A Safer Online Public Square

Jonathan Reeve, Colin Muller

Contents

Introduction

Perhaps the most challenging step in developing social or technological tools for promoting a 'safer online public square' is defining what sort of speech constitutes a threat to the civility and safety of members of online communities.

Where do we draw the line between off-color and offensive content and content that inflicts harm or hate upon an individual or group? Additionally, how can a third party moderator (or algorithm) gauge the context of the speech as well as the the subjective perception of the speech by both speaker and target of the speech?

Some critics of attempts to filter out harmful speech online claim that it violates first amendment free speech principles. These arguments vary, but generally point to the fear that content moderation may silence dissenting voices or unpopular opinions. On the other hand, defenders of content moderation assert that social media platforms have the right to remove content at their own discretion as they are not government agents who are held to first amendment standards. First amendment legal scholars have split the debate of free speech and content intermediaries between the right to speak and the right to hear/be heard , with some arguing that the mere right to speak is insufficient if media outlets are able to suppress certain speech on all platforms. Conversely, an individual's abilit to rapidly draft and publish a hateful tweet should require the target of the speech to view the harmful post.

Other concerns about content moderation include threats to anonymity and privacy, no platforming, silencing marginalized voices, amplifying hate speech by calling it out, and fixing virtual identites to legal identities through 'real name' policies.

There are many categories of speech that fall under the umbrella of harmful speech (The main categories are harassment, bullying, hate speech, and dangerous speech), but the boundaries between these different categories are neither rigid nor clear. The ambiguity posed by these various categories has led some to abandon formal definitions and seek different approaches to classifying harmful speech (see section below). Beyond this issue of classifying speech, there remains the problem of what one subjectively perceives as harmful speech; one study found only 33% overall agreement between students of different races asked to assess the degree to which tweets were racially offensive.

Statement of purpose of our preliminary report (to discuss at meeting)

Approaches to identification

Taxonomies and definitions

There is significant corpus of writing which attempts to formulate a definition and taxonomy of harmful speech online. These classifications will often vary based on the purpose of the definition which varies from academic research, legal recommendations, or advocacy. These different purposes result in varying breadth and scope of definitions, and disagreement over definitions is one of the main challenges in compiling data from different sources on the frequency of harmful speech online (as the specific behavior being monitored in different studies varies widely). Despite these challenges, it is useful to outline the different categories of harmful online behavior, as these different types manifest themselves differently and will require different approaches for intervention. Several organizations also have lists describing sub-categories and the different forms online abuse can take. The most exhaustive lits are from Women's Media Center and the Digital Rights Foundation.

Hate speech

Andrew Sellars conducted an overview of various attempts to define hate speech by academics, legal experts, and online platforms. In lieu of defining hate speech he identifies 8 emerging themes and common traits that are used in defining hate speech:

1. Targeting of a group, or individuals as a member of a group
2. Content in the message that expresses hatred
3. The speech causes a harm
4. The speaker intends harm or bad activity
5. The speech incites bad actions beyond the speech itself
6. The speech is either public or directed at a member of a the group
7. The context makes violent response possible
8. The speech has no redeeming purpose

Sellars points out that there are two divergent theoretical fields that deal with hate speech: free speech theory and critical race theory, and there is little overlap between the two.

Academic definitions often vary based on whether the goal is to apply the indention to legal sanctions or to simply understand the social phenomenon of hate speech. Some definitions focus more on the content of the speech, whereas others place more emphasis on intent of speech, and a third approach ignores the intention of the speaker and instead focuses on the outcome and effects of speech.

Alice Marwick and Ross Miller have synthesized these varying approaches, and proposed the following three general elements used to define hate speech:

1. Content-based elementary
2. Intent-based elementary
3. Harms-based element

Perhaps the clearest definition of hate speech is Susan Benesch's: - "An expression that denigrates or stigmatizes a person or people based on their membership of a group that is usually but not always immutable, such as an ethnic or religious group. Sometimes other groups, defined by disability or sexual orientation, for example, are included." Source: Benesch, Susan. Defining and diminishing hate speech. 2014.

The important difference between hate speech and other forms of harmful behavior online, is that hate speech is targeted at a group of people; even when it takes the form of hate speech against an individual it is targeting them for their belonging to a specific group

Bullying

Definitions of cyberbullying largely rely on definitions of offline bullying which are characterized by 3 factors: 1) psychological torment 2) repeated behavior 3) carried out with intent. Social Psychologist Robert Tokunaga defines cyberbullying as "any behavior performed through electronic or digital media by individuals or groups that repeatedly communicates hostile or aggressive messages intended to inflict harm or discomfort on others." While in traditional settings the factor of repeated behavior is more clear cut, there is some disagreement over what is considered repetition online, as a single post can be shared or otherwise interacted with repeatedly.

While the definition of offline and online bullying are nearly identical, (Tokunaga 2010) and others have identified several differences in the nature of offline and online bullying. Studies have found that individuals who don't engage in traditional bullying are more likely to do so online since there is a lower threat level of being caught and reprimanded. While schools have clear norm-enforcing agents, there are no clear authority figures online who will regulate behavior. Cyberbullying is therefore considered as a more "opportunistic offense." Additionally online bullying is less likely to occur at home, and bullies online are able to act anonymously and bully others they don't know in 'real life.' However, most cyberbullying is done against people the bully knows.

Findings from research on traditional bullying can also be informative for future studies on cyberbullying and approaches to intervention. One finding is that kids often don't identify with the term "bullying" often considering bullying acts to be "drama" or "beef." As a result in surveys that asked about cyberbullying, between 20-40% of youth reported being victimized, but in surveys that ask about "mean things" online, 70% of youth reported such experiences. This is

important to bear in mind when collecting survey data, but it also highlights another feature of bullying which is its cyclical nature. Teens often refer to bullying as drama which blurs the line between serious and nonserious conflict. This is because there is not always a clear distinction between categories of bully, victim, and bystander. In many surveys half of young respondents report being involved in both bullying and victimizaiton. This is referred to as "reactive" bullying, where a target of bullying retaliates with similar bullying behavior. It is important to bear the cyclical nature of bullying in mind when crafting intervention models.

Harassment and stalking

Online harassment is a term used to describe a variety of online behaviors that target an individual or group with the intention of harming them, getting them to remove content, or discouraging from expressing themselves online. Behaviors that constitute online harassment include "threats, continued hateful messages, doxxing, DDoS attacks, swatting, defamation, and more." According to HeartMob "what separates online harassment from healthy discourse is the focus on harm: including publishing personal information, sending threats with the intention to scare or harm, using discriminatory language against an individual, and even directly promoting harm against a person or organization." Source: HeartMob

Dangerous speech

Dangerous speech is a form of online abuse that has high stakes. Susan Benesch characterizes it as "speech that can inspire or catalyze intergroup violence." The speaker of dangerous speech often hold positions of power with influence over a large base. The audience dangerous speech is addressed to often hold "grievances and fear that the speaker can cultivate," and the speech plays on this in order to call on that group to act violently. Dangerous speech is most likely to occur in social or historical contexts that are already marked by or susceptible to violence, such as situations of civil war.

Source: Counterspeech on Twitter: A Field Study & http://www.worldpolicy.org/sites/default/files/Dangerous%2

Alternative approaches to classifying harfmul speech:

Implicit v. Explicit & Generalized v. Directed Abusive Language

Besides the approach of defining and taxonomizing different forms of hate speech, some have formulated other vectors for recognizing harmful speech online. One particularly useful approach is that developed by Waseem et al. Rather than attempting to define various terms such as hate speech, cyberbullying, and cyber harassment, they propose two primary vectors for typologizing abusive

language—Implicit vs. Explicit and Generalized vs. Directed abusive language. To classify abusive speech they recommend asking the following questions: Is the language directed towards a specific individual or entity or is it directed towards a generalized group? Is the abusive content explicit or implicit? Typologizing based on these two factors is useful as there can be a lot of overlap and ambiguity when relying on stricter definitions. However, one shortcoming of this typology is that by making the distinction between directed/generalized attacks, the research may downplay the fact that hate speech, even when verbally directed at an individual (ex: calling someone a racial slur) is an offense against an entire sub-group of people.

- Source: Waseem, Zeerak, et al. "Understanding Abuse: A Typology of Abusive Language Detection Subtasks." arXiv preprint arXiv:1705.09899 (2017).

What is left out of definitions

There are two forms of harmful speech that are often left out of discussions of harmful speech. The first is self-directed harmful speech and the second is harmful speech expressed privately between a group of like-minded people where the target of the harmful speech is not a party to the discussion. Neither is easily categorized with other forms of harmful speech and they therefore need separate attention.

Harmful speech directed at one's self

Self-directed harmful speech includes online posts that depict eating disorders or suggest self-harm. It is infrequently included in discussion of harmful online behavior, but it is a serious issue that needs to be dealt with in a different way than outward oriented harmful speech. ##### Harmful speech between a group that does not include the target of the speech

Harfmul speech between an "in-group" includes things like a private message channel where racists exchange racist messages about a group of people who are not a part of the message channel. Upon first consideration this speech may seem more like private thought than it does speech in the public realm (this is how some legal scholars frame it), but such conversation can have real life consequences on the well being of those discussed such as in private message forums where doxxing information is collected

Legal approaches

Justice Stewart's rule: "I know it when I see it"

Supreme Court Justice Stewart famously asserted "I know it when I see it" when referring to identifying obscenity. It seems to be the consensus that this approach

is not applicable to identifying hate speech due to the variety of forms of speech and contexts which one could identify as hate speech.

There are instances where specific epithets or insults are used and an outsider or scholar may see hate speech but the speaker/recipient do not. Henry Louis Gates Jr. for this reason asserted that we should not "spend more time worrying about speech codes than coded speech." Since a lot of hate speech can be coded or masked by symbols.

The discussion of Stewart's "I know it when I see it" points to a central difficulty in defining hate speech since it requires assessing the subjectivity and intention of both the perpetrator and the victim. However, only some definitions include the component of intention on the part of the perpetrator, and definitions also vary on how they define harm to the victim.

Brandenburg test

The 1969 supreme court case Brandenburg vs. Ohio narrowed the scope of unprotected speech under the first amendment. The court found that speech advocating illegal action is only prohibited when it is "directed to inciting or producing imminent lawless action and is likely to incite or produce such action." Prior to this ruling the wording was applicable to vaguer, more generalized advocacy of illegal action. The case revolved around whether or not a KKK leader saying "it's possible that there might have to be some revengeance [sic] taken" in a speech full of racist epithets constituted speech that intended to advocate violent illegal action. The Supreme court ended up deciding that the statement was protected by the first amendment since it was abstract advocacy of violence or illegal action. The court set higher standards for prohibiting speech that advocates for illegal action, requiring that it be likely that the illegal action actually occur in the near future.
What emerged from this case is the "Brandenburg test" which requires three elements for speech to be considered unprotected by the first amendment: intent, imminence, and likelihood. The court established that speech may be prohibited if it is: 1. "directed to inciting or producing imminent lawless action" 2. "likely to incite or produce such action" source: https://www.law.cornell.edu/wex/brandenburg_test

Elons v US

In 2015 the supreme court ruled on a case about Anthony Elonis who had posted rap lyrics on his facebook that violently threatened his ex-wife in graphic detail. In lower courts Elonis had been convicted of threatening a person over interstate lines, but the supreme court reversed this ruling. Their decision revolved around whether or not the standard of a "reasonable person" feeling threatened by the post was sufficient evidence to sentence Elonis. The supreme court found that there was also the necessary element of a "guilty mind" or "mens rea" on the

part of the speaker which they said was absent because Elonis had consistently held that he posted the threatening lyrics for therapeutic purposes. While this was the first time the Supreme Court heard a case related to speech on social media, the court largely avoided discussing first amendment related questions in the majority opinion. The judgment was only narrowly about whether or not mens rea was a requisite component of threatening a person.
In Samuel Alito's concurrence there is more discussion of First Amendment on social media. Alito distinguishes between rap lyrics and social media posts saying: "lyrics in songs that are performed for an audience or sold in recorded form are unlikely to be interpreted as a real threat to a real person." Whereas "Statements on social media that are pointedly directed at their victims, by contrast, are much more likely to be taken seriously."

Section 230 and Federal Discrimination Law

In the US Section 230 of the Communications Decency Act says that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." This rule shields social media platform moderators from getting in trouble for posts by others on their websites and removes their legal responsibility of moderating posts based on their legality.

Legal scholar Mary A Franks has called for section 230 of CDA to be amended so that it treats internet entities the same way it treats employers or school administrators under federal discrimination law. This means treating them as intermediaries who can exert control over those settings. This could be done simply by including explicit language in section 230 of the CDA about compliance with federal discrimination law, since the immunity that web hosts currently have is already limited by other federal laws such as copyright violations..

Other US Laws

Federal Criminal Law related to stalking criminalizes using the internet with "the intent to kill, injure, harass, intimidate, or place under surveillance with intent to kill, injure, harass, or intimidate another person" this includes placing a "person in reasonable fear of the death of or serious bodily injury" or causing, attempting to cause, "or would be reasonably expected to cause substantial emotional distress to a person"

Other laws people have tried using to prosecute online abuse includes defamation,copright (for non-consensual porn), harassment, coercion, menacing. Hate crime laws are hard to use because hate crime requires that another specific offense be comitted against someone that is motivated due to an immutable characterstic.

EU's Framework

The European Union has a "Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law." In 2016 a cooperative agreement was reached between private online platforms and the European Commission to regulate hate speech on the websites. The framework defines hate speech as containing one of the following three elements:

- "public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, colour, descent, religion or belief, or national or ethnic origin;
- the above-mentioned offence when carried out by the public dissemination or distribution of tracts, pictures or other material;
- publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes as defined in the Statute of the International Criminal Court (Articles 6, 7 and 8) and crimes defined in Article 6 of the Charter of the International Military Tribunal, when the conduct is carried out in a manner likely to incite violence or hatred against such a group or a member of such a group."

In Germany big social media platforms are facing fines for failing to take down content that violates their laws fast enough.

Industry approaches

Most social media sites have a bifurcated approach to presenting their policy on harmful speech to their users. On the one hand, their formal documents such as the terms of service tend to broadly prohibit harassment on the site, but these formal documents rarely elaborate on what constitutes harassment. On the other hand, the websites' informal documentation such as 'community guidelines' will go into more detail about what constitutes misconduct on the website, but falls short of formally defining harassment, giving the websites lee-way to determine what sort of content they remove. In both the formal and informal policy of social media platforms, harassment is usually lumped together with other prohibited activity such as spamming and hacking.

Source: Pater, Jessica Annette, et al. "Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms." GROUP. 2016.

Many have pointed to the gap between the publicly available policy on harfmul speech and the internal guidelines that websites provide to their content moderators (this is largely thanks to leaked internal documents).These internal guidelines are often very detailed, but they are considered arbitrary by critics who argue that the strict criteria results in instances of harmless speech getting removed while some actually harmful speech is left unmoderated. An often cited example of this is Facebook's leaked rules which classified white men as a protected category but not black children. These rules allowed offensive posts targeting black children to stay online and also resulted in the removing of posts

by anti-racist activists that addressed white men. Additionally, there is a gap between the written rules (internal or public) and the actual practices of content moderators who are often given only a few seconds to decide if a post should be removed.

See Appendix 2 for more details.

Statistics

A 2014 Pew Research Center survey found that 73% of adult Internet users have witnessed harassment online, and 40% have experienced it personally (Duggan 2014). A later 2016 report by the Data and Society Research Institute claims that, of American Internet users, 72% have witnessed harassment or abuse, and almost half (47%) have personally experienced it (Lenhart et al. 2016).

The types of harassment include calling of names (reported by 60% of witnesses) purposeful embarrassment (53%), physical threats (25%), sexual harassment (19%), and stalking (18%) (Duggan 2014). Of those who had personally experienced harassment, 8% had been physically threatened or stalked, and 6% had been sexually harassed. These typically took place on social media platforms, although also in comments sections or in multiplayer games.

The Pew report finds that young women, aged 18-24, are disproportionately targeted in all categories except for those of purposeful embarrassment and the calling of offensive names. A 2013 report by the WHOA organization (Working to Halt Online Abuse) echoes this finding. In their analysis of 4,043 self-reported cases of abuse in American from 2000-2013, they find that 70% of victims were female, with a 42% majority between the ages of 18 and 30. The abusers, they find, are more likely to be men (47%) than women (30%). LGB Internet users, as well, are more likely to experience harassment (Lenhart et al. 2016, 37).

Possible social-psychological causes of harmful behavior online

While theory building on the underlying causes of harmful speech online is generally underdeveloped and often not rigorously proven with empirical research, (Tokunaga 2010), existing social-psychological theories helps us better determine effective forms of intervention. This vein of research is particularly useful in emphasizing the impact that an individual's harmful speech can have on their social group (including those not directly targeted by the speech), since "cyber-bystanders" witnessing of abusive language online impacts their understanding of acceptable online norms. This research then reminds us that when intervening in the name of a safer online public sphere, it is not only the speaker and recipient that we must pay attention to, but also those bystanders and digital onlookers who may also happen to witness the encounter.

John Suler has theorized the "online disinhibition effect" whereby online users compartmentalize their "online self" and "real life self," and the normal cognitive processes that guide their "real life" behavior are suspended when they are online. Online disinhibition is a result of anonymity, asynchronous communication, and empathy deficit (Suler 2004).

Similar to Suler's theory, the Barlett and Gentile Model argues that individuals' likelihood of engaging in bullying is a product of their attitudes towards cyberbullying and their perceived anonymity (Barlett and Gentile 2012). Users who realize that they have anonymity online will dissociate their online actions from their "real" self which will contribute to positive cyberbullying attitude and increased cyberbullying. The model hypothesizes that perceived anonymity gives potential cyberbullies a sense of impunity online as well as empowerment since they can attack individuals they would not be able to attack offline (either because of differences in physical strength or because they do not actually know that person offline). Perceived anonymity leads cyberbullies to distance themselves from their own actions and feeds a more positive perception of cyberbullying thus creating a feedback loop that encourages further cyberbullying.

Barlett and Gentile propose intervening against cyberbullying by informing "Internet users that they are not anonymous" through techniques such as showing "them evidence of IP address tracking and how History folders operate, then perhaps cyberbullying will decrease" (178) (Barlett, Gentile, and Chew 2016). However this form of intervention would not be effective in cases where users are actually able to be anonymous online (such as by using a mix network architecture) and could even embolden them. This points to a limitability to Suler, Barlett and Gentile's theories that argue that anonymity contributes to harmful online behavior. Both theories have ambiguous definitions of anonymity that do not distinguish between pseudonymity, perceived anonymity, or partial anonymity.

Suler, Barlett and Gentile's theories suggest that a user online undergoes a sort of deindividuation, where they experience a loss of their sense of individuality or sense of self. Other researchers have pushed back against the idea that online activity leads to deinidivuation and instead have put forward the notion of depersonalization as a better explanation that does not pathologize all online behavior as decidedly negative. Depersonalization is defined as a "process through which individuals perceive that their certain group identity is more salient than other identities in a particular context, termed as 'the emergence of group in the self' (Huang 399) (Huang and Li 2016). The principle difference is that deindividuation implies a loss of rationality that leads to necessarily harmful results whereas depersonalization is stripped of some of the value judgments and provides an explanation for both positive and negative group behavior online. Huang et al in their study ("The effect of anonymity on conformity to group norms in online contexts: a meta-analysis")[http://ijoc.org/index.php/ijoc/article/view/4037] put forward an argument that online anonymity results in depersonalization and group conformity online (Huang and Li 2016).

The deindividuation v. depersonalization debate can be traced to debates over crowd psychology that date as far back as Gustave Le Bon's 1895 The Crowd: A Study of the Popular Mind. Le Bon's work was followed by a corpus of discussions about a variety of factors including agency, control, reasoning, and emotional manipulation of (crowds)[https://logicmag.io/01-the-madness-of-the-crowd/].

By understanding how anonymity online can lead to greater group conformity, we can see how this dynamic can be used to both combat and contribute to harmful speech online. If an individual is part of an online community where harmful speech is perceived as unacceptable and uncommon, then they are less likely themselves to engage in harmful speech. This is referred to as a mimicry effect, and the theory holds that boosting positive content will foster an environment where other members of the website will contribute positive content

A clear example of depersonalizaiton and mimicry effect being mobilized to combat harmful speech online is the website (HeartMob)[https://iheartmob.org/about] which allows for online bystander intervention and provides an immediate support group for people who have faced online harassment. Another example is Instagram's "kind comments" where the company has encouraged users to write kind comments on posts to foster an environment that encourages conformity to positive norms

Studies in temporal clustering demonstrate the role of peer-influence on harmful speech. Studies have shown that online abuse is often clustered around the time of a certain event, such as in reaction to a heavily mediatized crime. They are especially likely to snowball in instances where there is a viral hashtag. Hashtags that have been associated with prejudicial crimes include #killallmuslims #ferguson #charliehebdo #brussels #banislam #baltimore #mizzou. Temporal clustering research also teaches us that some efforts to mitigate harmful speech are very time sensitive/specific (King and Sutton 2013).

Political psychologists Nyhan and Reifler have researched the "backfire effect" where attempts at correcting misconceptions or misinformed beliefs results in firmer beliefs in the misconception (Nyhan and Reifler 2010). It is important to consider the backfire effect in looking at effective forms of intervention/counterspeech, since often presenting facts or engaging in logical/reasoned debate is not the most effective strategy. Rather than being swayed be reasoned fact-based challenges to their views, individuals will become further entrenched in their views and even incorporate the contrary speech into their world view. A similar psychological phenomenon is "motivated reasoning" where people make a strong effort to support the conclusions they seek despite being exposed to contradictory facts

Approaches to Intervention and Reporting

Organizations and Advocacy Groups

A number of organizations exist to study and combat harassment, hate speech, and related phenomena. There are a number of approaches taken by these group including education, legal or psychological resources, group support, legal aid, and direct action. Most organizations employ a variety of these approaches.

Most groups publish educational material including studies and reports on online harassment's different forms, its adverse effects, and ways to mitigate such speech. Such organizations include the Women's Media Center's Speech Project and Take Back the Tech. Additionally, the Berkman Klein Center for Internet and Society at Harvard is in its third year of its "Harmful Speech Online" project which has yielded several valuable publications. Groups also provide education on how to secure online identities to prevent or limit online harassment—this includes employing cybersecuirty measures to prevent hacking, doxxing, and leaking of nonconsensual pornography. Such groups include HackBlossom a DIY feminist cybersecurity group and HeartMob's Technical Safety Guide. Heart mob also has detailed Social Media Safety Guides for Facebook, Twitter, Reddit, Tumblr, and Youtube. They also have an extensive Supportive Organizations page.

Organizations also publish legal and psychological resources for individuals trying to cope with and mitigate the effects of harassment. These resources include the anti-harassment non-profit Without My Consent's page which outlines laws in all 50 states and federal level that could be used to prosecute various forms of online harassment. Without My Consent's 50 State Project also summarizes lawsuits related to those laws and their outcomes.

Beyond providing resources for targets of harassment, some groups have professional or amateur support networks. Online SOS and Cybersmile for example has professional support that provides crisis coaching that helps people document and cope with harassment and seek legal and/or psychological counsel. Similarly, human rights activist Nighat Dad founded the Digital Rights Foundation in Pakistan which has a cyberharassment helpline that provides support and walks callers through the steps of filing legal complaints—they published a report outlining the nature of the calls they received in 2016-17. Instead of solely providing professional advice, Heartmob (established by anti-harassment group Hollaback) is a platform that creates a community of individuals who provide support for others targeted by harassment. Heartmob's site allows people to tell their story and receive "understanding messages, helpful resources, and practical assistance." The Cyber Civil Rights Initiative conducts legal research on online harassment and even provides model legislation for drafting "revenge porn" laws.

A slightly more controversial approach is taken by popular social media accounts that call out online abuse. This includes the twitter and instagram accounts Yes

You're Racist, Yes You're Sexist, and Bye Felipe. These accounts retweet or post screen shots of racist or sexist tweets, sometimes with the name of the account name of the person being called out. While some are critical of this approach in the name of protecting the person being amplified by these accounts, most of the things they are reposting were already posted online publicly. Even more controversial is the Tumblr blog "Racists Getting Fired" which posts screenshots of racist posts made on social media along with the poster's employer's info so that people contact the company to get them fired. The blog came under heat from the Washington Post for being a form of undiscerning internet vigilantism that on at least one occasion has gotten innocent people fired. The blog has a page where they outline their successes.

Inoculation is a long-term method for fighting against hate speech that takes some time. It involves instilling values in a society that oppose hate speech, and deals especially with building the social-psychological tools necessary so that groups of people don't fall victim to the pressures of engaging in hate speech or being incited by it. An example of a group that deals with Inoculation is Radio la Benevolencija (RLB) a dutch nonprofit that produces entertainment for countries in central africa that deals with the psychology underlying incitement to hate and violence.Citron and Norton suggest that internet intermediaries and society at large (especially public schools) play a stronger role in fostering digital citizenship in attempt to inncoulate hate speech.

For an extensive list of organizaitons dealing with online abuse see Heartmob's Supportive Organizations resource.

Initiatives by Social Media Platforms

Most social media platforms have similar approaches to removing abusive content, relying on a mix of flagging by users and detection software. As discussed above, most sites have vaguely worded terms of service and community standards that prohibit harmful speech without explicitly defining it, and once a post is flagged moderators determine whether or not to delete a post based on private, internal regulation.

When sites decide not to remove the content, they still provide users with the options of blocking, unfriending, or muting the poster or hiding the individual post/comment. Beyond these options, sites have developed advanced and creative approaches that are available to users.

Twitter launched a feature where users can export lists of people they've blocked and share them this list with friends who wish to block the same people. Twitter explained that the feature is for "those who experience high volumes of unwanted interactions on Twitter [and] need more sophisticated tools. That's where this new feature comes in. You can now export and share your block lists with people in your community facing similar issues or import another user's list

into your own account and block multiple accounts all at once, instead of blocking them individually."](https://blog.twitter.com/official/en_us/a/2015/sharing-block-lists-to-help-make-twitter-safer.html).

Instagram introduced an option where users can "Enable Offensive Comment Filter". Users can also create a customized list of keywords they want filtered out of posts (this includes emojis). Users can also choose to use Instagram's default keywords. Instagram also has a feature which allows users to anonymously report friends' posts that suggest self-harm. Once they are reported Instagram sends the reported user a message of support that includes a phone number to a help line.

Instagram (along with Tumblr) seem to be the most proactive in fostering a safe and kind environment on their social media websites. While other websites emphasize their commitment to free expression, instagram instead emphasizes their desire for the platform to be safe, kind, and inclusive. Therefore they don't shy away from asserting their right as a platform to remove or boost certain posts. Instagram launched a #KindComments campaign in 2017 where they encouraged users to generate kind content. The campaign also involved comments identified as kind being made more visible to other users. Instagram and Tumblr both seem to also be more proactive about connecting their users with mental health and anti-bullying resources.

It is interesting to note that while Facebook owns instagram, many of the initiatives launched on instagram have not been transferred to Facebook. This is at least partially due to Instagram's younger user base.

In response to what the far-right characterizes as an assault on free speech, the social media website Gab.ai was founded with the purpose of creating "Free Speech for Everyone." Their logo is a frog, reminiscent of the alt-right icon Pepe the frog, and they have primarily drawn in far right members with a good deal of racist, sexist, and xenophobic content. It has been described as the far right's "digital safe space"

For a more granular look into approaches by platforms, HeartMob has introduced detailed Social Media Safety Guides for 5 of the most used social media platforms (Twitter, Facebook, Tumblr, Reddit, and Youtube). They worked closely with staff from all five platforms in preparing the reports.

Counterspeech

Counterspeech is speech that directly addresses the perpetrator of hate speech or the cyberbystanders who sees the harmful speech occurring online. For hateful speakers with deeply ingrained hate, counterspeech is less effective, but it can still alter their discourse to be less overtly hateful which in itself can be seen as positive. The guiding principle of counterspeech is the liberal ideal that more speech is the best remedy to harmful speech. This is partially informed by the

idea that if you delete hateful content from one platform that there is always somewhere else to go. It is important to remember that counterspeech does not just target the speaker, but also the bystanders whose discourse norms are shaped by seeing hate speech go uncontested.

There are specific recommended and discouraged approaches to responding to hateful speech. Some of the discouraged approaches risk putting the target of the speech in greater danger.

The following summary of counterspeech tactics draws largely on the research by Susan Benesch and her colleagues. Notably their works:

- Counterspeech on Twitter: A Field Study, by Susan Benesch, Derek Ruths, Kelly Dillon, Haji Mohammad Saleem, and Lucas Wright
- Considerations for Successful Counterspeech, by Susan Benesch, Derek Ruths, Kelly Dillon, Haji Mohammad Saleem, and Lucas Wright

Recommended Counterspeech Strategies

Warning of consequences

Counterspeakers are encouraged to remind the hateful speaker of the harm done by their speech. They should also remind them of the long term consequences of their speech which (may be permanently stored in a company's data center) as well as immediate online consequences (blocking, reporting, suspended account). The strategy of just warning the speaker of the consequences of their speech is primarily effective at getting the speaker to delete their posts, but it does not necessarily change the speakers point of view.

Shaming and Labeling

Shaming and labeling the speech as bigoted, misogynistic, or otherwise hateful can be effective. However, it is important that the shaming and labeling is targeted a the speech and not the speaker (ex: say "saying X sounds racist" instead of "you're racist"). Labeling speech as harmful can also be helpful to bystanders and the person targeted. Also, in some cases the speaker 'may not have known' the harm in what they were saying so it is better not to make a personal attack and to only to label the speech has harmful.

Empathy and Affiliation

Counterspeakers are encouraged to change the tone of the discussion to friendly or empathetic. They are also encouraged to establish common ground that leads the harmful speaker to identify with them (ex: "I am also a conservative" or "I am also a white man"). It is also effective for the counterspeaker to identify with the targeted group (ex: "what you said hurt me as a Muslim").

Humor

In order to neutralize dangerous or intimidating speech it can be useful to deploy a humorous message. This can attract a larger audience in support of the counterspeech and also soften a message of counterspeech that could otherwise come off as hostile or aggressive (ex: it's time to stop posting cat)

Images

Similarly to the humorous approach to counterspeech, using images can make counterspeech go viral. Furthermore, counterspeech is much more effective when the counterspeaker appeals to emotions instead of trying to engage in a rational/logical argument. Therefore images can be a good way to "send people along emotive pathways"

Discouraged Counterspeech Strategies

Using a hostile, aggressive, or insulting tone can worsen a situation and cause the speaker to strike back with more harmful speech.

As was mentioned, Fact checking or attempting to engage in a reasoned debate is unlikely to influence the original speaker (see discussion of "backfire effect" and "motivated reasoning" above in the "Possible social-psychological causes" section). Often when challenged with facts, people will dig their heels in and become even more firmly committed to their views. However, fact-checking and pointing out hypocrisy can be influential on bystanders.

Successful counterspeech is indicated by:

- Speaker shifts their discourse if not also their beliefs
- Speaker apologizes, recants, or deletes original hate speech
- Discourse norms of the cyberbystanders are positively affected
- Hate speech narrative is delegitimized (even if speaker is not swayed)
- More counterspeech is elicited from the audience (this is good until it turns into harassment/dogpiling)

Typology of counterspeech

A useful method of typologizing harmful speech online is by distinguishing between the types of exchanges (vectors). (This is based on the models put forward in Counterspeech on Twitter: A Field Study) - One-to-one: one person deploying counterspeech against one person's hate speech - One-to-many: one person deploying counterspeech against many people's hate speech - Many-to-one: many people deploying counterspeech against one person's hate speech - Many-to-many: many people deploying counterspeech against many people's hate speech

Kevin Munger's Study as an example of counterspeech

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Kevin Munger conducted a study where he identified 231 twitter accounts operated by white men that regularly used racial slurs against black twitter accounts. Munger created a variety of twitter 'bots' with different amounts of followers, and some with a white male avatar and others with a black male avatar. The bots would send the following tweet in response to detected racist tweets by the identified twitter accounts: "@[subject] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language."

Munger found that accounts confronted by the white male twitter bot with a lot of followers were most likely to alter their language in future posts. Around 27% of users stopped using the n-word in their posts the following weeks. Munger's findings corroborate existing theories on counterspeech that tell us that counterspeech is most effective when done by someone with whom the harmful speaker identifies (in this case white males identify with other white men). On the other hand, Munger's findings cut against existing theories about how anonymity affects online behavior; the study found that accounts with personal information (name, photo, location, etc.) that used the n-word were significantly less likely to change their behavior after being rebuked by the bot, and in some cases would increase their usage. It was primarily accounts with little or no personal information that would redress their behavior. This goes against the theory that anonymity contributes to online hate speech, since those that were most anonymous were the most likely to alter their behavior.

Databases and Datasets

Some organizations maintain structured databases of abuse. The No Hate Speech Movement's Hate Speech Watch database is a database for user-submitted reports of hate speech on the Internet. The database contains descriptions of websites containing hate speech, social media posts, and abusive users on social media. The website Trolldor: the global blacklist of twitter trolls similarly collects data on Twitter trolls, and maintains a "top 10 worldwide trolls" list. Hatebase.org, which bills itself as the "world's largest online repository of structured, multilingual, usage-based hate speech," is a database of terms or phrases that its users report as hate speech. Although not all of these databases seem to have public APIs, if their data might nonetheless be used to train language or metadata categorizers.

Public datasets also exist which might be used as training data for categorizers. Many authors of papers in this area of computational linguistics also release their training data, much of which is manually labeled (Kolhatkar and Taboada 2017, @ott_finding_2011, @samghabadi_detecting_2017, @waseem_hateful_2016, @wulczyn_ex_2017). For an analysis of the Gamergate controversy, Phillip Polefrone collected a dataset of roughly one million tweets. The 2012 Kaggle

task provides a CSV with hand-labeled abusive short messages.The Wiki DeTox project of the Wikimedia Foundation provides language from their Wikipedia Talk Pages, human-annotated for "toxicity" and "aggression." A separate dataset of theirs provides personal attack annotations of over 100K Wikipedia comments, manually annotated by about 4,000 annotators.

Finally, organizations such as Bing and Google maintain public "bad word lists." Shutterstock, for instance, unofficially maintains a "List of Naughty, Obscene, and Otherwise Bad Words" in 25 languages, including Esperanto and Klingon. These lists are currently used to filter and automatically replace offensive language, and might be used to inform feature vectors for categorizers.

Organizations and Projects Employing Machine Learning

A few projects and institutions already use a machine learning categorization techniques to identify abusive language. Arguably the most well-known of these is the Google incubator Jigsaw, whose self-described mission is to "tackle some of the toughest global security challenges facing the world today—from thwarting online censorship to mitigating the threats from digital attacks to countering violent extremism to protecting people from online harassment." Their project Perspective provides an API for querying a "toxicity" language model, one trained on data manually labeled on a spectrum from "very healthy" to "very toxic." They define "toxic" as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." Much of their code is freely available online, on GitHub. Among these code repositories are moderator support programs, that use the Perspective API and automatically label contributions for toxicity on comment platforms such as Reddit, WordPress, and Discourse. Those that have tested the API, however, report mixed results, a fact which has led one commentator to posit that the AI "mistakes civility for decency".

The related Wikimedia Foundation project Detox publishes previously mentioned public data sets, and also provides an API for querying two of their models: of personal attacks and aggressive tone (Wulczyn, Thain, and Dixon 2017). Our tests of the API, like those of Perspective above, show problems handling ambiguity: the threatening phrase "be careful, you might find some white powder in an envelope come in the mail one day" was rated as 1% aggressive, while the more playfully celebratory "I'm going to nominate you for the Nobel prize, you brilliant man" was rated 61% aggressive.

The Mozilla-funded Coral Project publishes a comment moderation assistance tool called Talk, in use by the New York Times, in which language containing banned words is rejected, and suspect language is flagged or highlighted to aid moderators. It is unclear whether the project uses machine learning techniques or simply a dictionary-based approach, however.

More proprietary systems also exist. A number of US patents—US5796948, US8868408, US8473443, US7818764, and US20080109214—describe "offensive

message interceptors," "methods for word offensiveness processing," "inappropriate content detection," "methods for monitoring blocked content," and "methods for computerized psychological content analysis," respectively.

Computational Detection of Abusive Language, Behaviors, or People

General Classification Studies

Most computer science and computational linguistics studies in the detection of abusive language treat the problem as one of document categorization. This is a very common pattern of machine learning analysis in which an algorithm decides whether to place documents in one of two or more categories. These categories are almost never identically defined across studies, but are variously termed "abusive / non-abusive," "hate speech / non-hate speech," "high-quality / low-quality," and so on.

Almost all of the categorization experiments we examined employ some form of supervised machine learning. Supervised learning approaches involve training a machine learning algorithm on a set of pre-labeled training data, such as tweets or comments that human annotators have labeled as "abusive" or "non-abusive." The algorithm learns which textual or metatextual features best correlate with their categories, and then uses these weighted features to predict the probability of an unlabeled document falling into a category. The studies we examined vary greatly in their approaches: the features they choose, the classification algorithms they use, and the categories themselves.

There are two main categories of features used in these studies: textual and metatextual features. Textual features are properties of the abusive language itself, without reference to their context; metatextual features are contextual: users' self-descriptions, IP addresses, and operating systems, just to name a few. Some document categorization studies employ only textual features (which, in rare cases, are the only features available), but most employ a mix of both.

Textual features are either words, characters, or groupings of words or characters known as n-grams, where n refers to the number of words or characters. Character 4-grams, for instance, refers to groupings of four characters, and word trigrams refers to groups of three words. Skip-trigrams refers to groups of words representing every other, or every third word in a series.

Choosing what constitutes a word, and defining its boundaries, is an important first step in the creation of textual features called tokenization. This may seem like an inconsequential first step in text analysis, and is for this reason often considered *pre*-processing, but careful tokenization, informed by domain-specific knowledge and familiarity with the textual corpus, can make significant differences in the

outcomes of the categorization experiment. A related task, termed normalization, involves transforming words into their canonical morphologies, either through lemmatization (collapsing words into their dictionary forms, i.e. "went" to "ran"), or stemming (transforming word variants into their stems, i.e. collapsing "translate" and "translation" into "translat").

Normalization often also involves domain- or platform-specific transformations. Treating emoji as words, for instance, or transforming them into words with sentiment valence, often greatly influences the outcomes of sentiment analysis. (See, for example, (Castillo, Mendoza, and Poblete 2013, samghabadi_detecting_2017)). Collapsing obfuscated words, like transforming "w o r d" into "word," and "ta$k into"task," also proved to be useful to participants in the Kaggle contest, Detecting Insults in Social Commentary. Related normalization techniques involve expanding abbreviations such as "r u" into "are you," and collapsing long vowel representations, transforming "cooooooool" into "cool."

The results of this tokenization and normalization are language features in the form of n-grams. These features are then vectorized and weighted using a variety of techniques. In some cases, binary representations of words are used (either 1 or 0 for the presence or absence of a word in a document, see (Sood, Churchill, and Antin 2012)), but more frequently, term frequencies are used (ratios of the words in each document), and even more frequently, TF-IDF, or term frequencies adjusted for inverse document frequencies (see (Samghabadi et al. 2017, @diakopoulos_editors_2015)). These vectorizations are often used in statistical studies of linguistic style (stylometry) and are sometimes considered proxies for the stylistic fingerprints of individual or authorial voices. However, the limitation of these vectors is that they require a critical mass of text (>500 words) for term frequencies to be statistically practical.

Other vectorization techniques take into account the meanings or functions of the words. Word embeddings, for instance, transform words into high-dimensional vectors that encode the probabilities of their co-occuring with other words in a large corpus. Embeddings from the Stanford GloVe vectors, for instance, encode semantic information into high-dimensional vectors.

To this collection of token vectors, often other language measurements are added. General measurements such as document length are usually among these features. More specific measurements may include ratios of capital letters (a proxy for all-caps emphasis) and ratios of punctuation marks such as exclamation points. (The features used by the 2014 Stanford Literary Lab Pamphlet 7, "Loudness in the Novel", are similar, and are used as proxies for what they term "loudness.")

Once these features are constructed, the classification task itself may begin, which will infer categories (abusive/non-abusive, high/low quality, and so on) using the features. A wide variety of classification algorithms are used in these experiments. In many, Support Vector Machines (SVM) are used, as in (Siersdorfer et al. 2010, samghabadi_detecting_2017) and many of the entries of the 2012 Kaggle

task. In others, Long Short-Term Memory (LSTM) categorizers are used, as in (Kolhatkar and Taboada 2017), or Convolutional Neural Networks, as in (Gambäck and Sikdar 2017). Importantly, the categorizers that perform best in these experiments seem to vary greatly according to the data set and domain. It is for this reason that some of the most successful experiments in the Kaggle task used meta-categorizers. Meta-categorizers test the efficacy of a number of categorization algorithms, making the choice of categorization algorithm one of the tasks of the categorization experiment. Many of the Kaggle entrants, for example, used cross-validation grid searches (used also for parameter tuning), Random Forest regressors, or other stack-regressors.

Detection of Quality, Formality

Some related content-based approaches to language categorization include the identification of language quality. "Quality" can refer to the subjective usefulness of speech towards the goals of a particular website or online community, grammatical quality (correctness), credibility, or formality, among other definitions. (Siersdorfer et al. 2010) define quality as YouTube comments with good feedback (high numbers of user upvotes), and construct a categorization experiment where 6.1M training comments are vectorized using the most distinctive words of each category (TF-IDF), their SentiWordNet sentiment synonyms, and then classified using support vector machine classifiers. (Agichtein et al. 2008) also define quality based on user-reported and metadata-based reputation scores, like PageRank and ExpertiseRank, in an experiment categorizing Yahoo Answers conteng. They test a number of features of each answer, including n-grams of length 1-5, their POS representations, and metadata such as number of clicks, and categorize these using stochastic gradient boosted trees.

A related metric is language formality. (Heylighen and Dewaele 2002) propose a formality-score based on proportions of parts of speech. Based on the anthropologist Edward T. Hall's concept of "high-context" and "low-context" speech, and on the linguistic concept of deixis, they divide their lexicon into deictic words, such as pronouns, adjectives, and interjections, and non-deictic words, such as nouns and adjectives. The formality score is then proportional to the sum of deictic word frequencies subtracted from the sum of non-deictic word frequencies. This score is used, though not successfully, in a categorization experiment in (Agichtein et al. 2008), who also use a variety of other language quality metrics, like grammatical correctness. Similar measurements, like spelling, uppercase frequency, and lexical entropy, are used for comment classification in (Brand and Van Der Merwe 2014).

The presence of profanity has also been shown to correlate with abusive language, as in (Brand and Van Der Merwe 2014). Many of the top entries in the Kaggle contest, for instance use a "bad words list" as a seed feature set. (Samghabadi et al. 2017) uses Google's bad words dictionary, and combines it with a list from another researcher. "Bad" words themseles, though, are of course never

unambiguously bad. (Kapoor 2016), for instance, describes an attempt to differentiate between "casual" and abusive swearing. They find that the high-severity swears are likely to occur in abusive contexts.

Sentiment Analysis

Another useful language analysis, and one which often provides an additional categorization feature, is sentiment analysis. Now a veritable subfield of computational linguistics, and digital humanities more generally, sentiment analysis has its roots in industry NLP applications, where it is used, for example, to predict stock prices based on sentiments conveyed about companies in news media. Traditional approaches to sentiment analysis use lexical approaches: words with negative or positive valence (like "awful" or "amazing") are encoded as such in a lexicon such as SentiWordNet, , and this lexicon is then used to compute the overall sentiment for a given text. This sentiment is usually expressed numerically as two measurements: "PN-polarity" and "SO-polarity," for positive-negative sentiment, and subjective-objective (Baccianella, Esuli, and Sebastiani 2010). A word with very high positive sentiment associations might be, and which is highly subjective, might be encoded as (0.8, 0.8), for instance, and a negative-objective word might be encoded as (-0.8, -0.8).

Sentiment analysis has seen some success in categorization experiments. (Siersdorfer et al. 2010) find that sentiment scores, computed using the SentiWordNet, correlate with user ratings of comments on YouTube. (Castillo, Mendoza, and Poblete 2013) find sentiment scores to be among the best features that distinguish between "credible" and "non-credible" tweets. There are limits to lexical sentiment analysis, however, even among those studies that distinguish between homographs using word sense disambiguation. (Brand and Van Der Merwe 2014) test a number of features, including sentiment scores, and find that content- and quality-based features perform much better at categorization than sentiment. It has often been noted, as (Sood, Churchill, and Antin 2012) note, that "sentiment analysis is, in addition to being author, context, and community-specific, a domain-specific problem" (3). Futhermore, in psycholinguistics, an 1994 literature review of the "language of psychopathy" finds that one of the linguistic features of these patients is the lack of emotional markers (Rieber and Vetter 1994). This would suggest that sentiment analysis might not be effective in detecting abusive language from speakers with mental illness.

More recent approaches to sentiment analysis attempt to mitigate the domain problem of lexical approaches by using probabilistic methods, and training domain-specific models. After annotating a corpus by sentiment—this could be the traditional positive/negative scale, but also a set of sentiments, such as anger, happiness, disappointment, and so on—researchers train a categorizer such as a Recursive Neural Network on this corpus, before using the trained model to predict sentiment categories for unannotated text. For more on sentiment analysis, see (Liu and Zhang 2012).

Metadata Analysis

While many of the computational approaches described thus far have been concerned with detection of abusive speech through content analysis, metadata provides, in many cases, an even more useful feature set for categorization. Of course, the number and availability of these features is entirely dependent on the platform. Sites like Reddit provide comment upvote/downvote data, which sites like Twitter lack. Similarly, sites like Twitter provide data on the social networks of its users, which are missing on sites like Reddit. Metatextual features from a corpus of online news comments include the response times of comments to the news story and the "engagement" provoked by comments (i.e., the number of child comments) (Brand and Van Der Merwe 2014). (Castillo, Mendoza, and Poblete 2013), for instance, find that the presence of a Twitter user's self-description ("bio") correlates strongly with the likelihood of their authoring abusive tweets.

For those sites that provide public data about the social networks of its users, social network analysis is a useful tool in these categorization experiments. (Agichtein et al. 2008), for instance, in their analysis of Yahoo answers social relations, find social network analysis and trust propogation to be useful in predicting the quality of an answer. For more on trust propogation, see (Ortega et al. 2012) and (Rowe and Butters 2009).

A related subfield to these deals with the detection of paid malicious opinion manipulation "trolls." In this area, metatextual data is very predictive of trolls: as employees, their posts happened from 9-5, Monday through Friday (Mihaylov et al. 2015). The reply status of potential trolls, and the time of their replies were also strongly predictive factors. In cases where trolling is automated, some Twitter users have pointed out that malicious Twitter accounts are likely to be named with similar patterns, for example, eight random digits. These patterns are so widespread that the service Twitter Audit offers to tell you the proportion of your followers that are real and fake.

Some sites, like Twitter, maintain hidden tweet metadata, such as data about their users' operating systems, that could help to identify automated trolls. At least 10% of the #Gamergate tweets were written by accounts running bot operating systems, for instance. These could be used to easily identify malicious bots. For more on troll detection, see (Mihaylov, Georgiev, and Nakov 2015, @ortega_propagation_2012, and @kumar_accurately_2014).

Another related subfield is credibility detection. Motivated by a desire to create a filter for "fake news" or politically motivated misinformation, credibility detection uses machine-learning models trained on manually-annotated corpora do distinguish between credible and discreditable information. Using a training corpus of tweets labeled by volunteers as likely or unlikely to be true, (Castillo, Mendoza, and Poblete 2013) test a variety of features and methods for classification. They find that the features which best predict credible tweet threads

include: the average number of tweets posted by authors of tweets in that topic in the past, the average number of followers of the authors, the sentiment scores of the tweets, and whether the tweets contain a URL, emoticons, punctuation, or first-person pronouns (575).

Similarly, the subfield of deceptive opinion spam detection attempts to identify fraudulent product reviews on online shopping sites like amazon.com, reviews which are usually funded by the product's creators. (Ott et al. 2011), for instance, construct a "gold-standard" dataset by commissioning fraudulent opinion spam from freelance writers using Amazon Turk, and training a model on that dataset. The resulting classifier is roughly 90% accurate at detecting deceptive opinion spam, while human judges detect it only at around 50%. Some of these techniques (commissioning gold-standard data, for instance) might be applied to the detection of abusive language, as well.

There are also a few completely different academic disciplines that offer patterns of abusive language. Quantitative psycholinguistics, for instance, is interested in the language patterns of psychological conditions such as mental illness and states of heightened emotion. (Gawda 2013), for instance, for instance, studies narratives written by prison inmates diagnosed with Antisocial Personality Disorder (ASPD), as compared with a control group, and those diagnosed as not having the disorder. They find that emotional words are higher in general among those with ASPD, but negative words, for instance, might have lower than normal scores for narratives that describe hate. When seen in the context of our project of the computational identification of abusive language, this finding suggests that negative words on their own may not be markers of abuse, at least that originating from those with ASPD. Similarly, (Rieber and Vetter 1994), a literature review of "the language of psychopathy" finds that often one of the distinguishing linguistic features of these patients is the *lack* of emotional markers in certain contexts. Here again, this indicates that strong emotional valence, as measured by sentiment analysis, might not on its own be a useful feature for a categorizer, and that contextually contrasting emotional content might perform better.

Future Directions

Recommendation: a Twitter Bot

There are many opportunities for improving the computational detection of abusive language, and potentially also for constructing an automated technique for intervention. Since categorizers are most effective when they are domain-specific, a preliminary experiment would focus on a single domain and social media platform. We might choose to begin with Twitter, and with journalism, for instance. The construction of the experiment would proceed according to these steps:

- Step 1. We would start by compiling a small preliminary corpus of abusive language, gathered from harassed journalists. This corpus wouldn't need to be exhaustive, as it would be augmented later. We would study the corpus to determine the most appropriate features. Twitter metadata (bios, social networks, etc.) could be used, along with linguistic features, including emoji, uppercase, parts-of-speech, or the presence of certain words.
- Step 2. Using this corpus, we would train a model. The model's features would include all those discussed above: lexical features as well as metatextual features, and also including analytic metrics such as sentiment and formality. The model's categorizers would be selected from a meta-categorization layer, which would select the best performing categorizer. If possible, we could even use existing categorizers, like Jigsaw's Perspective API, in this layer.

- Step 3: Once the model predicts abusive language, which it would do with an associated certainty it reaches out to humans for verification. In particular, it asks the target of the potentially abusive language whether it finds the language to be abusive. "Hi. I'm a friendly abusive language detection bot," it might say. "Was this tweet offensive to you?"
- Step 4: Once the human replies, the bot uses this response to learn from the interaction. In this way, positive responses like "yes" would add the tweet to the corpus of abusive language, thereby improving the bot's detection abilities. A response like "no" would have the opposite effect, and a response like "a little" might add the tweet to the corpus in a week or weighted sense. A response like "I'm not sure" wouldn't do any of the above, but might trigger a reeavaluation of the algorithm.
- Step 4.5: Unsolicited submissions of offensive tweets might be accepted from the community. These could also be added to the collection of abusive language, probably after undergoing a human vetting process, and could then be used to improve the feature set.

The process then repeats, learning from each interaction, and regenerating its model accordingly.

There are many opportunities here for applications of methods across disciplines. The findings of quantitative psycholinguistics, for instance, are almost never discussed in the categorization experiments discussed in computer science journals. The linguistic properties (proportions of parts-of-speech, for instance) characteristic of the writing of those with certain mental illnesses could be used as potential features, as could be the properties of emotional speech.

Combinations of existing approaches could also be explored. While some studies have used formality metrics as features, and others have used social network theory, experiments have not been performed which combine these. There are many more such combinations to be tested. Grid search techniques and parameter tuning could be used to identify the best features for the domain and dataset, and this could even be done dynamically, at each of the bot's iterations.

Quality ranking, credibility detection, and malicious spam detection might also be leveraged in some way.

Automated Counterspeech

Future directoins: Addressing cyberbystanders & encouraging meta-discussions about the regulation of discussion

A figure that is constantly present in the theoretical discussions of harmful online behavior but rarely accounted for in approaches implemented by social media platforms is that of the cyberbystander. These are the third parties that witnesses an online interaction. Most users of social media sites have witnessed some form of harmful online behavior as a cyberbystander. In these situations it is more common for people to do nothing, but ideally they would engage in effective counterspeech.

Studies consistently show that individuals are more likely to engage in bullying/hateful behavior if they are in an environment that they perceive to tolerate or even reward such behavior. This is why sites are quick to remove harmful content before it is widely viewed by other users. If too many people see the post, then they are likely to think that it is acceptable to post similar content on the platform. A problem with the current approach of simply removing a post is that the cyberbystanders who saw the post before it was deleted have no way of later learning that this post was deemed inappropriate for the website. Therefore, seeing the post before it was deleted has shaped their understanding of the acceptable norms on the platform, making them more likely to mimic this harmful behavior. Therefore it would be beneficial for platforms to retroactively inform cyberbystanders that content that they engaged with (even just scrolling past) has been removed with an explanation of why it violated the community standards. On some platforms this explanation is provided to the person who reported and posted the content, but it is never shared with those who saw it.

This points to the broader need for more transparency and inclusion of platform users in discussing and establishing community standards and policy. As was discussed in the report, public facing standards and policies have vague definitions which give leeway to the platform and don't leave much room for interpretation and discussion by users. However, we know from leaked internal documents that platforms have detailed and (probably too) rigid guidelines for their content moderators to follow. Platforms should strike a better balance between the looseness of their outward-oriented documents and the rigidity of their internal documents, and in doing so should actively promote discussions on their platform about the rules and regulations. Something like this exists to some extent on Facebook already. facebook.com/help/community/ is a sort of social network of its own with an entirely different layout. On Facebook's Help Community, users can post questions they have about any of facebook's policies and their questions

can get upvoted and commented on by other quotidian users. Eventually if a post gets enough attention, a Facebook employee will provide an official answer that is pinned to the top of discussions. The layout of the site looks more like Reddit or stack overflow than facebook and most posts are about malfunctions or difficulty in doing things like sharing a post. However there is an entire section dedicated to "reporting abuse" where users post questions seeking advice for reporting abuse. In the reporting abuse section there are the following subcategories: "Inappropriate Content, Bullying & Harassment, Managing a Deceased Person's Account, Threats To Share Private Images, Someone Is Pretending To Be Me, Intellectual Property, Minors, About Our Policies."

Facebook's Help Community is a good first step, but it is clear that it is designed to feel entirely distinct from the regular Facebook experience. In fact, most users probably have never even been to this section since it requires clicking on a lot of small-print buttons to find, and it is not even the default page users are directed to when they have a query about rules or if they want to report something. If Facebook were to incorporate the Help Community section with the portions of the website that daily users experience (i.e. including help community questions in the Timeline) then it would encourage more active discussion and debate between users of the platform about the rules of the platform. This could help transform the social media platform into what Christopher Kelty describes as a "recursive public." A recursive public is a public that is not only involved in generating public speech and deliberation, but also deliberates about the techniques and conditions which governs their speech. A recursive public is " vitally concerned with the material and practical maintenance and modification of the technical, legal, practical, and conceptual means of its own existence as a public." By integrating the Help Community section of Facebook with the Timeline, users will be able to discuss rules and regulations contextually alongside the posts that are themselves being regulated.

Such an approach of encouraging meta-discussions of the rules by which discussions occur on platforms would also allow for richer counterspeech, making it less awkward for cyberbystanders to intervene and talk about the rules and norms of the platform with others. We know from the literature on 'real life' bullying that harsh punishment is rarely effective, and many who research in the domains of criminal justice and educational discipline advocate for a more restorative approach to justice. Untrained cyberbystanders are likely to deploy discouraged forms of counterspeech where they attack the speaker directly. This is highlighted by research in cyberbullying that shows that often people engaged in cyberbullying are deploying harmful speech in response to harmful speech directed at them. In schools that practice restorative justice, a form of counterspeech is crucial. Instead of having one-on-one disciplinary hearings, students that act out are brought in a "restorative circle" with peers that discuss the problems with what they did wrong. These "restorative circles" are effective because they are part of a more regular practice within restorative justice-oriented environments where classrooms have (at least) weekly discussion circles about classroom related issues including discussions of behavior and etiquette. By making group discussion of

rules and norms a common practice, restorative circles are effective because they are in the context of deliberate community building. Similarly, if social media websites integrated their sections where rules were discussed with the rest of the platform, then counterspeakers would be more accustomed to discussing rules and norms with others, thus making the counterspeech more effective.

Such a push towards transparency that invites public participation is unlikely, but some of the new tools on Instagram and Twitter are promising and do allow users to become more directly involved with the regulation of the platform. On Instagram users can customize and opt-in to blocking a list of words from comments on their posts, and on Twitter users can share lists of blocked accounts with each other. Both of these new policies bring users closer to direct engagement with the rules of the platform.

Appendices and Bibliography

Appendix: Patents

- Patent US5796948 - Offensive message interceptor for computers
- Patent US8868408 - Systems and methods for word offensiveness processing using aggregated...
- Patent US8473443 - Inappropriate content detection method for senders
- Patent US7818764 - System and method for monitoring blocked content
- Patent US20080109214 - System and method for computerized psychological content analysis of...
- Patent US20110191105 - Systems and Methods for Word Offensiveness Detection and Processing Using ...

Bibliography

Agichtein, Eugene, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. "Finding High-Quality Content in Social Media." In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 183–94. ACM. http://dl.acm.org.ezproxy.cul.columbia.edu/citation.cfm?id=1341557.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In *LREC*, 10:2200–2204. 2010.

Barlett, Christopher P, and Douglas A Gentile. 2012. "Attacking Others Online: The Formation of Cyberbullying in Late Adolescence." *Psychology of Popular Media Culture* 1 (2). Educational Publishing Foundation:123.

Barlett, Christopher P, Douglas A Gentile, and Chelsea Chew. 2016. "Predicting Cyberbullying from Anonymity." *Psychology of Popular Media Culture* 5 (2).

Educational Publishing Foundation:171.

Brand, Dirk, and Brink Van Der Merwe. 2014. "Comment Classification for an Online News Domain." http://scholar.sun.ac.za/handle/10019.1/96148.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. 2013. "Predicting Information Credibility in Time-Sensitive Social Media." *Internet Research* 23 (5):560–88. http://www.emeraldinsight.com.ezproxy.cul.columbia.edu/doi/abs/10.1108/IntR-05-2012-0095.

Diakopoulos, Nicholas A. 2015. "The Editor's Eye: Curation and Comment Relevance on the New York Times." In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1153–7. ACM. http://dl.acm.org.ezproxy.cul.columbia.edu/citation.cfm?id=2675160.

Duggan, Maeve. 2014. "Online Harassment," October. http://www.pewinternet.org/2014/10/22/online-harassment/.

Gambäck, Björn, and Utpal Kumar Sikdar. 2017. "Using Convolutional Neural Networks to Classify Hate-Speech." *ACL 2017*, 85. http://www.aclweb.org/anthology/W17-30#page=97.

Gawda, Barbara. 2013. "The Emotional Lexicon of Individuals Diagnosed with Antisocial Personality Disorder." *Journal of Psycholinguistic Research* 42 (6):571–80. https://doi.org/10.1007/s10936-012-9237-z.

Heylighen, Francis, and Jean-Marc Dewaele. 2002. "Variation in the Contextuality of Language: An Empirical Measure." *Foundations of Science* 7 (3):293–340. http://www.springerlink.com.ezproxy.cul.columbia.edu/index/p08225g588771321.pdf.

Huang, Guanxiong, and Kang Li. 2016. "The Effect of Anonymity on Conformity to Group Norms in Online Contexts: A Meta-Analysis." *International Journal of Communication* 10:18.

Kapoor, Hansika. 2016. "Swears in Context: The Difference Between Casual and Abusive Swearing." *Journal of Psycholinguistic Research* 45 (2):259–74. https://doi.org/10.1007/s10936-014-9345-z.

King, Ryan D, and Gretchen M Sutton. 2013. "High Times for Hate Crimes: Explaining the Temporal Clustering of Hate-Motivated Offending." *Criminology* 51 (4). Wiley Online Library:871–94.

Kolhatkar, Varada, and Maite Taboada. 2017. "Constructive Language in News Comments." *ACL 2017*, 11. http://www.aclweb.org/anthology/W17-30#page=23.

Kumar, Srijan, Francesca Spezzano, and V. S. Subrahmanian. 2014. "Accurately Detecting Trolls in Slashdot Zoo via Decluttering." In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, 188–95. IEEE. http://ieeexplore.ieee.org.ezproxy.cul.columbia.edu/abstract/document/6921581/.

Lenhart, Amanda, Michele Ybarra, Kathyrn Zickuhr, and Myeshia Price-Feeney. 2016. "Online Harassment, Digital Abuse, and Cyberstalking in America."

Liu, Bing, and Lei Zhang. 2012. "A Survey of Opinion Mining and Sentiment Analysis." *SpringerLink*, 415–63. https://doi.org/10.1007/978-1-4614-3223-4_13.

Mihaylov, Todor, Georgi Georgiev, and Preslav Nakov. 2015. "Finding Opinion Manipulation Trolls in News Community Forums." In *CoNLL*, 310–14. https://www.aclweb.org/anthology/K/K15/K15-1.pdf#page=334.

Mihaylov, Todor, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015. "Exposing Paid Opinion Manipulation Trolls." In *RANLP*, 443–50. https://pdfs.semanticscholar.org/5923/f6241acad641d21255c7ad8ae7a1594864e9.pdf.

Nyhan, Brendan, and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32 (2). Springer:303–30.

Ortega, F. Javier, José A. Troyano, Fermín L. Cruz, Carlos G. Vallejo, and Fernando Enríquez. 2012. "Propagation of Trust and Distrust for the Detection of Trolls in a Social Network." *Computer Networks* 56 (12):2884–95. https://doi.org/10.1016/j.comnet.2012.05.002.

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. "Finding Deceptive Opinion Spam by Any Stretch of the Imagination." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 309–19. Association for Computational Linguistics. http://dl.acm.org.ezproxy.cul.columbia.edu/citation.cfm?id=2002512.

Rieber, R. W., and Harold Vetter. 1994. "The Language of the Psychopath." *Journal of Psycholinguistic Research* 23 (1):1–28. https://doi.org/10.1007/BF02143173.

Rowe, Matthew, and Jonathan Butters. 2009. "Assessing Trust: Contextual Accountability." *ESWC, Heraklion*. http://ceur-ws.org/Vol-447/paper2.pdf.

Samghabadi, Niloofar Safi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, and Thamar Solorio. 2017. "Detecting Nastiness in Social Media." *ACL 2017*, 63. http://www.aclweb.org/anthology/W17-30#page=75.

Siersdorfer, Stefan, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. "How Useful Are Your Comments?: Analyzing and Predicting Youtube Comments and Comment Ratings." In *Proceedings of the 19th International Conference on World Wide Web*, 891–900. ACM. http://dl.acm.org.ezproxy.cul.columbia.edu/citation.cfm?id=1772781.

Sood, Sara Owsley, Elizabeth F. Churchill, and Judd Antin. 2012. "Automatic Identification of Personal Insults on Social News Sites." *Journal of the Association for Information Science and Technology* 63 (2):270–85. http://onlinelibrary.wiley.com.ezproxy.cul.columbia.edu/doi/10.1002/asi.21690/full.

Suler, John. 2004. "The Online Disinhibition Effect." *Cyberpsychology & Behavior* 7 (3). Mary Ann Liebert, Inc.:321–26.

Tokunaga, Robert S. 2010. "Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization." *Computers in Human Behavior* 26 (3). Elsevier:277–87.

Waseem, Zeerak, and Dirk Hovy. 2016. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." In *SRW@ HLT-NAACL*, 88–93. http://anthology.aclweb.org/N/N16/N16-2.pdf#page=98.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. "Ex Machina: Personal Attacks Seen at Scale." In *Proceedings of the 26th International Conference on World Wide Web*, 1391–9. International World Wide Web Conferences Steering Committee. http://dl.acm.org.ezproxy.cul.columbia.edu/citation.cfm?id=3052591.