A Unified Typology of Harmful Content

Michele Banko, Brendon MacKeen and Laurie Ray
Sentropy Technologies
380 Portage Avenue
Palo Alto, CA 94306
{mbanko,brendon,laurie}@sentropy.io

Abstract

The ability to recognize harmful content within online communities has come into focus for researchers, engineers and policy makers seeking to protect users from abuse. While the number of datasets aiming to capture forms of abuse has grown in recent years, the community has not standardized around how various harmful behaviors are defined, creating challenges for reliable moderation, modeling and evaluation. As a step towards attaining shared understanding of how online abuse may be modeled, we synthesize the most common types of abuse described by industry, policy, community and health experts into a *unified typology of harmful content*, with detailed criteria and exceptions for each type of abuse.

1 Introduction

Content moderation, the practice of monitoring and reviewing user-generated content to ensure compliance with legal requirements, community guidelines, and user agreements, is important for creating safe and equitable online spaces. While traditional content moderation systems rely heavily on human reviewers who use a set of proprietary guidelines to determine if content is in violation of policy, the use of algorithmic approaches has become a part of moderation workflows in recent years. While not a full replacement for human content moderators, the use of AI promises to reduce trauma and cost incurred by purely human-centric workflows.

As a result, the ability to recognize abusive content using data-driven approaches has attracted attention from researchers in the computational and social sciences. To study, model and measure systems designed to recognize online abuse, researchers typically create labelled datasets using crowdsourcing platforms or in-house annotators. While the number of research datasets continues to grow (Vidgen and Derczynski, 2020), the research community has not reached a consensus on how common abuse types are defined. Despite the use of best practices that leverage multiple annotators, definitional ambiguity can lead to the creation of datasets of questionable consistency (Ross et al., 2017; Waseem, 2016; Wulczyn et al., 2017). Furthermore, without thorough domain understanding, research datasets built to capture abusive content may be prone to unintended bias (Wiegand et al., 2019). Together, these shortcomings create challenges for reliable modeling and study of abuse as it occurs in the real world.

Harmful content has also drawn attention from internet companies, such as those in the social media, online gaming, and dating industries, who seek to protect their users from abuse. These companies typically employ a Trust and Safety organization to define and enforce violations of content policies, and to develop tools aimed at identifying instances of abuse on their platforms. Several online platforms which have seen large volumes of harmful content on their platforms, have created content policies that can be useful in specifying definitions of various abuse classes. In the absence of a standard upon which content policies can be based, community standards within digital platforms are largely shaped by users who report abuse they have experienced firsthand. Additionally, some aspects of content policies are informed by requirements handed down from local law enforcement agencies wishing to prosecute users engaging in illegal activity online.

Recently, the demand for internet companies to more aggressively reduce the spread of cyberbullying, radicalization, deception, exploitation, and other forms of dangerous content has been increasingly called for by governmental and civil society organizations. The proposed Online Harms Bill in the UK and amendments to Section 230 in the United States call for stricter accountability, trans-

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 125–137
Online, November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

parency, and regulations to be imposed on companies hosting user-generated content. Civil society organizations have yielded numerous proposals for better describing types of harmful content online so that internet companies may better understand the nature and impact of such content.

In this paper, we enumerate, consolidate and define the most common types of abuse described in content policies for several major online platforms and white papers from civil society organizations. We look for commonalities in both *what* types of abuse have been identified and *how* they are defined. Our goal is to provide a unified typology of harmful content, with clear criteria and exceptions for each type of abuse. While hate speech and harassment have attracted attention from the natural language processing community in recent years, upon close study, we find that the domain of harmful content is broader than many may have realized. We hope this typology will benefit content moderation systems by:

- Defining abuse types that are readily usable by content moderators, both human and algorithmic
- Encouraging the construction of accurate, complete and unbiased datasets used for model training and evaluation
- Creating awareness of types of abuse that have received limited attention in the research community thus far

2 Background and Related Work

2.1 Abuse Typologies

Several efforts to categorize online abuse began by closely studying specific types of harm. With a focus on cyberbullying, Van Hee et al. (2015) published a scheme for annotation, which considers the presence, severity, role of the author (harasser, victim or bystander) and a number of fine-grained categories, such as insults and threats. Waseem et al. (2017) discussed the lack of consensus around how hate speech is defined, noting that messages labeled as hate speech in some datasets are only considered to be offensive in others. They devised a two-fold typology that considers whether hate is directed at a specific target (as opposed to taking the form of a general statement), and the degree of explicitness. Anzovino et al. (2018) studied misogynistic social media posts, and modeled seven types of abuse, most of which extend beyond abuse directed at women. Similar to these works, we break apart class definitions into fine-grained categories when possible in an attempt to disambiguate potentially underspecified requirements. We build upon this body of work by considering a larger set of abuse types, as opposed to just cyberbullying or hate speech.

More broadly, Vidgen et al. (2019) noted the difficulties in categorizing abusive content, and proposed a three dimensional scheme for defining abuse classes. They suggest to consider (1) the type of the abuse target (e.g. individual, identity, entity or concept), (2) the recipient of the abuse (e.g. a specific individual, women, capitalism), and (3) the manner in which the abuse is articulated (e.g. as an insult, aggression, stereotype, untruth). We consider target type and manner in our categorization and take the suggested scheme one step further by instantiating a large set of what the authors refer to as subtasks. In some cases, where there is no clear or uniform target (e.g. misinformation) we found it helpful to organize types based on topic or possible outcome that can be easily reasoned about by those impacted by moderation systems.

2.2 Hate Speech

The natural language processing community has largely focused on detection of hate speech and cyberbullying (Schmidt and Wiegand, 2017). As a result, a number of research datasets have been produced (Vidgen and Derczynski, 2020), yet none have used the same definition, or have annotated only partial phenomena (e.g. annotating racist and sexist speech, but not hate speech directed at all groups who require protection).

While building a hate speech corpus from Twitter data, Ross et al. (2017) investigated how the reliability of the annotations is affected by the provision of accompanying definitions. They compared annotations in which the annotators were provided Twitter's definition of hate speech versus no definition. While annotators shown the definition were more likely to ban the tweet, the authors found that even when presented with Twitter's definition, inter-annotator agreement, measured using Krippendorf's alpha, was at best 0.3, depending on the question asked.[1] Ross et al. concluded that more detailed coding schemes are needed to be

[1] Krippendorff (2004) suggests that for annotations to be considered reliable, a minimum score of 0.80 is desirable, with 0.667 being the lowest conceivable limit

able to distinguish hate speech from other content.

Other works which report the difficulty of achieving high levels of interannotator agreement when compiling hate speech datasets include Davidson et al. (2017), who found that 5% of tweets were coded as hate speech by the majority of annotators with only 1.3% being annotated unanimously as containing hate speech. The creators of the 2018 Kaggle Toxic Comment Classification Challenge (Wulczyn et al., 2017) report that while the challenge dataset was built using ten annotators per label, agreement was weak (Krippendorff alpha of 0.45).

Demonstrating the importance of having well-defined annotation guidelines, Waseem and Hovy (2016) articulated an eleven-point definition of gendered and racial attacks. The use of detailed criteria yielded a high level of agreement. The authors measured inter-annotator agreement, defined using Cohen's kappa, to be 0.84.

2.3 Other Forms of Harmful Content

In recent years, machine learning has been used to recognize forms of self-harm such as pro-eating disorder content in social media posts (Chancellor et al., 2016; Wang et al., 2017) and suicidal ideation (Burnan et al., 2015; Cao et al., 2019).

Snyder et al. (2017) developed an automated framework for detecting dox files, i.e. files which reveal personally identifiable data without consent, and measuring the frequency, content, targets, and effects of doxing on popular dox-posting sites.

Detection of sexually explicit content includes efforts to recognize instances of child sexual abuse (Lee et al., 2020) and human trafficking (Dubrawski et al., 2015; Tong et al., 2017).

Another endeavor related to online harm that has gained attention within the research community is the detection of misinformation, which is surveyed by Su et al. (2020).

3 Methodology

To develop a unified typology of harmful content, we employed a grounded theory approach, in which we synthesized inputs from several sources:

- Community guidelines and content policy made public by large online platforms, specifically Discord,[2] Facebook,[3] Pinterest,[4] Reddit,[5] Twitter,[6] and YouTube[7]
- The International Covenant on Civil and Political Rights,[8] an international human rights treaty developed by the United Nations
- Proposals from members of civil society organizations such as the Women's Media Center; Internet and Jurisdiction Policy Network (2019) and Benesch (2020)
- Recommendations from experts and health organizations who study psychological and physical impact of abuse, including the American Association of Suicidology and the Conflict Tactics Scale

3.1 Principles

While qualitatively analyzing the data mentioned above, we used the following principles to guide the creation of the typology:

Avoid the use of subjective adjectives as core definitions. As prior research has shown, annotation tasks that make use of underspecified or subjective phrases such as "hateful," "toxic" or "would make you leave a conversation," without further explanation are likely to be interpreted differently depending on the annotator. Enumerate problematic content types using precise objective criteria when possible.

Prefer fine-grained classes over those spanning multiple behaviors. Behavior that is casually described as "toxic" or "bullying" may contain a mix of identity-based hate speech, general insults, threats and inappropriate sexual language. Narrowly defined classes simplify annotation requirements and provide a level of explainability that is missing from underspecified labels.

Consider the type of the subject of abuse. To keep definitions well-scoped, we consider the subject of the attack, and avoid mixing subject types within a single definition when possible. For instance, instead of having a generic class aimed at recognizing sexually explicit content, we advocate for annotating sexually charged content directed at an individual separately from content advertising for adult sexual services. However, we find that

[2] https://discord.com/guidelines

[3] https://www.facebook.com/communitystandards/

[4] https://policy.pinterest.com/en/community-guidelines

[5] https://www.reddithelp.com/hc/en-us/sections/360008810092-Account-and-Community-Restrictions

[6] https://help.twitter.com/en/rules-and-policies

[7] https://support.google.com/youtube/topic/2803176?hl=en&ref$_topic = 6151248$

[8] https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx

there are some instances in which a broad form of harm can not be uniformly defined by the type of the target, and employ a topical approach that may be more understandable to moderators and users of social platforms. An example of this is *Misinformation*.

Consider potential downstream actions. If a type of behavior is universally associated with an outcome, avoid definitions that mix behaviors that do not share the outcome. For instance, child sexual abuse content is not tolerated under any circumstances in most countries and must be reported to law enforcement, whereas insults that make use of sexual terms are unlikely to have legal ramifications. A platform may have strict policies against attacks on protected groups but permit mild forms of non-identity based insults. Distinguishing between the two simplifies the ability to enforce policy and therefore, improves the usefulness of a moderation system.

Despite our preference for fine-grained classes, they are not defined to be mutually exclusive. Additionally, hierarchical arrangement of types is not always possible. As a result, there are cases where multiple types may apply to a single input. For example, *Time to shoot this n*****, where the last word represents a racial slur, should be classified as both *Identity Attack* and *Threat of Violence*. *Time to shoot up this school* is a violent threat without an identity-based attack. *N***** aren't welcome here* is a non-violent *Identity Attack*.

3.2 Severity

All forms of abuse are problematic and require some means to identify and address them in order to mitigate their impact on users. While the strength of statements involving abuse may be interpreted differently depending on the recipient and context, some forms of online harm present immediate or lasting danger to individuals or stand in violation of the law. For each type of abuse we present, we establish qualifications for what may be considered *severe abuse*. The ability to detect severe abuse is critical for content moderation systems seeking to identity extreme or time-sensitive violations quickly.

The concept of "severe toxicity" is annotated in the Kaggle Toxic Comment Classification Challenge (2018), where it is defined it as "rude, disrespectful, or unreasonable comments that are very likely to make people leave a discussion." Moving away from the use of subjective adjectives, we consider the following attributes when determining severity:

- Use of language expressing direct intent (severe) vs. use of language that is passive or merely wishful (not severe)
- Time-sensitive or immediate threats of harm are considered to be severe
- Consequences of, or degree of harm associated with, the abuse, i.e. actions resulting in death or long-lasting physical or psychological trauma shall be treated as severe
- Vulnerability of the target, e.g. attacks directed at members of groups that have been historically marginalized, dehumanized or objectified are considered to be severe
- Violations of personal privacy and consent are treated as severe
- Violations of applicable laws, including internationally recognized policies are handled as severe

4 Abuse Class Definitions

Using the data and guidelines described in Section 3, we arrived at the typology depicted in Figure 1. In the remainder of this section, we describe each type in detail. Within each section, we present the types in lexicographic order.

4.1 Hate and Harassment

Hate and Harassment describes abuse directed at a specific individual or group of people (e.g. identity) meant to torment, demean, undermine, frighten or humiliate the target. Abuse directed at institutions or abstract concepts is not included in this set of definitions. In the remainder of this section, we present criteria for defining common forms of hate and harassment: *Doxing*, *Identity Attack*, *Identity Misrepresentation*, *Insult*, *Sexual Aggression*, and *Threat of Violence*.

4.1.1 Doxing

Doxing is a form of severe abuse in which a malicious party tries to harm an individual by releasing personally identifiable information about the target to the general public. During a doxing attack, sensitive information is typically distributed on web sites that permit anonymous posting and

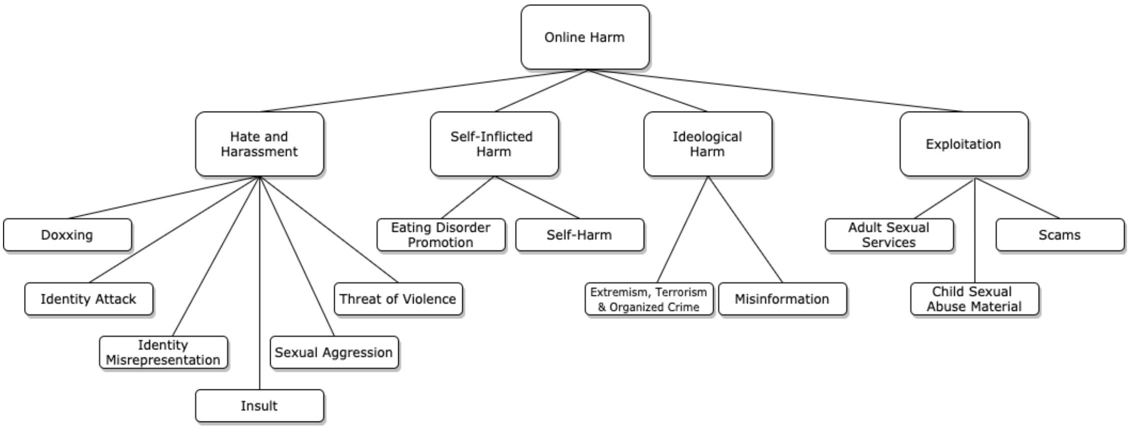

Figure 1: A Typology of Harmful Content

do not proactively remove harassing content, such as onion sites, torrents, IRC, and anonymous text sharing websites such as pastebin.com, 4chan and 8chan. Doxing can lead to another form of harm known as SWATing, in which someone calls law enforcement with false reports of violence at an address in order to cause harm at the target's residence (e.g. a SWAT team kicking in their door).

Personal data that should not be shared without the consent of others includes:

- Physical or virtual locations such as home, work and IP addresses, or GPS locations
- Contact information such as private email address and phone numbers
- Identification numbers such as Social Security, passport, government or school ids
- Digital identities such as social network accounts, chat identities, and passwords
- Personal financial information such as bank account or credit card information
- Criminal and medical histories

Mentions of data already in the public domain such as one's place of education or employment, email addresses that have been voluntarily shared (such as on a personal homepage) are not considered instances of doxing, nor are cases where people willingly share their own private information.

4.1.2 Identity Attack

Identity Attack is a form of online abuse where malicious actors severely attack individuals or groups of people based on their membership in a protected or vulnerable group. During an *Identity Attack*, a bad actor will use language reflecting the intent to dehumanize, persecute or promote violence based on the identity of the subject. The use of slurs and/or derogatory epithets may be present but is not a requirement.

While there is neither agreement in what constitutes hate speech across academic datasets, nor is there any industry or legal standard of this definition, constructions of *Identity Attack* definitions typically attempt to: 1) protect vulnerable groups, 2) protect specific characteristics or attributes of individuals, 3) prohibit hate speech but fail to offer a definition. For platforms that fall into either of the first two categories the policies described protect users from attacks and violence on the basis of identity-based attributes including age, disability, ethnicity, gender identity, military status, nationality, race, religion, and sexual orientation. Some platforms offer additional protections for vulnerable groups such as immigration status, socioeconomic class or the presence of a medical condition.

Jigsaw (Kaggle, 2018) defines *Identity Attack* as: "Negative or hateful comments targeting someone because of their identity." Underspecified definitions such as this are difficult to use in practice due to the open-endedness of how "negative" or "hateful" may be interpreted by annotators or community moderators. As mentioned in our survey of related work, we promote the use of more fine-grained classes, focusing here on severe attacks (e.g. those with dehumanizing and/or violent intent) and defining a separate class for more mild phenomena such as spread of negative stereotypes or misinformation related to vulnerable groups. Another distinction we suggest is to ensure the subject of the attack refers to a human or group thereof, as opposed to institutions or organizations. For example, an attack on people who practice a religion (e.g. Jews, Muslims) falls under the class, but attacks on religion itself (e.g. Judiasm, Islam) itself do not. As a result, statements such as *You deserve to be euthanized, you dirty* **** and *****s deserve to be euthanized"*[9] would be treated as *Identity Attack*, whereas **** *is a religion that should cease to exist* would not. In some cases, it is possible that use of organizations are replacements for the individuals belonging to them, but for the first version of the typology we propose to maintain this distinction.

Here we summarize types of content that warrant a classification of *Identity Attack*, all of which are considered severe forms of abuse:

- Explicit use of slurs and other derogatory epithets referencing an identity group
- Violent threats or calls for harm directed at an identity group
- Calls for exclusion, domination or suppression of rights, directed at an identity group
- Dehumanization of an identity group, including comparisons to animals, insects, diseases or filth, generalizations involving physical unattractiveness, low intelligence, mental instability and/or moral deficiencies
- Expressions of superiority of one group over a protected or vulnerable group
- Admissions of hate and intolerance towards members of an identity group
- Denial of another's identity, calls for conversion therapy, deadnaming
- Support for hate groups communicating intent described above

The following should be considered non-examples for the *Identity Attack* abuse class:

- Attacks on institutions or organizations (as opposed to the people belonging to them)
- Promotion of negative stereotypes, fear or misinformation related to an identity group (defined as *Identity Misrepresentation* in Section 4.1.3)
- In-group usage of slurs and their variants, reclamation of hateful terms by the those who have been historically targeted
- Discussion of meta-linguitic nature or education related to slurs or hate speech
- Accounts of the speech behavior of parties external to the immediate conversational context

4.1.3 Identity Misrepresentation

Identity Misrepresentation is defined by statements or claims that are used to convey pejorative misrepresentations, stereotypes, and other insulting generalizations about protected or vulnerable populations. As with *Identity Attack*, protected groups are defined by attributes including age, disability, ethnicity, gender identity, military status, nationality, race, religion, and sexual orientation. Vulnerable groups such as those defined by immigration status, socio-economic class or the presence of a medical condition, may also be offered protection.

Statements belonging to this class fall below the severity of *Identity Attack*. They may be presented as fact but may lack supporting evidence or be opinions in disguise. Criteria outlined in the definition of *Identity Attack* belong uniquely to that class. For example, a stereotype suggesting that group of people is inferior (e.g. has low IQ) would fall under the definition of *Identity Attack*, whereas generalizations regarding food preferences (e.g. eats foods that are unappealing to others, without conveying explicit hatred towards the group), non-dehumanizing assumptions about physical appearance (e.g. wearing a style of facial hair implies support for extremism) or stereotypes about spending habits (e.g. frugality) should be treated as *Identity Misrepresentation*.

A summary of qualifying criteria for the *Identity Misrepresentation* class is as follows:

- Dissemination of negative stereotypes and generalizations about a protected or vulnerable group, apart from those that involve explicit dehumanization or claims of inferiority
- Statements about protected or vulnerable groups presented as declarative truth without supporting evidence
- Microaggressions, subtle expressions of bias towards a protected or vulnerable group
- Intent to spread fear of protected or vulnerable groups, without calls for violence

[9] As per the WOAH guidelines we use **** in place of any group identifier to avoid reproducing harm

Positive criteria defined for *Identity Attack* and *Insult* should be considered non-examples of *Identity Misrepresentation*.

4.1.4 Insult

Two datasets shared by Kaggle (2012, 2018) have provided guidelines used to determine whether or not content can be considered insulting. The latter defines *Insult* as: "insulting, inflammatory, or negative comment towards a person or a group of people." The earlier task provides more detail, whereupon insults are constrained to be person-to-person speech acts in which the target is assumed to be active in the conversation. This definition allows for the presence of profanity, racial slurs and other offensive terms. Using these specifications, it is not obvious how to distinguish between an *Insult* and an *Identity Attack* (which is also defined in the 2018 challenge). We propose an important distinction in that statements that make use of identity-based slurs and epithets are to be elevated to the level of *Identity Attack*.

With the assumption that the subject of an *Insult* is a participant in the conversation, an *Insult* is defined as:

- General name-calling, directed profanity and other insulting language or imagery not referencing membership in a protected group or otherwise meeting the criteria for *Identity Attack*
- Content mocking someone for their personality, opinions, character or emotional state
- Body shaming, attacks on physical appearance, or shaming related to sexual or romantic history
- Mocking someone due to their status as a survivor of assault or abuse
- Encouraging others to insult an individual
- Images manipulated with the intent to insult the subject

The following should be considered non-examples of insults:

- Insults strictly based on the target's membership in a group with protected status, including use of slurs
- Insults aimed at non-participant subjects, such as celebrities and other high-profile individuals
- Self-referential insults and self-deprecation
- Insults directed at inanimate objects
- Harassment education or awareness

With regard to the scale of severity, *Insults* are not elevated to the level of severe abuse, as they do not explicitly threaten physical safety, contain identity-based attacks, compromise personal privacy or involve criminal behavior.

4.1.5 Sexual Aggression

Various forms of sexual content are present online, such as pornography, nudity, and offers for adult services. Here we focus on a type of person-to-person abuse, *Sexual Aggression*. This type of content includes unwanted sexual advances, undesirable sexualization, non-consensual sharing of sexual content, and other forms of unsolicited sexual conversations. *Sexual Aggression* is defined as:

- Threats or descriptions of sexual activity, fantasy or non-consensual sex acts directed at an individual
- Unsolicited graphic descriptions of a person (including oneself) that are sexual in nature
- Unwanted sexualization, sexual advances or comments intended to sexually degrade an individual
- Solicitations or offers of non-commercial sexual interactions
- Unwanted requests for nude or sexually graphic images or videos
- Sharing of content depicting any person in a state of nudity or engaged in sexual activity created or shared without their permission, including fakes (e.g. revenge porn)
- Sharing of content revealing intimate parts of a person's body, even if clothed or in public, created or posted without their permission (e.g., "creepshots" or "upskirt" images)
- Sextortion, threat of exposing a person's intimate images, conversations or other intimate information

Sexual Aggression does not refer to:

- Pornography created with consent of all participants
- Solicitation or offers of commercial sex transactions
- Definitions of sexual terms
- Sexual health and wellness discussions
- Non-graphic use of words associated with sex
- Insults that make use of sexual terms
- Flirting, compliments, or come-ons that are not sexually graphic or degrading

Using our criteria for severe abuse, the following subset of content meeting the definition of *Sexual Aggression* is to be considered severe:

- Threats of non-consensual sexual activity
- Sharing of sexual content without consent from an involved party
- Demand for sexual activity
- Graphic and/or violent descriptions of sex

4.1.6 Threat of Violence

Many online platforms do not permit users to state a desire to kill or inflict physical harm towards others. Statements which celebrate, encourage or condone violent acts are also prohibited, as they may incite others to commit violent acts. *Threat of Violence* refers to content that contains at least one of the following:

- Desire to physically harm a person or group of people, including violent sexual acts
- Call for the death, serious injury or illness of a person or group of people
- Encouragement of another individual to commit self-harm or suicide
- Incitement to commit acts of violence
- Glorification of violence or violent events

Non-examples of this class include:

- Anecdotal or personal accounts of violence without glorification (e.g. survivor stories, criminal rehabilitation accounts)
- Historical descriptions or research studies of violence
- Hyperbolic or metaphorical violence

Taxonomies of violence (Straus et al., 1996) treat physical threats as more severe than verbal or psychological threats. While threats made via online communication are technically verbal, severe threats are those in which there is credible belief that the aggressor could or would carry out a threat physically. Severe forms of violent threats intend to do at least one of the following:

- Create the fear or belief that the violent act will occur in real life
- Threaten acts that result in serious consequences, such as a long-term injury or illness or fatality
- Convey the abuser's desire to carry out the threat personally

Mild forms of physically violent threats have at least one of the following features:

- The abuser's intent is to insult or dismiss the target, with little to no harmful consequences (e.g. "I could easily take you down")
- Threaten acts that result in minor or no lasting harm to a person's health or well-being, (e.g. "I'll slap you if you don't stop")
- Passive threats stated as wishes or hopes for an unfortunate event or illness to occur

4.2 Self-Inflicted Harm

Self-Inflicted Harm describes forms of harmful behavior, both physical and psychological, directed at one's self. The detection of content belonging to this class is intended to flag such behaviors in order to provide help to those in distress and prevent the spread of dangerous behavior within online communities. In the remainder of this section, we discuss definitions for two common forms of self-inflicted harm: *Eating Disorder Promotion* and *Self-Harm*.

4.2.1 Eating Disorder Promotion

Eating disorders (EDs) are mental disorders characterized by abnormal eating habits and attitudes towards food. Many online platforms explicitly prohibit pro-ED content in order to prevent the spread of unhealthy behavior. While the DSM-V offers clinical definitions of such disorders, here we summarize types of dangerous content related to the way eating disorders may be discussed online:

- Promotion of eating disorders as legitimate lifestyle choices (e.g. pro-ana, pro-mia content)
- Glorification of slim or emaciated bodies (e.g. thinspiration)
- Content featuring high-fat food or overweight people intended to induce disgust (e.g. reverse thinspiration)
- Sharing instructions for unhealthy weight loss methods

The following should be considered non-examples of *Eating Disorder Promotion*:

- Research, advocacy, and education related to eating disorders
- Discussion of recovery mechanisms and resources to prevent eating disorders
- Anecdotes of individuals who have suffered from eating disorders in a manner that does not glorify the disorder

Pro-ED content potentially creates long-lasting impact to one's health and therefore is considered to be severe.

4.2.2 Self-Harm

Self-Harm is a behavior in which a person purposefully physically hurts themself using methods such as cutting with a sharp object, burning, biting, and pulling out hair. Practitioners of such behavior do so in order to cope with emotional distress. While according to the DSM-V, people who exhibit self-harming behaviors do not intend to cause long-term, serious harm or fatality, suicide is an additional, albeit extremely different, form of self-harm, which we include in our definition.

Self-Harm includes the following content:

- Discussion of current or recent acts of deliberately harming one's own body.
- Suicidal ideation, discussing details of a suicide plan, or stating that one intends to commit suicide
- Requests for instructions on how to conduct or hide self-harm or suicide
- Describing emotions or symptoms of mental illness explicitly related to self-harm, or traumatic experiences and triggers
- Promotion of or assistance with self-harming behaviors

Self-Harm content does not refer to:

- Anecdotes of personal recovery, treatment
- Sharing coping methods for addressing thoughts of self-harm or suicide
- Support for individuals who are considering or are actively harming themselves
- Recollection of self-harming behaviors or suicidal attempts that occurred at least 12 months in the past that does not promote self-harm or suicide
- Research or education related to prevention of self-harm or suicide
- Discussion of depression or other mental illnesses, symptoms, or depressed thoughts and feelings that are not explicitly tied to self-harm or suicide

Identifying severe expressions of *Self-Harm* primarily rests on determining the individual's intent. An individual who is cutting or punching walls is doing so in order to help them cope with emotional pain. Suicidal individuals are not attempting to cope but rather responding to unbearable physical or emotional pain by ending their lives.

Severe forms of *Self-Harm* include:

- Suicidal ideation and planning
- Threatening to take action to kill, cut or otherwise hurt oneself
- Asking for or providing instructions or how to commit suicide or self-harm
- Positive reflections on death and dying or the perceived benefits of the individual's death

Less severe forms of *Self-Harm* include:

- Advice on hiding evidence of non-suicidal self-harm
- Showing off self-harm scars or positive reflections on self-harm behaviors

- Admitting to active or recent acts of non-suicidal self-harm

- Discussing events or objects that have recently "triggered" an individual to harm one's self

- Discussing reductions in recent non-suicidal self-harming behaviors without clear evidence of cessation

4.3 Ideological Harm

Ideological Harm describes the spread of beliefs that may lead to real world harm to society at large over time. Content belonging to this class may include statements without an explicit human target at the time of creation, for example, statements that openly question health or government policies that may lead to public crises, or expressions of praise for ideologies associated with crime, violence or exclusion. In this section, we present definitions of two common forms of ideological harm: *Extremism, Terrorism and Organized Crime* and *Misinformation*.

4.3.1 Extremism, Terrorism and Organized Crime

While to date there is no internationally agreed upon definition of terrorism, the UN General Assembly defines it as "criminal acts intended or calculated to provoke a state of terror in the public, a group of persons or particular persons for political purposes are in any circumstance unjustifiable, whatever the considerations of a political, philosophical, ideological, racial, ethnic, religious or any other nature that may be invoked to justify them." Various national governments and international organizations maintain lists of organizations they officially recognize as terrorist.

While terrorist groups are predominantly associated with violent behaviors, extremism refers to both violent and peaceful forms of expression. Organized crime groups, which frequently engage in violent criminal behavior, are not typically driven by political or ideological goals, but instead operate for economic gain.

Harmful content related to with *Extremism, Terrorism and Organized Crime* includes:

- Recruiting for a terrorist organization, extremist group or organized crime group

- Praise and promotion of organized crime, terrorist or extremist groups, or acts committed by such groups

- Assisting a terrorist organization, extremist group or organized crime group

- Content that includes symbols known to represent a terrorist organization, extremist group or organized crime group

At its core, every goal or belief of this class fits the criteria of severely abusive content. Either through exclusion, segregation, eradication or criminal activity, severe harm is intended.

White Supremacist Extremism One notable subtype of this type that we draw attention to is *White Supremacist Extremism* (*WSE*). The United States Congress recently identified white supremacist extremism as the most significant domestic terrorism threat facing the United States.[10]

WSE describes content seeking to revive and implement various ideologies of white supremacy. Content policies developed to address white supremacist ideologies are often established as part of a broader "hate speech" definition. While certain *WSE* statements attacking individuals based on religion, race or immigration status indeed overlap with our definition of *Identity Attack*, the motivation to elevate *WSE* to its own type of abuse is driven by a few factors. *WSE* content is often marked by various ideologies and linguistic patterns not expressed in direct person-to-person abuse. Attributes of the abuser are often in focus (e.g. whiteness and national identity), as opposed to characteristics of the abused. Additional features of *WSE* language include the use of dog whistle phrases and emoji, nostalgic references to "better times" in history, and the promotion of conspiracies and pseudo-science related to race, religion and sexuality.

WSE content can be generalized as belonging to one or more of the following ideologies:

- Neo-Nazism: idolization of Adolph Hitler, praise of Nazi policies or beliefs, use of Nazi symbols or slogans

- White racial supremacy: belief in white racial superiority, promotion of eugenics, incitement or allusions to a race war, concerns about "white genocide," cynicism towards interracial relationships and miscegenation

[10] https://www.congress.gov/116/bills/s894/BILLS-116s894is.xml

- White cultural supremacy: promotion of a white ethnostate, xenophobic attitudes, nostalgia for times of segregation

- Holocaust denial, propagation of Jewish conspiracy theories

- Recruitment or requests for financial support for WSE ideology, incitement of extreme physical fitness as a readiness measure for race-driven conflict

4.3.2 Misinformation

Simply stated, *Misinformation* is false or misleading information. It may be spread by users who are unaware of its credibility and lack a deliberate intent to harm. Disinformation, a subset of misinformation, refers to the knowing spread of misinformation. The intent behind disinformation is malicious, such as to damage the credibility of a person or organization, or to gain political or financial advantage.

Types of *Misinformation* include fake news, false rumors, conspiracy theories, hoaxes, and opinion spam. Increasingly more forms of misinformation are disallowed on many online platforms, including:

- Medically unproven health claims that create risk to public health and safety, including the promotion of false cures, incorrect information about public health or emergencies

- False or misleading content about members of protected or vulnerable groups

- False or misleading content that compromises the integrity of an election, or civic participation in an election

- Conspiracy theories

- Denial of a well-documented event

- Opinion spam, fabricated product reviews

- Removal of factual information with intent to erode trust or inflict harm, such as the omission of date, time or context

- Manipulation of visual or audio content with the intent to deceive

The spread of misinformation poses risks to society, erodes trust, hurts decision-making abilities, and may even lead to harmful global health or political events. *Misinformation* that may result in physical harm, civil unrest or health crises should be considered severe.

4.4 Exploitation

In order to provide safe online spaces, content created by users seeking to benefit by causing harm to others financially, sexually or physically is not permitted within digital communities. Forms of *Exploitation* include *Adult Sexual Services*, *Child Sexual Abuse Material* and *Scams*.

4.4.1 Adult Sexual Services

Certain forms of sexual solicitation and commerce cross over into illegal behavior that exploit often vulnerable participants, and are thus treated as a type of severe abuse. *Adult Sexual Services* includes:

- Promotion or solicitation of illegal sexual services such as prostitution, escort services, paid sexual fetish/domination services and sensual massages

- Organization of human trafficking

- Recruitment for live sex performances, sex chat

4.4.2 Child Sexual Abuse Material

Child Sexual Abuse Material (*CSAM*), sometimes referred to as “child pornography,” is defined as content involving sexual abuse and exploitation of anyone under the age of eighteen. Materials included in the definition of *CSAM* have expanded beyond sexual images involving minors to include exploitative text content as well. *CSAM* is a severe form of abuse that includes:

- Images and videos which depict minors in a pornographic, sexually suggestive, or sexually violent manner, including illustrated or digitally altered pornography that depicts minors (e.g. lolicon, shotacon, or cub)

- Sharing adult pornography or CSAM with a minor

- Grooming of minors (the development of relationships of trust with the intent to sexually exploit)

- Sexual remarks directed at minors

- Arranging real-world sexual encounters or direct solicitation of sexual material from a minor
- Providing advice for or advocacy of child sexual abuse

4.4.3 Scams

Online scams are attempts to trick a person into providing funds or sensitive information using deceptive or invasive techniques. The perpetrator of a scam may attempt to build insincere relationships over the course of a conversation or misrepresent themselves as someone with skill or authority. Types of *Scams* that are commonly prohibited from digital communities include:

- Attempts to trick users into sending money or sharing personal information (e.g. phishing)
- Promise of funds in return for a smaller initial payment via wire transfer, gift cards, or prepaid debit card (e.g. money-flipping)
- Offers promising cash or gifts, such as lottery scams
- Romantic and military impersonation
- Promises of debt relief or credit repair
- Recruitment into pyramid schemes

5 Conclusions and Future Work

Upon careful synthesis of content policies, human rights treaties and recommendations from experts in physical and psychological harm, we have presented a typology of harmful content along with a set of best practices for developing precise definitions of types. In the future, we plan to report on the impact of how the proposed definitions impact the quality of datasets and models built using them, and to share public datasets based on this typology that may be used by the research community. We have published the typology at `https://gitlab.com/sentropy-technologies/typology-of-online-harm` and encourage those who study online abuse to contribute.

Acknowledgments

We wish to thank Cindy Wang, Taylor Rhyne, Bertie Vidgen and the WOAH reviewers for providing detailed feedback on this work.

References

Internet and Jurisdiction Policy Network. 2019. Content and jurisdiction program: Operational approaches, norms, criteria, mechanisms.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. *Automatic Identification and Classification of Misogynistic Language on Twitter*, pages 57–64.

Susan Benesch. 2020. Proposals for Improved Regulation of Harmful Online Content.

Pete Burnan, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, page 75–84, New York, NY, USA. Association for Computing Machinery.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. 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)*, pages 1718–1728, Hong Kong, China. Association for Computational Linguistics.

Stevie Chancellor, Zhiyuan (Jerry) Lin, and Munmun De Choudhury. 2016. "this post will just get taken down": Characterizing removed pro-eating disorder social media content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1157–1162, New York, NY, USA. Association for Computing Machinery.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*.

Artur Dubrawski, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. 2015. Leveraging publicly available data to discern patterns of human trafficking activity. In *Journal of Human Trafficking*.

Kaggle. 2012. Detecting insults in social commentary.

Kaggle. 2018. Toxic comment classification challenge.

Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage.

Hee-Eun Lee, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. 2020. Detecting child sexual abuse material: A comprehensive survey". *Forensic Science International: Digital Investigation*, 34:301022.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *Proceedings of the Workshop on Natural Language Processing for ComputerMediated Communication*.

Anna Schmidt and Michael Wiegand. 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*, pages 1–10.

Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. 2017. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *Proceedings of the 2017 Internet Measurement Conference*, page 432–444, New York, NY, USA. Association for Computing Machinery.

Murray A. Straus, Sherry L. Hamby, Sue Boney-McCoy, and David B. Sugarman. 1996. The revised conflict tactics scales (cts2): Development and preliminary psychometric data. *Journal of Family Issues*, 17(3):283–316.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: An nlp perspective. *Natural Language Processing Research*, 1:1–13.

Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combating human trafficking with deep multimodal models. *CoRR*, abs/1705.02735.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 672–680.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *ArXiv*, abs/2004.01670.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. Association for Computational Linguistics.

Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International conference on web search and data mining*.

Zeerak Waseem. 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*, pages 138–142.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Women's Media Center. Online abuse 101.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.