Relevant hate crime/harassment law

- USA
    - Without My Consent, an anti-harassment non-profit has a 50 state project where they've compiled relevant criminal and civil laws in all 50 states and federally that could be used to prosecute various forms of online harassment. http://www.withoutmyconsent.org/50state
    - Nationally
        - Title IX of civil rights act prohibits discrimination in any education program receiving federal financial assistance
        - Title VII of civil rights act prohibits discrimination in workplace
        - Copyright law has been used to prevent the dissemination of nonconsensual pornography
        - 2009 Matthew Shepard and James Byrd, Jr. Hate Crimes Prevention Act: Matthew Shepard act introduced hate crimes based on gender, sexual orientation, disability and gender identity into the list of federally protected groups in hate crimes. It also removed the prerequisite that the victim be engaging in a federally protected activity for the crime to be considered a hate crime, and it also required the FBI to track statistics on hate crimes based on gender/gender identity in addition to the other groups that were already being tracked. The act also increased federal law enforcement's ability to intervene in state prosecution.
        - 1968 Congress passed, and LBJ signed into law, the first federal hate crimes statute. The statute made it a crime to use, or threaten to use, force to willfully interfere with any person because of race, color, religion, or national origin and because the person is participating in a federally protected activity,
        - 1968 civil rights act had six federally protected activities where discrimination is unlawful (Matthew Shepard act in 2009 extended hate crime beyond these protected activities)
            - enrolling in or attending any public school or public college;
            - Participating in government program
            - Applying for employment
            - Serving as juror
            - Traveling or using any interstate facility (motor, rail, water, or air)
            - Patronizing public place: Enjoying goods and services of hotels, restaurants, gas stations, theaters or other establishments that serve the public
        - 1968, Congress made it a crime to use, or threaten to use, force to interfere with housing rights because of the victim's race, color, religion, sex, or national origin;
        - federal criminal law as well as state laws mainly function as "penalty enhancement" whereby penalties for bias-motivated crimes are

increased. This means that online hate speech usually doesnt constitute crime in itself, but could enhance the penalty for associated crimes

- NTIA report in 1993 https://www.ntia.doc.gov/legacy/reports/1993/TelecomHateCrimes1993.pdf
    - The National Telecommunication and Information Administration (part of dept of commerce) did a study on the role of Telecommunication in hate crimes. Their findings were that Telecommunication technology do not significantly contribute to hate crimes and that the best solution in their view to hate speech was more speech that was non-hateful. They opposed any regulation of telecommunication protecting against hateful, bigoted, or racist speech out of fear of 1st amendment violations. Instead of legal reforms, their recommendations encouraged government officials and private media companies to speak out against hate speech. The NTIA report seems to consider hate crimes to be limited to 'real world' crimes such as murder or arson, so the role of telecommunication seems just to be a means to express the prejudice that motivates the real hate crime.
    - The NTIA report mentions that as early as 1983 there were reported computer bulletin boards for hate groups such as "aryan nation liberty net" and other boards for neo-nazis and skinheads. The boards had hateful propaganda as well as a "hit list" of targeted jews, blacks, and 'traitors.' Alan Berg, a jewish radio talk show host was murdered in 1984 after his name was added to one of these hit lists.

- States
    - There are 5 states with no hate crime laws: Georgia, Wyoming, South Carolina, Indiana, and Arkansas. Other states vary in what groups are protected. Race, religion, and ethnicity are the most covered, disability is covered by 31 states. Fewer protect gender or gender identity. And only 3 protect homelessness. Recently there has been a push to protect against "blue racism" against police officer.
    - New York
        - Article 485 of NY Penal law: Hate crime is committed if a person commits a specified offense against someone due to their (the perpetrator's) "belief or perception regarding the race, color, national origin, ancestry, gender, religion, religious practice, age, disability or sexual orientation of a person, regardless of whether the belief or perception is correct"
            - Menacing and harassment are included in the list of "specified offenses"
            - Does not seem like it has been used for online hate yet

        - At the federal level, the US supreme court overturned the conviction of a man who threatened to slit his estranged wife's throat on facebook, because the prosecuters were unable to prove intention to actually commit the crime http://www.pewresearch.org/fact-tank/2015/06/01/the-darkest-side-of-online-harassment-menacing-behavior/
    - No separate tort relating to cyberbullying crimes--victims would instead have to use defamation law or infliction of emotional distress law
        - There are laws related to cyberbullying education/intervention in public schools https://charactercounts.org/cyberbullying-law-passed-in-new-york/
    - N.Y. Penal Law § 135.60 Coercion--compelling individual to engage in conduct by instilling fear that if they do not comply the perpetrator will expose or publicize a secret that will subject the victim to hatred, contempt or ridicule
        - *People v. Piznarski*, 2013 NY Slip Op. 61967(U) (N.Y. Ct. App. 3rd Dep't. Jan. 15, 2013)[1]
            - Defendant filmed his girlfriend performing consensual act and coerced her into another sexual act with the threat of revealing the video. He was charged with coercion in the second degree and unlawful surveillance
    - Stalking law: intentionally stalking with no legitimate purpose, and the course of conduct will instill reasonable fear of material harm in the victim.
    - Menacing: intentionally placing or attempting to place another person in reasonable fear of physical injury, serious physical injury or death
    - Harassment and criminal nuisance has also been used in cases of cyber harassment
- Europe
    - Council of Europe created a commission against racism and intolerance which publishes reports and policy recommendations
    - The European Union drafted the "Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law." This is the framework used for the recent cooperative agreement reached between private online platforms and the European Commission to police hate speech on the websites' platforms.[1]

[1]https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code
https://www.theguardian.com/media/2017/jun/30/germany-approves-plans-to-fine-social-media-firms-up-to-50m

    - The framework defines hate speech as one of three things:
    - (1)“Public incitement to violence or hatred directed against a group of persons or a member of such group defined on the basis of race, [color], descent, religion or belief, or national or ethnic origin;”
    - (2) The same, when done through “public dissemination or distribution of tracts, pictures, or other material;”
    - (3)“publicly condoning, denying or grossly trivializing crimes of genocide, crimes against humanity, and war crimes [as defined in EU law], when the conduct is carried out in a manner likely to incite violence or hatred against such group or a member of such group.”
- International
    - Article 20 of the International Covenant on Civil and Political Rights (ICCPR) says that “any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law” (20). The UNHCR creates the “Rabat Plan” which sets out a six part test to assess the severity of hate speech:
    - (1) the social and political context in which the statement is made;
    - (2) the position or status of the speaker in society;
    - (3) the specific intent to cause harm;
    - (4) the degree to which the content of the speech was “provocative and direct,” and the “nature of the arguments deployed in the speech”;
    - (5) the extent and reach of the speech and the size

****

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

This paper looks at the differences between 15 widely used social media platforms. They found that in general, harassment is not specifically defined but is instead mentioned alongside other prohibited behavior such as spamming or hacking. The study analyzed formal and informal documents on the social media sites, and found that on average the documents had an FKGL score of 11.2 (meaning that they were readable by people with a high school Junior’s reading level). None of the policy documents define harassment outright, but Instagram and Twitter provide specific behaviors or activities that could constitute harassment. They also found that informal documents such as “community guidelines” were more likely to mention harassment than the formal documents. For the most part, the discussion of sanctions and responses to harassment were lumped in with punishment for other prohibited activity such as spamming and hacking. While all sites expressed their willingness to work with law enforcement, Instagram and tumblr were the most proactive in working with third parties, allowing users to interface directly with anti-bullying and suicide prevention groups.

An important question raised by the study is how to differently deal with cases of inward harassment/harm such as suicidal posts and posts indicating eating disorders. Another issue with the current policies is that there is no procedures for dealing with witnesses or bystanders. This is an important issue because bystanders reading hate speech/harassment shapes their norms on the site as well.

Overall instagram seems to have the most written about harassment in their policy documents, and is the most proactive in addressing harassment (including self-harm). This may be because of their very young user base along with the fact that the CEO is being rather proactive in making the app more pleasant for users.

****

Waseem, Zeerak, et al. "Understanding Abuse: A Typology of Abusive Language Detection Subtasks." *arXiv preprint arXiv:1705.09899* (2017).
http://www.aclweb.org/anthology/W17-3012

This article from the ACL workshop on abusive language proposes two primary factors for typologizing abusive language. To a large extent I agree with them that their two factors are more useful vectors than attempting to define various terms such as abusive language, hate speech, cyberbullying, cyberharassment etc. They propose the following two factors:

1. Is the language directed towards a specific individual or entity or is it directed towards a generalized group?
2. Is the abusive content explicit or implicit?

This is useful because the definition of terms like cyberbullying, harassment, and hate speech often overlap. While there is overlap, the two factors also allow researchers to more clearly define their terms of research. For example cyberbullying deals with directed attacks, but the abusive content could be explicit or implicit. The distinction between explicit and implicit abuse is compared with Barthes' distinction between denotation and connotation.

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 a crime against a sub-group of people.

	Explicit	*Implicit*
Directed	"Go kill yourself", "You're a sad little f*ck" (Van Hee et al., 2015a), "@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga" (Davidson et al., 2017), "Youre one of the ugliest b*tches Ive ever fucking seen" (Kontostathis et al., 2013).	"Hey Brendan, you look gorgeous today. What beauty salon did you visit?" (Dinakar et al., 2012), "(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles" (Hine et al., 2017), "you're intelligence is so breathtaking!!!!!" (Dinakar et al., 2011)
Generalized	"I am surprised they reported on this crap who cares about another dead n*gger?", "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!" (Nobata et al., 2016), "So an 11 year old n*gger girl killed herself over my tweets? ^_^ thats another n*gger off the streets!!" (Kwok and Wang, 2013).	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." (Burnap and Williams, 2015), "most of them come north and are good at just mowing lawns" (Dinakar et al., 2011), "Gas the skypes" (Magu et al., 2017)

Table 1: **Typology of abusive language.**

****

Clarke, Isobelle, and Jack Grieve. "Dimensions of Abusive Language on Twitter." *Proceedings of the First Workshop on Abusive Language Online*. 2017.

http://www.aclweb.org/anthology/W17-3001

This study uses Multi-Dimensional Analysis and Multiple Correspondence Analysis to find the main dimensions of linguistic variation in abusive language. They use a set of 1,486 tweets, 628 are sexist tweets and 858 are racist tweets. They find 3 main dimensions of linguistic variation which are characterized by the degree of interactive, antagonistic, and attitudinal language exhibited by tweets.

The study finds that sexist tweets on average are more interactive and attitudinal than racist tweets, meaning that sexist tweeters are more likely to interact with their targets and trivialize their target's speech. On the other hand racism is more likely to be reproduced by story-telling and argumentation which may seem less attitudinal and present itself more as facts that justify their racist ideology.

A limit of the study is that they only look at racism and sexism (they include islamophobia in their definition of racism), and that all the tweets in the sample were abusive tweets with no non-abusive tweets as a point of comparison.

Table 1: The positive and negative features strongly contributing to the Dimensions

Dim	Coord	Features
2	+	Question mark (4), Question do (3.9), Accusative case (3.8), absence of Prepositions (3.5), absence of Nouns (3.3), 2nd person pronoun (3.1), absence of Proper nouns (2.9), Emoticons (2.4), absence of Articles (2.4), Nominative case (2.3), Other pronouns (2.2), WH-words (2.1), absence of Attributive adjectives (2.1), Initial DO (2), absence of Be as main verb (1.8), absence of Coordinating conjunctions (1.2), 1st person pronouns (1.2), Subject pronouns (1.1), Initial verbs (.9), WH-clause (.9), Exclamation marks (.8), Quotation marks (.7), absence of Mentioning (.7), Hashtags (.7), Interjections (.6)
	-	Existentials (5.5), Place adverbials (5.4), BE as main verb (3.3), Coordinating conjunctions (2.3), Proper nouns (2.3), absence of Nominative case (2), Articles (1.9), Quantifiers (1.9), Attributive adjectives (1.6), Synthetic negation (1.5), Predicative adjectives (1.2), Contrastive conjunctions (1.2), absence of Other pronouns (1.1), Nominalisations (1.1), Prepositions (1), Numerals (.9), absence of 2nd person pronouns (.9), absence of Accusative case (0.9), Perfect aspect (.7), Determiners (.7), absence of Question marks (.7)
3	+	Question DO (9), Question marks (6.8), 2nd person pronouns (6.8), absence of Subject pronouns (4.4), Initial DO (3.7), Initial verbs (3.2), Determiners (3), Nominalisation (2), Synthetic negation (2), Possessive pronouns (1.9), absence of 1st person pronouns (1.8), Other pronouns (1.7), absence of Nominative case (1.1), absence of Third person pronoun (1), Pro-verb DO (.9), Emoticons (.8), Existentials (.8), BE as main verb (.7)
	-	Subject pronouns (8.7), 1st person pronouns (6.2), Auxiliary BE (3.2), 3rd person pronouns (2.8), Object pronouns (2.5), absence of 2nd person pronouns (1.9), Progressive aspect (1.8), absence of Determiners (1.7), Verbs of perception (1.6), Nominative case (1.3), absence of Mentioning (1.2), absence of Question marks (1.2), absence of Other pronouns (.9), Passives (.8)
4	+	Predicative adjectives (4.5), Existentials (4.4), absence of Prepositions (3.7), absence of Proper nouns (3.5), BE as main verb (3.4), Place adverbials (3), Emoticons (2.5), absence of Nouns (2.3), Synthetic negation (2.3), absence of Capitalisation (2), Subject pronouns (1.9), 1st person pronouns (1.9), absence of Past tense (1.4), Interjections (1.3), absence of Auxiliary BE (1.2), Comparatives (1.1), absence of Articles (1), Requests (.9), absence of URLs (.8), Nominative case (.8)
	-	Auxiliary BE (7.3), Progressive aspect (4.6), Hashtags (3.9), Capitalisations (3.2), By-passives (3.3), URLs (3.1), Proper nouns (2.8), Public verbs (2.1), absence of BE as main verb (1.8), Past tense (1.5), Numerals (1.5), Question DO (1.3), Passives (1), Prepositions (1), Perfect aspect (1), absence of Subject pronouns (1), Articles (0.8), absence of Nominative case (0.7), absence of Predicative adjectives (0.7), Infinitives (0.7)

The table shows the features that are positively and negatively correlated with each dimension (2: interactive, 3: antagonistic, 4: attitudinal).

****

****

At Columbia this past week student activist Meghan Brophy wrote an OP-ED for the spectator describing her involvement in student activism. In response there have been flyers with her name disseminated around campus that link to this website (http://fuckspec.com/) which only

contains a copy of her article with comments on the article that criticize her writing, politics, and make sexual innuendos.
Original article; http://columbiaspectator.com/opinion/2017/09/06/dont-just-be-a-spectator/

****
****

Twitter bots can reduce racist slurs—if people think the bots are white

https://arstechnica.com/science/2016/11/twitter-bots-can-reduce-racist-slurs-if-people-think-the-bots-are-white/
11/15/16
Munger found 231 twitter accounts that tended to use the n-word in a targeted manner, they were all, or almost all white men accounts.
Munger made a variety of accounts, some with white male avatars and other with black male avatars, and each account varied in number of followers. The bots would send the tweet: "@[subject] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language."

> Munger's data shows that a rebuke from an apparent white user with a high follower count had the most impact, and this impact carried more weight with the most anonymous Twitter users. In these cases, future posts containing the n-word dropped by roughly 27 percent compared to a control group in the following week. That drop-off leveled out somewhat in two-week and one-month follow-ups, but it remained. (As Munger puts it, "the 50 subjects in the most effective treatment condition tweeted the word 'nigger' an estimated 186 fewer times in the month after treatment.")

Munger's finding about anonymity is particularly interesting/useful. He found that accounts with personal information (name, photo, location etc) using the n-word were significantly less likely to change their behavior after being rebuked by the bot, and in some cases would increase the usage. This seems to go against other psychological studies on online anonymity's effect on hate speech.

Munger is an Nyu PhD student who it would be good to contact for a conference

https://github.com/kmunger/Replication-Materials-for-Tweetment-Effects-on-the-Tweeted

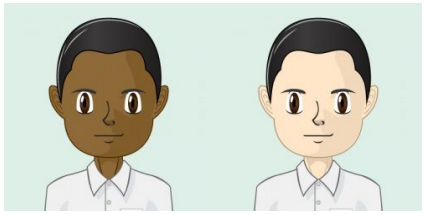
Tweetment Effects on the Tweeted: Experimentally

Reducing Racist Harassment

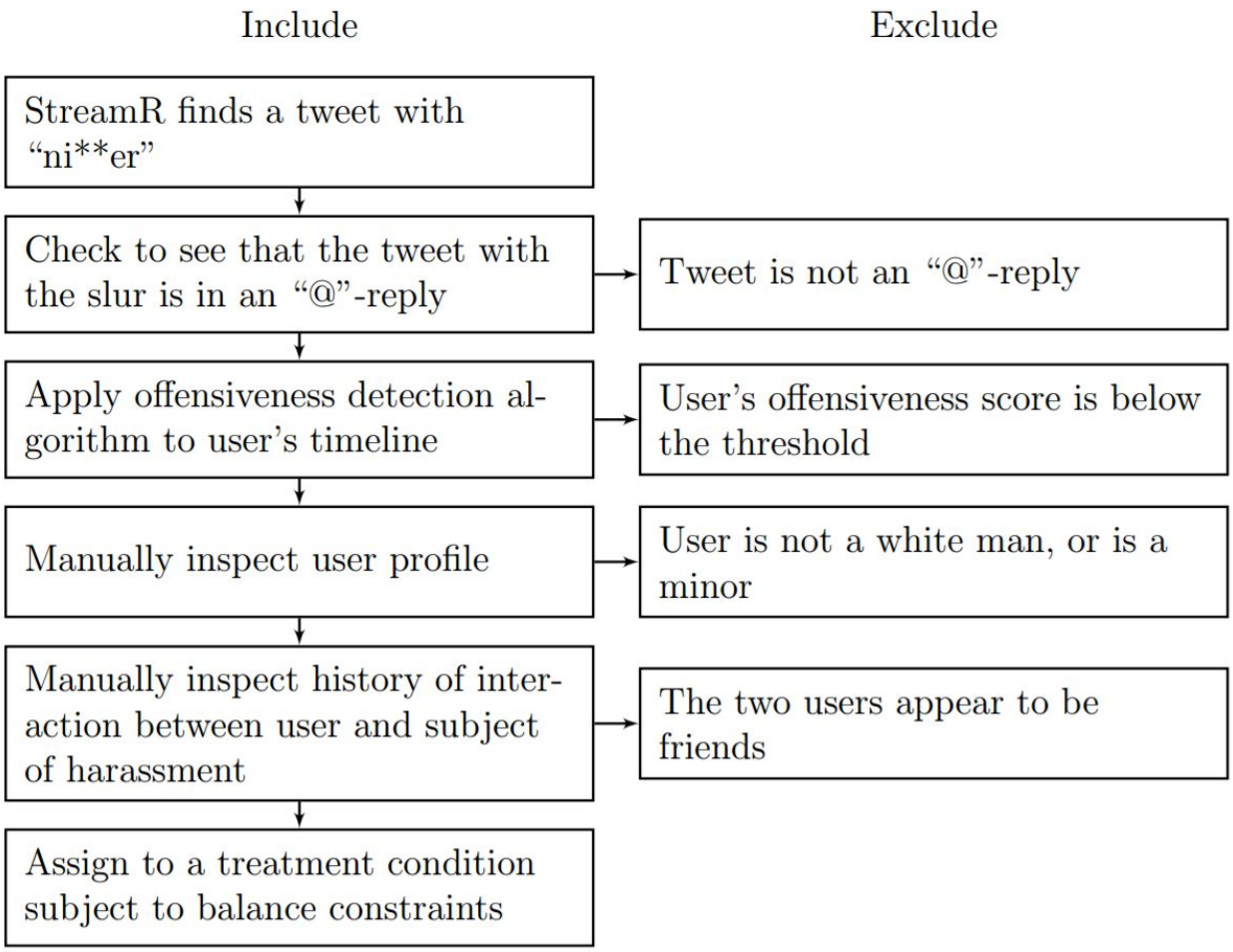
https://link-springer-com.ezproxy.cul.columbia.edu/content/pdf/10.1007%2Fs11109-016-9373-5.pdf

Kevin Munger

I conduct an experiment which examines the impact of group norm promotion and social sanctioning on racist online harassment. Racist online harassment de-mobilizes the minorities it targets, and the open, unopposed expression of racism in a public forum can legitimize racist viewpoints and prime ethnocentrism. I employ an intervention designed to reduce the use of anti-black racist slurs by white men on Twitter. I collect a sample of Twitter users who have harassed other users and use accounts I control (‘‘bots’’) to sanction the harassers. By varying the identity of the bots between in-group (white man) and out-group (black man) and by varying the number of Twitter followers each bot has, I find that subjects who were sanctioned by a high-follower white male significantly reduced their use of a racist slur. This paper extends findings from lab experiments to a naturalistic setting using an objective, behavioral outcome measure and a continuous 2-month data collection period. This represents an advance in the study of prejudiced behavior



@Rasheed and @Greg twitter bots



How munger selected accounts

****

https://www.theverge.com/2014/8/8/5981565/cyberbullying-prevention-project-trisha-prabhu
https://getinspired.mit.edu/sites/default/files/documents/ST307_Report.pdf

14 year old Trisha Prabhu created an anti-cyberbullying app called Rethink that detected harmful comments before they were published and asked users if they were sure that they wanted to post it and encouraging to consider their harmfulness. She found that a large majority (94%) of people who were asked to reconsider their harmful decided not to post the harmful post.

****

****

Twitter lets you avoid trolls by muting new users and strangers
https://techcrunch.com/2017/07/10/twitter-mute/