used to explain the severity of abuse (Hosseini et al., 2017; Gröndahl et al., 2018). However, it is not fully clear how well these probabilities correspond to the human perception of severity and in what ways they might be affected by sampling methods. Recently, Vidgen et al. (2020b) showed that the output scores of classifiers can be re-calibrated to better align with human evaluations. Also, Perspective API calibrates the output scores of its classifier to convert them to approximate probabilities. The final probabilities are interpreted as the percentage of people that would consider the comment to be toxic (PerspectiveAPI, 2017).

One approach to explainability is through more comprehensive data annotation so that more particulars can be learned directly from training data. For example, models trained on the Kaggle Toxicity dataset labelled for five sub-categories of toxicity can provide more information than the models trained on the previous versions of this dataset annotated with binary labels (Wulczyn et al., 2017). Another example is the OffensEval dataset that includes annotations for the target of abuse (individual, group, or other) (Zampieri et al., 2019b). Sap et al. (2020) employed modern large-scale language models in an attempt to automatically generate explanations as social bias inferences for abusive social media posts that target members of identity groups. They asked human annotators to provide free-text statements that describe the targeted identity group and the implied meaning of the post in the form of simple patterns (e.g., "women are ADJ", "gay men VBP"). This work showed that while the current models can accurately predict whether the online post is offensive or not, they struggle to effectively reproduce human-written statements for implied meaning.

3.4 Privacy

Privacy is an important theme in any discussion of ethical AI systems. In this context, privacy encompasses both the use of user data to train machine learning models for online abuse detection, and individual users' agency to control their personal data.

In research, one area of concern is the creation and distribution of datasets for the purpose of benchmarking abusive language detection systems. While it is scientifically valuable to compare systems using identical training and test sets, this may conflict with the user's right to privacy. For example, most of the abusive language datasets collected from Twitter or reddit involved scraping publicly available data from those platforms without the explicit consent of the users that their data be used for this purpose (Pitropakis et al., 2020). Furthermore, in the process of developing abusive language classifiers, researchers may infer personal information about the users that the users did not intentionally share, such as gender and location (Waseem & Hovy, 2016; Unsvåg & Gambäck, 2018). One aspect of privacy that is gaining increased attention is *erasure*, or the "right to be forgotten". The research community has been moving towards protecting this right by distributing datasets as a set of post IDs, for example, rather than the complete texts. This way, if users delete the post or their entire profile, the next researcher to download the corpus will not be able to access those texts. Thus, the theme of privacy is clearly connected with professional responsibility, as practitioners are responsible for the collection, usage, and storage of personal data.

A different set of privacy issues emerges when we consider the commercial deployment of an abusive language filter. In contrast to research studies, which typically involve relatively small convenience samples of public data, a large-scale system in deployment would require access to the personal data of all users on a given platform. However, this privacy issue is not specific to abusive language detection, and numerous solutions have been proposed, including federated learning (Konečnỳ et al., 2016; Yang et al., 2019) and edge computing (Shi et al., 2016). The basic principle behind these techniques is to avoid sending user information to the cloud (i.e., an external server) for processing. Instead, training *data* for the model remains on each individual's device, and only the model *parameters* are stored in the cloud, sent to the device, updated on the user's data, and sent back to the cloud using encrypted communication. Another approach to protecting user privacy is Secure Multi-Party Computation, which was used by Reich et al. (2019) to demonstrate an example of privacy-preserving hate speech detection in personal text messages.

Another issue related to user privacy and online abuse is the practice of "doxxing", or publishing private information about a person online (such as their home address), typically in order to subject that person to harassment. In this sense, privacy is closely related to safety and security. Very little research so far has considered the automated detection of such behaviour. Jurgens et al. (2019) argue that the NLP community has thus far focused on a too-narrow definition of online abuse, and should branch out to both more subtle forms of abuse, such as microaggressions, and more severe threats to safety and personal privacy, such as doxxing.

3.5 Safety and Security

In sensitive applications, where critical decisions are made, safety and security challenges are key obstacles to the wide-scale adoption of emerging technologies (Darvish Rouani et al., 2019). Safety and security measures are crucial elements for building reliable systems: systems that are safe, in that they perform as intended, and also secure, in that they are not vulnerable to being compromised by unauthorized third parties (Fjeld et al., 2020). Ensuring the safety and security of AI systems is one of the crucial pillars of the accountability and professional responsibility for systems' designers and developers. Further, users need to be able to take control of the technology when safety and security are at risk.

Previous work defines safety in machine learning systems as minimizing the possibility and probability of expected harms (Varshney, 2016; Varshney, 2020). Saria and Subbaswamy (2019) identified three principles of reliability engineering to ensure the safety of developed systems: failure prevention, failure identification and reliability monitoring, and fixing or addressing the failures when they occur. In the context of online abuse detection, especially for cases when the end goal is content removal, both false positive and false negative errors can have significant consequences for users as one threatens the freedom of speech and affects people's reputations, and the other ignores hurtful behaviour. The risk of deploying an automatic system depends heavily on the practical circumstances of the application in hand. To minimize the safety risks, models need to be systematically tested on a variety of inputs and language phenomena. Towards this goal, Röttger et al. (2021) proposed HateCheck, a suite of functional tests to identify weaknesses of a hate speech detection model in handling various expressions of hate and contrastive non-hateful utterances. Another safety risk for automatic abuse detection is the mismatch between the training and test environments. To minimize this risk, the trained systems have to be maintained and retrained as the language of the users evolves over time.

In real-world scenarios, a system is expected to function accurately not only in the presence of regular users, but also in the presence of adversaries and malicious users that might try to deceive the system. Several studies have shown that trained abuse detection systems can be deceived or attacked by malicious users. Hosseini et al. (2017) demonstrated that an adversary can query the system multiple times and find a way to subtly modify an abusive phrase resulting in a significantly lower confidence that the phrase is abusive. Gröndahl et al. (2018) showed that adding a positive word such as *'love'* to an abusive comment can flip the model's predictions. They studied seven models trained for hate speech detection and concluded that although character-based models are more resistant to attacks, model variety is less important than the type of training data and labelling criteria. They also found that all detection techniques are brittle against adversaries who can (automatically) insert typos, change word boundaries, or add innocuous words to the original hate speech. One common recommendation to tackle such vulnerabilities is to use sub-word information instead of word-based features. However, Kurita et al. (2020) observed that in spite of rich sub-word representations, a BERT-based classifier can be deceived by inserting a specific rare word into an abusive sentence. Kalin et al. (2020) introduced a method for identifying vulnerabilities of the system after it was deployed. They implemented this technique for Perspective API and showed that simple attacks such as vowel substitution and duplication lead to significant reduction in the toxicity score for toxic comments, changing the prediction to non-toxic. They also developed a framework for securing models against such attacks. Mou et al. (2020) proposed to improve the robustness of a hate speech classifier against adversarial attacks by using subword information, word-level semantics, and the significance of words calculated by an attention mechanism.

3.6 Accountability

Accountability refers to the concerns about who is accountable for automatically made decisions as well as the potential impacts of the technology on the social and natural world (Fjeld et al., 2020). It includes the issues of verifiability and replicability, impact assessment, evaluation and auditing requirements, ability to appeal, and liability and legal responsibility. Further, the principle of accountability is strongly connected to safety and security, transparency and explainability, and human control of technology.

It is generally agreed that the organizations that develop and deploy AI systems should be responsible for the systems' outcomes and impacts. AI systems themselves are not legal agents, and, making them legally accountable may be unnecessary and troublesome (Bryson et al., 2017). Some ethical AI guidelines distinguish between the liability of the developers of an AI system and the liability of the organizations that deploy the system. At the level of development, the most appropriate measures of accountability are typically considered to be transparency and codes of professional responsibility.

Audit for ethical compliance, both internal or external, is another requirement for accountability at both levels of development and deployment. Some ethical AI principles documents, such as the Toronto Declaration, assert the necessity of an external (third-party) audit for systems that have a significant risk of human rights violation. Along with a technical component, audit can include an institutional component to verify institutional practices in order to prevent improper use and negative impacts on society. External audit assesses the risk that a system will cause harm to individuals or society from outside the system, and tends to be conducted after deployment, when some harms can already be done (Green & Chen, 2019). Raji et al. (2020) proposed a framework for internal algorithmic auditing that supports the system development end-to-end through the full development and deployment life-cycle. Their framework includes five distinct stages: Scoping, Mapping, Artifact Collection, Testing and Reflection (SMACTR). At each stage, a set of audit documents is produced, that together form an overall audit report that assesses the fit of design and implementation decisions within the organization's values and ethical guidelines.

Currently, 'online information intermediaries' or, in other words, social media platforms, are almost solely responsible for their own content moderation. They decide which posts to remove or downrank, and which user accounts to suspend or delete, based on alleged violations of the platform's policies and terms of use. In some jurisdictions, the social media platform corporations are legally responsible for removal of 'dangerous' content, such as incitement to violence or expression of hatred directed against a protected group. However, many users, civil society organizations, and policy makers consider common content moderation practices ineffective and often detrimental (York & McSherry, 2019). They argue that content moderation provided by social media platforms, such as Twitter and Facebook, is inconsistent and confusing, the appeal system is inadequate, and transparency is minimal.

Several documented cases showed that such content moderation can cause harm to users and businesses by unnecessary restriction of posts with certain words and imagery.[9]

The social media corporations have little accountability for either automatic or human decisions regarding content moderation. Given the unprecedented impact of social media corporations on the public sphere, new mechanisms for platform governance are required (Leonard, 2019). A set of minimum standards for content moderation, the Santa Clara Principles on Transparency and Accountability in Content Moderation, has been proposed by organizations and academics.[10] They call for better transparency to the public about the processes (including automatic decision making) and results of content moderation, meaningful opportunities for users to appeal any content or account suspension or removal, and justification for any content removal decisions. Other mechanisms, for example, evaluation and engaging of policymakers in social media platforms' rule-making activities or 'procedural accountability', can also be effective (Bunting, 2018). Further, new institutions, such as social media councils and e-courts, can be established to discuss terms of use, adjudication processes, and fundamental ethical questions (Tworek & Tenove, 2020).

3.7 Human Control of Technology

Human control of technology refers to the ability of users to appeal automated decisions and request human review, or even opt out of automated decisions entirely. Given the ambiguity of language and the need to protect freedom of speech, this is an important ethical principle in relation to abusive language detection and moderation. As Duarte et al. (2018) point out, many research studies report accuracies in the 80% range – this means that 1 in 5 automated decisions will be "incorrect" (and even "correct" decisions may be disputable due to the highly subjective nature of the task). This necessitates human review of uncertain or contested decisions. Human participation in the decision making can be viewed as one way to achieve various other objectives, such as safety and security, transparency and explainability, fairness and non-discrimination, and promotion of human values.

Most of the research studies in this area focus simply on the detection of abusive language, without stating explicitly what can or should be done with that content once detected, or how to deal with false positives (innocuous posts inadvertently flagged as abusive) and false negatives (harmful language that is not detected). To some extent, these decisions will vary depending on the platform and the communities it serves (Gorwa et al., 2020). However, many online platforms have determined a need for human moderators in addition to algorithmic toxicity detection (Cecillon et al., 2019). For example, despite a recent move to increase the amount of automated content moderation, Facebook still employs 15,000 content moderators, particularly to ensure the most violent or disturbing material does not reach the public eye.[11] Additionally, users still have the option to appeal the decision to have their content removed, and have a human review the decision.

This example illustrates two different locations "in the loop" where human review can be deployed in an automated system: first, the automated system can flag potentially

[9] https://onlinecensorship.org/content/infographics

[10] https://santaclaraprinciples.org

[11] https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus/

problematic content (or cases in which it has low confidence in its prediction) and send it for human review before posting (pre-moderation). This process has the benefit that harmful content will not be exposed to the public, but it can result in a long lag time before posts are published, leading to frustration and disruption to the conversation. Second, the initial moderation can be fully automatic, but users can request that the automated decision be reviewed by a human decision-maker (post-moderation). For users to be able to challenge the system's outputs, the outcomes have to be presented in an easy-to-understand form with information on the factors and logic that influenced the decision. Furthermore, in either pre- or post-moderation, ideally the feedback from the human moderator can be used to improve automated classifier decisions in the future.

3.8 Professional Responsibility

As NLP researchers, we have a professional responsibility to act ethically. In the Harvard AI Ethics Framework, this responsibility encompasses tenets such as ensuring the accuracy of the systems we build, adopting principles of responsible design, considering the long-term effects of our work, engaging stakeholders who may be affected by our systems, and upholding scientific integrity. These principles are reinforced by the Professional Code of Ethics of the Association for Computing Machinery (ACM),[12] Institute of Electrical and Electronics Engineers (IEEE),[13] and Association of Internet Researchers (AoIR),[14] among others. The Association for Computational Linguists (ACL) has adopted the ACM Code of Ethics, which begins by stating, "Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good."

Blodgett et al. (2020) put forward several questions for NLP researchers and developers to consider throughout the software development cycle, to help situate the work within a broader societal and historical context, and uncover the implicit assumptions and normative values being reinforced. In the context of abusive language detection, these might include questions around: which groups will be affected by this system, and how? Will detecting abusive language dismantle social hierarchies and language ideologies, or serve to reinforce them? Have we engaged with members of the groups we hope to help with such systems, to ensure this is addressing their needs in an acceptable manner? Are we solving real problems, or just the ones that are convenient to solve with the methods or data we have at hand? What are the possible future applications of such a system – could it be used to silence political dissidents speaking out against their government? Or marginalized groups discussing their own lived experiences? Who decides what constitutes *offence* or *hate*?

While these are tough questions and may not have easy, one-size-fits-all answers, awareness has been building that we cannot proceed blindly with our research in the "ivory tower" of academia, without taking the time to become informed about society, and to critically assess the potential impact of our technology on a global scale. In particular, as Fjeld et al. (2020) emphasize, developments in AI have thus far out-paced the ability of governments to implement legal and regulatory frameworks for AI governance, placing increased respon-

[12] https://www.acm.org/code-of-ethics

[13] https://www.ieee.org/about/corporate/governance/p7-8.html

[14] https://aoir.org/ethics/

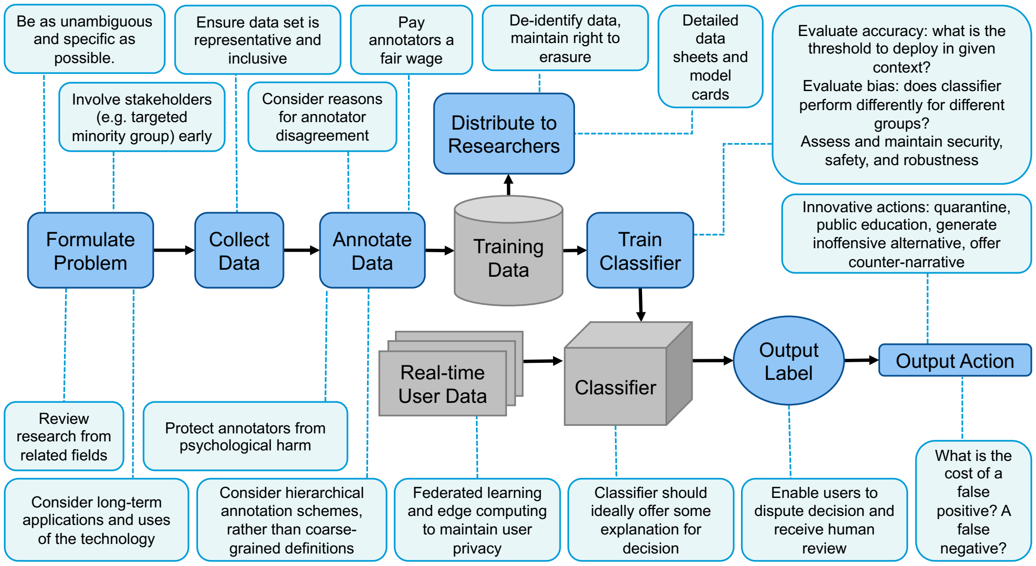
Be as unambiguous and specific as possible.
Ensure data set is representative and inclusive
Pay annotators a fair wage
De-identify data, maintain right to erasure
Detailed data sheets and model cards
Evaluate accuracy: what is the threshold to deploy in given context?
Evaluate bias: does classifier perform differently for different groups?
Assess and maintain security, safety, and robustness
Involve stakeholders (e.g. targeted minority group) early
Consider reasons for annotator disagreement
Distribute to Researchers
Formulate Problem
Collect Data
Annotate Data
Training Data
Train Classifier
Innovative actions: quarantine, public education, generate inoffensive alternative, offer counter-narrative
Real-time User Data
Classifier
Output Label
Output Action
Review research from related fields
Protect annotators from psychological harm
What is the cost of a false positive? A false negative?
Consider long-term applications and uses of the technology
Consider hierarchical annotation schemes, rather than coarse-grained definitions
Federated learning and edge computing to maintain user privacy
Classifier should ideally offer some explanation for decision
Enable users to dispute decision and receive human review

Figure 1: A high-level overview of some of the ways ethical considerations can be incorporated throughout the pipeline of abusive language detection.

sibility on practitioners to positively influence the values and ethics in our field. This can include taking actions such as: choosing not to work on projects that do not support the social good or that have the potential for long-term harm, engaging with stakeholders and taking their feedback seriously, and being open and transparent about the limitations and failures of our technology, including publishing negative results.

4. Ways to Move Forward

In this section, we outline several emerging research themes where the AI community can contribute to developing ethics-aware technologies to tackle online abuse. We start with challenging the common task formulation as a binary or multi-class classification problem and discuss alternative ways of mitigating the negative impacts of abusive behavior. We highlight some promising research directions to effectively confront online abuse, educate the public on evolving social norms and the potential harms of abusive behavior, and make AI systems and their outputs transparent and intelligible to various stakeholders. We stress that online abuse is a social problem and urge AI researchers to ground their work in theories and findings from other disciplines, like social sciences, communication studies, and psychology. Finally, engaging the affected communities in technology design and development is critical to produce robust, fair, and trustable systems. Figure 1 shows some of the ways that these and other, previously mentioned ethical considerations can be incorporated at different stages of the design, development, and deployment of AI solutions.

4.1 Reimagining the Task of Confronting Abusive Language Online

As we showed in the previous section, most of the NLP work dealing with online abuse views the task as a binary classification problem: determine whether a social media post is abusive or not. The definition of the 'abusive' class varies from project to project, and sometimes multiple categories of 'abusive' are included. However, it becomes increasingly evident that abusive language is much more nuanced and such a task formulation significantly limits the applicability of the developed technology in real life.

4.1.1 Moving Away from Coarse-Grained Definitions of Abuse

Online abusive content embodies a spectrum of practices that differ in motivation, expression, and consequences, and needs to be examined within regional and historical context (Shepherd et al., 2015; Pohjonen & Udupa, 2017; Thylstrup & Waseem, 2020). This defies easy binary division into content that is acceptable and content that is not. Clearly, there is content that is explicit, severely offensive, can lead to violence, or prohibited under the law, and needs to be removed from public view. Current NLP technologies have been successful at recognizing explicitly abusive texts and can be of help here bringing such content to the attention of human moderators to ensure fast and effective management. However, some abusive content is implicit, and at times even produced unintentionally or with little conscious awareness of its impact. Still, such messages can cause social and psychological harms, especially when accumulated over a lifetime. Implicit abuse is very hard to detect automatically, and even when detected, it is often not removed as it does not directly violate platforms' terms of use or any legal codes. Different mechanisms for dealing with such content are required.

Recently, the complexity of the task formulation has started to be recognized by the research community, and as a first step, several studies have proposed multi-dimensional, multi-level frameworks to address the task. Waseem et al. (2017) mapped the different types of abuse to two dimensions: (1) whether the abuse is directed towards an individual or a generalized group, and (2) whether the language is explicit or implicit. Vidgen et al. (2019) extended this topology to three dimensions: (1) whether the abuse is directed towards an individual, an identity (based on belonging to a demographic category, social group, or organization), or concept (such as a belief system, country or ideology), (2) who receives the abuse (e.g., which identity group, moderators vs. content producers, friends vs. strangers), and (3) how abuse is articulated (e.g., aggression, insults, stereotyping; explicit vs. implicit). Kiritchenko and Nejadgholi (2020) suggested a two-dimensional multi-level classification structure that includes a hierarchical schema for subject matter of an utterance (or target of abuse) and fine-grained severity of abuse explicitly annotated through comparative methods. Sap et al. (2020) framed offensive language detection as a hierarchical task that combines structured classification with reasoning on social implications. Their classification schema includes offensiveness, intention to offend, lewdness, target group, in-group language, and implied meaning of the utterance. Assimakopoulos et al. (2020) formulated hate speech as hierarchical and multi-layer inferences on sentiment, target, expression of abuse, and violence. Overall, this recent shift towards multi-dimensional, hierarchical schemas allows for a more nuanced, fine-grained representation of abusive language, which is better able to handle the complexity of real-life data.

4.1.2 Moving Towards Flexible, Rights-Respecting Moderation

The current practices of dealing with abusive content by major social media platforms are also often binary: posts that are deemed to violate a platform's terms of use are permanently removed; all the other posts are shown to users. Such black-and-white decisions can lead to power abuse by the social media companies, restricting users' rights to freedom of speech and causing harm to individuals and businesses. Several alternative mechanisms have been proposed in the literature that provide a middle ground between permanent removal of some content and no content moderation at all.

Quarantining of potentially abusive posts is one such flexible solution (Ullmann & Tomalin, 2020). Posts that have been automatically identified as potentially abusive can be temporarily quarantined, and an alert would be sent to both sender and direct recipients warning them of potentially harmful content. Then, the recipients could decide if they want to see the message and if they want the message to be posted. Similarly, in the case of a public message, each user can decide for themselves if they want to see the message. This can protect vulnerable populations from harmful messages while not invading the sender's freedom of speech. To ease users' decisions, messages flagged as potentially harmful can come with related information on estimated severity of abuse, confidence of the automatic classifier, the sender's name and history of abusive behavior, etc.

Automatic content moderation can be deployed at the point of a message creation at the user's side. Yet, instead of banning abusive posts, techniques such as *nudging* and *value sensitive design* can be used to alter users' behavior (Vincent & Jane, 2017). Based on research in the social sciences and the psychology of human behavior, communication tools can be designed in such a way that default actions would result in desirable, socially acceptable interactions, while socially unacceptable behavior would still be possible, but require extra effort from the users. Such communication environments would not eliminate online abuse completely, but would discourage anti-social interactions, and, hopefully, significantly reduce their occurrence. For example, in the popular online multi-player game "League of Legends" the introduction of the requirement to explicitly opt-in to online chat between opposing teams significantly reduced the number of negative (and often abusive) conversations and increased the number of positive exchanges.[15] Alternatively, the communication tool can make it more difficult to post explicitly abusive messages by alerting the user that their message contains offensive content and asking for confirmation of their intentions. Technology can also monitor the user's emotions (using verbal and non-verbal cues) and can be set up in a way that it blocks sending any online messages for a period of time if the user feels angry, frustrated, or annoyed. Such settings would be controlled by the user and can prevent them from doing harm in the heat of the moment.

Another interesting direction is style transfer in texts. The goal here is to automatically translate an abusive text into a non-abusive one while preserving as much (non-offensive) meaning as possible (Nogueira dos Santos et al., 2018; Tran et al., 2020). This technology can be useful for mitigating harmful outputs produced by bots as well as for proposing alternative, non-offensive rephrasings to human messages. However, care must be taken not to advertently or inadvertently manipulate users' views or introduce further biases.

[15] `https://www.polygon.com/2012/10/17/3515178/the-league-of-legends-team-of-scientists-trying-to-cure-toxic`

4.1.3 Extending the Task Beyond Detection

In addition to preventing or at least making it harder for users to post abusive content, other mechanisms of mitigating the harmful effects of online abuse have been proposed. Counter-narrative (or counterspeech) can be very effective in addressing abusive behavior at the societal level (Benesch et al., 2016; Schieb & Preuss, 2016; Lepoutre, 2017). Counter-narrative is a non-aggressive response to abusive content that aims to deconstruct the referenced stereotypes and misinformation with thoughtful reasoning and fact-bound arguments. It does not impinge on freedom of speech, but instead intends to delegitimize harmful beliefs and attitudes. It has been shown that counterspeech, for example, can be more effective in fighting online extremism than deleting such content (Saltman & Russell, 2014). Defining, monitoring, and removing extremist content, as well as any abusive content, is challenging. Furthermore, after the content has been deleted, it can simply be re-posted on another online platform. Counterspeech is intended to influence individuals or groups, and can be spontaneous or organized. Social media campaigns (e.g., #MeToo, #YesAllWomen, #BlackLivesMatter) can effectively raise public awareness and educate on issues, such as misogyny and racism online. Spontaneous counterspeech, produced by victims or bystanders in response to online harassment, can also be successful and result in offenders recanting and apologizing (Benesch et al., 2016). Computational techniques, such as natural language generation and automatic fact-checking, can assist expert and amateur counter-narrative writers in creating appropriate responses (Qian et al., 2019; Tekiroğlu et al., 2020). Similarly, NLP technology can be used to assist users with other types of positive engagements, such as offering support to victims of online abuse.

AI technology can also be put to use to track the spread of information over social networks and predict the virality of social media posts (Jenders et al., 2013; Samuel et al., 2019). Abusive posts that go viral are arguably more dangerous and can lead to public unrest and atrocities offline. Therefore, automatically detecting posts approaching virality or having significant potential to go viral and checking their abusiveness with human moderators can help prevent the spread of the most dangerous content over the network and reduce its potential harm (Young et al., 2018).

There is an ongoing effort to educate public on the issues of diversity, inclusion, and discrimination, especially in school and workplace environments. Often abusive and discriminating behavior occurs without the realization of its potential harms on the victim, so educating the public on these issues is an important step towards the needed cultural shift. For example in the case of the League of Legends, users suspended for abusive behavior were provided explanations of which of their actions led to their temporary banning from the game and why. Many users acknowledged that such explanations helped them become aware of the potential impacts of their actions and positively affected their future behavior.[16]

Online projects, such as "Microaggressions: Power, Privilege, and Everyday Life"[17] and HeartMob[18], are examples of resources built to inform and educate the general public on issues of online and offline abuse, including its subtle and implicit forms. These platforms

[16] https://www.spectrumlabsai.com/the-blog/how-riot-games-is-used-behavior-science-to-curb-league-of-legends-toxicity

[17] https://www.microaggressions.com/

[18] https://iheartmob.org/

allow users to report their experiences of severe and subtle forms of harassment in everyday life and online. Surfacing such abusive interactions helps victims to validate their experiences, motivate bystanders to provide support, and eventually establish community norms on appropriate online and offline behavior (Blackwell et al., 2017).

Public education on the issues of abusive and hurtful online behavior, especially its subtle and inadvertent forms, can be seamlessly integrated into an everyday workflow. Personalized applications can be developed that would alert users if their written communications can be interpreted as offensive or disrespectful in specific settings (e.g., work environment). Such a system could watch for the tendency of a user to refer to stereotypical or otherwise negative portrayals of certain identity groups, or condescending behavior. Further, detailed explanations on why the utterance can be interpreted as offensive or unfriendly along with counter-narrative to challenge the user's stereotypical views can have a significant positive impact. Another AI research direction is to empirically study the characteristics of conflict discussions and predict the point where a conversation is likely to derail to negative, unproductive exchanges (Zhang et al., 2018a; Marcinowski & Ławrynowicz, 2020). These kinds of applications can be highly personalized and tunable for specific situations (friendly conversation, official business communication, etc.). However, such technology can raise privacy concerns that should be thoroughly assessed and adequately addressed before the deployment.

Redesigning the task of online abuse detection to take into account the complexity of the phenomenon and investigate alternative approaches to mitigating its harmful effects contributes to addressing several of the ethical issues: promotion of human values by balancing the promotion of the human right to free speech and protection of vulnerable populations, human control of technology by transferring the decision power to affected users, and fairness and non-discrimination by reducing the negative outcomes of censorship on marginalized communities.

4.2 Advancing Interpretability and Explainability

Interpretability and explainability are critical elements in addressing several ethical principles, including human control of technology, accountability, fairness, transparency, and safety and security. Different stakeholders, including designers and developers of the systems, data scientists, regulators, and end users, can benefit from explainability for reaching a number of objectives, such as debugging the system, validating its fairness, safety and security, or appealing the system's decision. However, these different stakeholders and their different objectives require divergent, tailored solutions (Vaughan & Wallach, 2020).

One of the main barriers to ensuring fairness, safety, and security of automatic systems is that even creators of modern algorithms do not necessarily understand their working mechanisms. Interpretability can be thought of as a diagnostic tool to empower machine learning researchers and practitioners to detect and quantify biases and other vulnerabilities in automatic systems. For example, interpretability methods can be used to link back the unfair behavior of system output to data imbalances and can guide towards a more representative data collection (Dixon et al., 2018). As another example, interpretability tools can be used by developers to identify the risk of learning spurious correlations (Cheng et al., 2019). Through preventing and mitigating this risk the classifiers will be more robust and generalizable to out-of-distribution data examples.

Also, the development of explainable systems is a way to earn users' trust by providing relevant explanations on why the system made a particular decision. The most accessible explanations are the ones that articulate the reasons behind predictions in plain language (Sap et al., 2020). However, the current state of the language technology cannot reliably generate such articulations of reasons. At the present time, the most achievable sense of explanation is to provide the user with a reliability score of the model's predictions (Miok et al., 2020; Vidgen et al., 2020b). It is the responsibility of the creators of machine learning systems to design accessible interfaces for the developers and lay users. Through such interfaces the users will be able to receive and correctly interpret the various forms of explanations, such as scores or visualizations.

4.3 Grounding Research in Work from Other Disciplines

The problem of online abuse cannot be solved by AI technology alone as, ultimately, online abuse is a social problem that can be either amplified or mitigated with the help of technology. Instead of only focusing on improving predictive model performance, AI researchers should also work together with social scientists, anthropologists, psychologists, criminologists, human rights activists, and ethicists to understand abusive online behavior, its motivations and expressions, and how it is propagated through social networks, and to design communication technologies that encourage ethical behavior and discourage unethical behavior (Prabhakaran et al., 2020).

Social sciences, communication sciences, psychology, and anthropology have been studying online communication mechanisms, psychological principles involved in online communication, and conflict theories to understand the motivations and various processes involved in abusive behavior. For example, the communication studies suggest that online communication requires a communicant to be "heard" before they can communicate (Shepherd et al., 2015). Therefore, users struggle to exist as communicating subjects and use extreme statements to draw attention. However, once the online identity is built, it needs to be maintained to be able to be recognized by others and to continue to exist online.

Another contributing factor is group identity. According to Social Identity Theory, the social groups with which people identify themselves are important for their positive self-concept (Tajfel & Turner, 1979). To enhance their self-image, members of an in-group often seek to find negative aspects of an out-group. When two groups hold different views on an issue, the assumption of situational differences (people having different life experiences) is quickly replaced with the assumption of dispositional differences (people being selfish and biased), which leads to attacking individuals' characteristics rather than their arguments. Further, current social media platforms, such as Twitter, have been designed for instant information sharing with no barriers to communication. As a result, such platforms foster competition rather than cooperation (Conbere, 2019). Online communication lacks "grounding", a process by which two parties achieve a shared sense of participation in a conversation (Friedman & Currall, 2003). Conversations on social media are often asynchronous, have participants unaware of each other's environments, and lack visual and audio cues. Without grounding, it is hard for participants to understand and connect with each other. Anonymity, which is a common characteristic of online exchanges, exacerbates these difficulties. According to the Social Identity Model of Deinviduation Effects (SIDE), anonymity changes the relative salience of personal vs. social identity, therefore sometimes facilitating anti-normative behavior (Reicher et al., 1995).

Geographical and temporal context is also very important. Social and political environments, specific contemporary events, direct vs. indirect communication, and the identity of the speaker significantly influence the dynamics and impacts of online abuse (Shepherd et al., 2015). However, these factors are rarely taken into account when designing AI tools.

Engaging scholars and scholarly works from different disciplines in collective efforts to develop socio-technical solutions would help AI experts to exercise professional responsibility, and address the issues of fairness and discrimination, and safety and security in a more informed way. It would also contribute to developing more comprehensive explanations for automatic or semi-automatic decisions. Further, civil society and human rights activists can inform AI design on the issues of human rights and promotion of human values.

4.4 Engaging Affected Communities

Most often, the targets of online abuse are minorities and marginalized communities. On the other hand, the development and deployment decisions for technological solutions are usually in the hands of the powerful majority. Thus, technology continually reinforces the structural power relations in society. To shift this power imbalance, technological solutions should be centered around the lived experiences and the needs of the victims of online abuse (Blodgett et al., 2020). For example, Arora et al. (2020) designed a system to protect women journalists from online harassment by interviewing representatives from this target group about their encounters of online harassment on Twitter and directly engaging them in the data collection and annotation process. Involvement of the affected communities in the design decisions, including the decision of whether to build a particular system at all, would help better position the task in the social and political context, account for its many nuances, analyze possible consequences of the system's deployment, and identify and mitigate potential ethical issues (Blackwell et al., 2017; Katell et al., 2020).

When it comes to effective and inclusive community engagement, there is a lot to learn from other research fields, such as health sciences. Successful community engagements are built around a solidaristic relationship between researchers and the community members, which creates mutual understanding and empathy (Pratt et al., 2020). Partnerships that advocate for equity for all and build trust receive the most effective responses (Alberti et al., 2020). Also, besides inclusive data representation, it is essential to include the voices of disadvantaged groups in setting the priorities of the research agenda (Pratt, 2019). Another way forward to productive community engagement is committing to a fair and transparent compensation instead of limiting the engagement to low income or volunteer work.

5. Conclusion

In this survey, we have attempted to bring together all the various sub-fields of abusive language detection and examine the field through the lens of ethics and human rights. We expect that researchers working to protect people from hate speech, cyber-bullying, racism, sexism, and other forms of online abuse are already motivated to make the world a better place, and our goal is not to diminish the progress that has been made thus far — rather, we identify several future directions for research and critical thinking. Much of the research effort has focused on the language processing and machine learning components of the pipeline; however, for these components to be truly applicable in the real world, it is also necessary to take a step back and look at the bigger picture. Focusing on the input: how has the problem been formulated, and by whom? Have the communities who are being affected been consulted? Where is the data coming from, and is it representative? Who is annotating the data, and what are their implicit biases and beliefs? Many of the issues surrounding fairness, non-discrimination, and the promotion of human values are rooted in the task formulation, data collection, and annotation stages of the problem.

We also need to focus on what happens after the classification: is a binary label of "abusive" or "not abusive" truly sufficient? How can the decision be explained? How can the decision be appealed, and can that new knowledge be fed back to the classifier? Issues of explainability, accountability, human control, and professional responsibility must be addressed. We also encourage a stronger collaboration with experts from other fields, to better understand how NLP practitioners can design systems to not only block abusive language, but actually reduce it. There have been innovative proposals relating to counter-narratives, automated re-wording, and educational applications that can help raise awareness of the underlying social inequities being expressed and reinforced through language. Younger users may not have complete awareness of the context and implications of their words, and throughout our lifetimes, the norms surrounding language use and what words are considered appropriate or inappropriate are constantly shifting as we collectively seek to move towards a more inclusive public discourse. While such novel applications will certainly not 'solve' the social problems underlying abusive language online, it may be a step in the right direction, and one to which NLP researchers can contribute.

References

Abu Farha, I., & Magdy, W. (2020). Multitask learning for Arabic offensive language and hate-speech detection. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools*, pp. 86–90.

Adams, C. (2018). New York Times: Using AI to host better conversations. Google Blog.

Afzal, M., & Panagiotopoulos, P. (2020). Smart policing: A critical review of the literature. In *Proceedings of the International Conference on Electronic Government*, pp. 59–70.

Al Kuwatly, H., Wich, M., & Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 184–190.

Alberti, P., Castaneda, M., Castrucci, B., Harrison, L., et al. (2020). Engaging with communities—lessons (re) learned from COVID-19. *Preventing Chronic Disease*, *17*.

Ali, O., Scheidt, N., Gegov, A., Haig, E., Adda, M., & Aziz, B. (2020). Automated detection of racial microaggressions using machine learning. In *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 2477–2484. IEEE.

Alonso, P., Saini, R., & Kovács, G. (2020). Hate speech detection using transformer ensembles on the hasoc dataset. In *Proceedings of the International Conference on Speech and Computer*, pp. 13–21.

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD).*

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 45–54.

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., & Varshney, K. R. (2019). Factsheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, *63*(4/5), 6:1–6:13.

Arora, I., Guo, J., Levitan, S. I., McGregor, S., & Hirschberg, J. (2020). A novel methodology for developing automatic harassment classifiers for Twitter. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 7–15.

Aroyo, L., Dixon, L., Thain, N., Redfield, O., & Rosen, R. (2019). Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Proceedings of the Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing*, pp. 1100–1105.

Arsht, A., & Etcovitch, D. (2018). The human cost of online content moderation. *Harvard Journal of Law and Technology.*

Assimakopoulos, S., Vella Muskat, R., van der Plas, L., & Gatt, A. (2020). Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5088–5097.

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of Twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, *38.*

Banko, M., MacKeen, B., & Ray, L. (2020). A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 125–137.

Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pp. 1–16.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 54–63, Minneapolis, Minnesota, USA.

Basu, T. (2019). Reddit's automoderator is the future of the internet, and deeply imperfect. *MIT Technology Review*.

Bender, E. M. (2019). A typology of ethical risks in language technology with an eye towards where transparent documentation can help. In *Future of Artificial Intelligence: Language, Ethics, Technology Workshop*.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604.

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., & Wright, L. (2016). Counterspeech on Twitter: A field study. *A report for Public Safety Canada under the Kanishka Project*.

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction*, *1*(2), 1–19.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500.

Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., & Maurizio, T. (2018). Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, Vol. 2263, pp. 1–9. CEUR.

Breitfeller, L., Ahn, E., Jurgens, D., & Tsvetkov, Y. (2019). Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1664–1674, Hong Kong, China.

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*(3), 273–291.

Bunting, M. (2018). From editorial obligation to procedural accountability: policy approaches to online content in the era of information intermediaries. *Journal of Cyber Policy*, *3*(2), 165–186.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., & Smith, A. (2019). From soft classifiers to hard decisions: How fair can we be?. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 309–318.

Carlson, C. (2017). Censoring hate speech in US social media content: Understanding the user's perspective. *Communication Law Review*, *17*(1), 24–45.

Cecillon, N., Labatut, V., Dufour, R., & Linarès, G. (2019). Abusive language detection in online conversations by combining content- and graph-based features. *Frontiers in Big Data*, *2*.

Cecillon, N., Labatut, V., Dufour, R., & Linares, G. (2020). WAC: A corpus of Wikipedia conversations for online abuse detection. In *Proceedings of The Language Resources and Evaluation Conference*.

Chakrabarty, T., Gupta, K., & Muresan, S. (2019). Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 70–79, Florence, Italy.

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–25.

Cheng, L., Guo, R., & Liu, H. (2019). Robust cyberbullying detection with causal interpretation. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 169–175.

Citron, D. K., & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, *91*.

Conbere, T. (2019). A wretched hive of scum and villainy: How Twitter encourages harassment (and how to fix it). Master's thesis, University of Oregon.

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In *Proceedings of the European Conference on Information Retrieval*, pp. 693–696.

Darvish Rouani, B., Samragh, M., Javidi, T., & Koushanfar, F. (2019). Safe machine learning and defeating adversarial attacks. *IEEE Security Privacy*, *17*(2), 31–38.

Davani, A. M., Omrani, A., Kennedy, B., Atari, M., Ren, X., & Dehghani, M. (2020). Fair hate speech detection through evaluation of social group counterfactuals. *arXiv preprint arXiv:2010.12779*.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 25–35, Florence, Italy.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Delgado, R., & Stefancic, J. (1997). *Must we defend Nazis?: Hate speech, pornography, and the new first amendment*. NYU Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Confer-*

ence of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), pp. 4171–4186.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-dimensional gender bias classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 314–331.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the International Conference on World Wide Web*, pp. 29–30.

Duarte, N., Llanso, E., & Loup, A. C. (2018). Mixed messages? The limits of automated social media content analysis. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT)*.

Duggan, M. (2017). *Online Harassment 2017*. Pew Research Center.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 1–29.

Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pp. 214–228.

Field, A., & Tsvetkov, Y. (2020). Unsupervised discovery of implicit gender bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 596–608.

Fišer, D., Erjavec, T., & Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*, pp. 46–51.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30.

Fortuna, P., Soler, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6786–6794.

Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2019). A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, pp. 105–114.

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Fraser, K. C., Nejadgholi, I., & Kiritchenko, S. (2021). Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, Online.

Friedman, R. A., & Currall, S. C. (2003). Conflict escalation: Dispute exacerbating elements of e-mail communication. *Human Relations*, *56*(11), 1325–1347.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pp. 85–90.

Gao, L., & Huang, R. (2017). Detecting online hate speech using context aware models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pp. 260–266, Varna, Bulgaria.

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., & Beutel, A. (2019). Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2018). Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*.

Gelber, K., & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324–341.

Gencoglu, O. (2020). Cyberbullying detection with fairness constraints. *IEEE Internet Computing*.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230.

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gnanasekaran, R. K., Gunasekaran, R. R., et al. (2017). A large labeled corpus for online harassment research. In *Proceedings of the ACM Conference on Web Science*, pp. 229–233.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1).

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 90–99.

Grimsley, C., Mayfield, E., & Bursten, J. R. (2020). Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 1780–1790.

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

Gunasekara, I., & Nejadgholi, I. (2018). A review of standard text classification practices for multi-label toxicity identification of online content. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pp. 21–25.

Guzman-Silverio, M., Balderas-Paredes, A., & López-Monroy, A. (2020). Transformers and data augmentation for aggressiveness detection in Mexican Spanish. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain.*

Haapoja, J., Laaksonen, S.-M., & Lampinen, A. (2020). Gaming algorithmic hate-speech detection: Stakes, parties, and moves. *Social Media+ Society*, *6*(2).

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65*, 399–423.

Helberger, N. (2020). The political power of platforms: How current attempts to regulate misinformation amplify opinion power. *Digital Journalism*, *8*(6), 842–854.

Hinduja, S., & Patchin, J. W. (2020). *Cyberbullying fact sheet: Identification, Prevention, and Response.* Cyberbullying Research Center.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API built for detecting toxic comments. *arXiv preprint arXiv:1702.08138.*

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the Instagram social network. In *Proceedings of the International Conference on Social Informatics*, pp. 49–66.

Huang, X., Xing, L., Dernoncourt, F., & Paul, M. (2020). Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 1440–1448.

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 46–57.

Isaac, M., & Browning, K. (2020). Lawmakers drill down on how facebook and twitter moderate content. The New York Times.

Ishida, S. (2020). What makes discrimination morally wrong? A harm-based view reconsidered. *Theoria.*

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 3543–3556, Minneapolis, Minnesota.

Jenders, M., Kasneci, G., & Naumann, F. (2013). Analyzing and predicting viral tweets. In *Proceedings of the International Conference on World Wide Web*, pp. 657–664.

Jin, X., Barbieri, F., Kennedy, B., Davani, A. M., Neves, L., & Ren, X. (2021). On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3770–3783.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online. Association for Computational Linguistics.

Jurgens, D., Hemphill, L., & Chandrasekharan, E. (2019). A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3658–3666, Florence, Italy.

Kalin, J., Noever, D., & Dozier, G. (2020). Systematic attack surface reduction for deployed sentiment analysis models. In *Proceedings of the 8th International Conference on Security, Privacy and Trust Management (SPTM)*.

Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., Bintz, C., Raz, D., & Krafft, P. M. (2020). Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 45–55.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*.

Kiritchenko, S., & Nejadgholi, I. (2020). Towards ethics by design in online abusive content detection. *arXiv preprint arXiv:2010.14952*.

Koetsier, J. (2020). Report: Facebook makes 300,000 content moderation mistakes every day. Forbes.

Konečnỳ, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 1–11.

Kurita, K., Michel, P., & Neubig, G. (2020). Weight poisoning attacks on pre-trained models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 166–172.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Leonard, A. (Ed.). (2019). *Models for Platform Governance*, A CIGI Essay Series. Centre for International Governance Innovation.

Lepoutre, M. (2017). Hate speech in public discourse: A pessimistic defense of counter-speech. *Social Theory and Practice*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mahajan, A., Shah, D., & Jafar, G. (2020). Explainable AI approach towards toxic comment classification. EasyChair Preprint.

Maitra, I., & McGowan, M. K. (2012). *Speech and harm: Controversies over free speech*. Oxford University Press on Demand.

Manzini, T., Chong, L. Y., Black, A. W., & Tsvetkov, Y. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 615–621.

Marcinkevičs, R., & Vogt, J. E. (2020). Interpretability and explainability: A machine learning zoo mini-tour. *arXiv preprint arXiv:2012.01805*.

Marcinowski, M., & Ławrynowicz, A. (2020). Predicting the outbreak of conflict in online discussions using emotion-based features. In *Proceedings of the International Conference on Web Engineering*, pp. 505–511.

Martinez, N., Bertran, M., & Sapiro, G. (2019). Fairness with minimal harm: A pareto-optimal approach for healthcare. *arXiv preprint arXiv:1911.06935*.

Matsuda, M. J. (2018). *Words that wound: Critical race theory, assaultive speech, and the first amendment*. Routledge.

May, C., Wang, A., Bordia, S., Bowman, S., & Rudinger, R. (2019). On measuring social biases in sentence encoders. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 622–628.

Miok, K., Skrlj, B., Zaharie, D., & Robnik-Sikonja, M. (2020). To BAN or not to BAN: Bayesian attention networks for reliable hate speech detection. *arXiv preprint arXiv:2007.05304*.

Mishra, P., Yannakoudakis, H., & Shutova, E. (2019). Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229.

Mou, G., Ye, P., & Lee, K. (2020). Swe2: Subword enriched and significant word emphasized framework for hate speech detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1145–1154.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection in online social media. In *Proceedings of the International Conference on Complex Networks and Their Applications*, pp. 928–940.

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pp. 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Nadeem, M., Bethke, A., & Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Naseem, U., Razzak, I., & Eklund, P. W. (2020). A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on Twitter. *Multimedia Tools and Applications*, 1–28.

Nejadgholi, I., & Kiritchenko, S. (2020). On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the 4th Workshop on Online Abuse and Harms*.

Niemann, M., Riehle, D. M., Brunk, J., & Becker, J. (2019). What is abusive language?. In *Proceedings of the Multidisciplinary International Symposium on Disinformation in Open Online Media*, pp. 59–73.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the International Conference on World Wide Web*, pp. 145–153.

Nogueira dos Santos, C., Melnyk, I., & Padhi, I. (2018). Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 189–194, Melbourne, Australia.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Ousidhoum, N., Song, Y., & Yeung, D.-Y. (2020). Comparative evaluation of label agnostic selection bias in multilingual hate speech datasets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2532–2542.

Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2799–2804, Brussels, Belgium.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). Toxicity detection: Does context really matter?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4296–4305, Online. Association for Computational Linguistics.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. *Sensors*, *19*(21).

PerspectiveAPI (2017). Perspective API documentation - score normalization and feedback. https://github.com/conversationai/perspectiveapi/blob/master/3-concepts/score-normalization.md.

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Pitropakis, N., Kokot, K., Gkatzia, D., Ludwiniak, R., Mylonas, A., & Kandias, M. (2020). Monitoring users' behavior: Anti-immigration speech detection on Twitter. *Machine Learning and Knowledge Extraction*, *2*(3), 192–215.

Pohjonen, M., & Udupa, S. (2017). Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, *11*, 19.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 1–47.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., & Bosco, C. (2017). Hate speech annotation: Analysis of an Italian Twitter corpus. In *Proceedings of the Italian Conference on Computational Linguistics (CLiC-it 2017)*, pp. 1–6.

Prabhakaran, V., Waseem, Z., Akiwowo, S., & Vidgen, B. (2020). Online abuse and human rights: WOAH satellite session at RightsCon 2020. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 1–6.

Pratt, B. (2019). Inclusion of marginalized groups and communities in global health research priority-setting. *Journal of Empirical Research on Human Research Ethics*, *14*(2), 169–181.

Pratt, B., Cheah, P. Y., & Marsh, V. (2020). Solidarity and community engagement in global health research. *The American Journal of Bioethics*, *20*(5), 43–56.

Price, I., Gifford-Moore, J., Flemming, J., Musker, S., Roichman, M., Sylvain, G., Thain, N., Dixon, L., & Sorensen, J. (2020). Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 114–124, Online. Association for Computational Linguistics.

Prost, F. (2020). Mitigating unfair bias in ML models with the mindiff framework. Google AI Blog.

Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 33–44.

Raufi, B., & Xhaferri, I. (2018). Modelling and implementation of machine learning techniques for hate speech detection in mobile applications. In *Proceedings of the International Conference on Information Technologies (InfoTech-2018)*.

Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive language detection using multi-level classification. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pp. 16–27.

Razo, D., & Kübler, S. (2020). Investigating sampling bias in abusive language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 70–78.

Reich, D., Todoki, A., Dowsley, R., De Cock, M., & Nascimento, A. C. (2019). Privacy-preserving classification of personal text messages with secure multi-party computation: An application to hate-speech detection. *arXiv preprint arXiv:1906.02325*.

Reicher, S. D., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, *6*(1), 161–198.

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., & Meira Jr, W. (2018). Characterizing and detecting hateful users on Twitter. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Risch, J., & Krestel, R. (2020). Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 55–61.

Rizos, G., Hemker, K., & Schuller, B. (2019). Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 991–1000.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, Online.

Safi Samghabadi, N., Patwa, P., PYKL, S., Mukherjee, P., Das, A., & Solorio, T. (2020). Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*.

Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, *11*(1), 3–24.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-g., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, *10*(1), 1.

Saltman, E. M., & Russell, J. (2014). The role of prevent in countering online extremism. *Quilliam Publication.*

Samuel, J., Garvey, M., & Kashyap, R. (2019). That message went viral?! exploratory analytics and sentiment analysis into the propagation of tweets. In *Proceedings of the Northeast Decision Sciences Institute (NEDSI) Conference.*

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Saria, S., & Subbaswamy, A. (2019). Tutorial: safe and reliable machine learning. *arXiv preprint arXiv:1904.07204.*

Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counterspeech on Facebook. In *Proceedings of the 66th ICA Annual Conference*, pp. 1–23.

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

Scott, M., & Kayali, L. (2020). What happened when humans stopped managing social media content. POLITICO.

Sen, A. K. (2009). *The idea of justice.* Harvard University Press.

Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5248–5264.

Shepherd, T., Harvey, A., Jordan, T., Srauy, S., & Miltner, K. (2015). Histories of hating. *Social Media+ Society*, *1*(2).

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, *3*(5), 637–646.

Sigurbergsson, G. I., & Derczynski, L. (2020). Offensive language and hate speech detection for Danish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 3498–3508.

Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *Proceedings of the AAAI Conference on Innovative Applications of Artificial Intelligence*, pp. 1058–1065.

Srivastava, S., & Khurana, P. (2019). Detecting aggression and toxicity using a multi dimension capsule network. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 157–162.

Stappen, L., Brunn, F., & Schuller, B. (2020). Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and AXEL. *arXiv preprint arXiv:2004.13850.*

Subramani, S., Michalska, S., Wang, H., Du, J., Zhang, Y., & Shakeel, H. (2019). Deep learning for multi-class identification from domestic violence online posts. *IEEE Access*, *7*, 46210–46224.

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

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In Austin, W. G., & Worchel, S. (Eds.), *The Social Psychology of Intergroup Relations*, chap. 3, pp. 33–47. Brooks/Cole Pub. Co.

Tekiroğlu, S. S., Chung, Y.-L., & Guerini, M. (2020). Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1177–1190.

Thylstrup, N., & Waseem, Z. (2020). Detecting 'dirt' and 'toxicity': Rethinking content moderation as pollution behaviour. Available at SSRN: `https://ssrn.com/abstract=3709719`.

Tran, M., Zhang, Y., & Soleymani, M. (2020). Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tworek, H., & Tenove, C. (2020). Dispute resolution and content moderation: Fair, accountable, independent, transparent, and effective. *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression.*

Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, *22*(1), 69–80.

Unsvåg, E. F., & Gambäck, B. (2018). The effects of user features on Twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online*, pp. 75–85.

Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online.*

Van Bruwaene, D., Huang, Q., & Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, 1–24.

Van Geel, M., Vedder, P., & Tanilon, J. (2014). Relationship between peer victimization, cyberbullying, and suicide in children and adolescents: a meta-analysis. *JAMA Pediatrics*, *168*(5), 435–442.

Varshney, K. R. (2016). Engineering safety in machine learning. In *Proceedings of the Information Theory and Applications Workshop (ITA)*, pp. 1–5.

Varshney, K. R. (2020). On mismatched detection and safe, trustworthy machine learning. In *Proceedings of the 54th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–4. IEEE.

Vaughan, J. W., & Wallach, H. (2020). A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence.*

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, *15*(12).

Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., Botelho, A., Hall, M., & Tromble, R. (2020a). Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 162–172.

Vidgen, B., Hale, S., Staton, S., Melham, T., Margetts, H., Kammar, O., & Szymczak, M. (2020b). Recalibrating classifiers for interpretable abusive content detection. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pp. 132–138.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 80–93, Florence, Italy.

Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., & Tromble, R. (2021). Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2289–2303, Online. Association for Computational Linguistics.

Vincent, N. A., & Jane, E. A. (2017). Beyond law: Protecting cyber victims through engineering and design. In Martellozzo, E., & Jane, E. A. (Eds.), *Cybercrime and its Victims: An International Perspective*, pp. 209–223. Routledge, Oxon.

Vu, X.-S., Vu, T., Tran, M.-V., Le-Cong, T., & Nguyen, H. (2020). HSD shared task in VLSP campaign 2019: Hate speech detection for social good. *arXiv preprint arXiv:2007.06493*.

Waldron, J. (2012). *The harm in hate speech.* Harvard University Press.

Wang, C. (2018). Interpreting neural network hate speech classifiers. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pp. 86–92.

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pp. 138–142.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pp. 78–84, Vancouver, BC, Canada.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pp. 88–93.

Waseem, Z., Thorne, J., & Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In Golbeck, J. (Ed.), *Online harassment*, pp. 29–55. Springer.

Weller, A. (2019). Transparency: motivations and challenges. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 23–40. Springer.

Wich, M., Al Kuwatly, H., & Groh, G. (2020). Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 191–199.

Wiedemann, G., Yimam, S. M., & Biemann, C. (2020). UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the 14th Workshop on Semantic Evaluation*, pp. 1638–1644.

Wiegand, M., Ruppenhofer, J., & Eder, E. (2021). Implicitly abusive language – what does it actually look like and why are we not getting there?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 576–587, Online. Association for Computational Linguistics.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 602–608.

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018a). Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1046–1056.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018b). Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, Vienna, Austria.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20.

Wilhelm, C., & Joeckel, S. (2019). Gendered morality and backlash effects in online discussions: An experimental study on how users respond to hate speech comments against women and sexual minorities. *Sex Roles*, *80*(7-8), 381–392.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1391–1399.

Wullach, T., Adler, A., & Minkov, E. M. (2020). Towards hate speech detection at large via deep generative modeling. *IEEE Internet Computing*.

Xu, J., Ju, D., Li, M., Boureau, Y.-L., Weston, J., & Dinan, E. (2020). Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1–19.

York, J., & McSherry, C. (2019). Content moderation is broken. Let us count the ways. Electronic Frontier Foundation Blog.

Young, J., Swamy, P., & Danks, D. (2018). Beyond AI: Responses to hate speech and disinformation. Research Report.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1415–1420, Minneapolis, Minnesota.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 75–86.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Zhang, J., Chang, J., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Taraborelli, D., & Thain, N. (2018a). Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1350–1361, Melbourne, Australia.

Zhang, Z., Robinson, D., & Tepper, J. (2018b). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Proceedings of the European Semantic Web Conference*, pp. 745–760.

Zhou, X., Sap, M., Swayamdipta, S., Choi, Y., & Smith, N. (2021). Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3143–3155, Online. Association for Computational Linguistics.

Ziems, C., He, B., Soni, S., & Kumar, S. (2020). Racism is a virus: Anti-Asian hate and counterhate in social media during the COVID-19 crisis. *arXiv preprint arXiv:2005.12423*.

Zufall, F., Zhang, H., Kloppenborg, K., & Zesch, T. (2020). Operationalizing the legal concept of 'incitement to hatred' as an NLP task. *arXiv preprint arXiv:2004.03422*.