A Just and Comprehensive Strategy for Using NLP to Address Online Abuse

David Jurgens
University of Michigan
School of Information
jurgens@umich.edu

Eshwar Chandrasekharan
Georgia Tech
School of Interactive Computing
eshwar3@gatech.edu

Libby Hemphill
University of Michigan
School of Information
libbyh@umich.edu

Abstract

Online abusive behavior affects millions and the NLP community has attempted to mitigate this problem by developing technologies to detect abuse. However, current methods have largely focused on a narrow definition of abuse to detriment of victims who seek both validation and solutions. In this position paper, we argue that the community needs to make three substantive changes: (1) expanding our scope of problems to tackle both more subtle and more serious forms of abuse, (2) developing proactive technologies that counter or inhibit abuse before it harms, and (3) reframing our effort within a framework of justice to promote healthy communities.

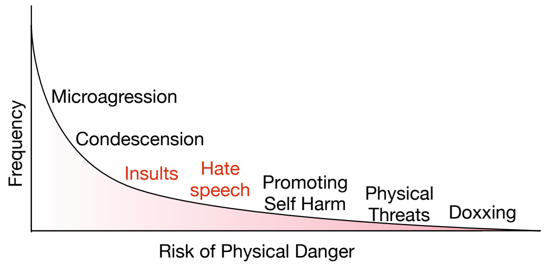

Figure 1: Abusive behavior online falls along a spectrum, and current approaches focus only on a narrow range (shown in red text), ignoring nearby problems. Impact comes from both the frequency (on left) and real-world consequences (on right) of behaviors. This figure illustrates the spectrum of online abuse in an hypothetical manner, with its non-exhaustive examples inspired from prior surveys of online experiences (Duggan, 2017; Salminen et al., 2018).

1 Introduction

Online platforms have the potential to enable substantial, prolonged, and productive engagement for many people. Yet, the lived reality on social media platforms falls far short of this potential (Papacharissi, 2004). In particular, the promise of social media has been hindered by antisocial, abusive behaviors such as harassment, hate speech, trolling, and the like. Recent surveys indicate that abuse happens much more frequently than many people suspect (40% of Internet users report being the subject of online abuse at some point), and members of underrepresented groups are targeted even more often (Herring et al., 2002; Drake, 2014; Anti-Defamation League, 2019).

The NLP community has responded by developing technologies to identify certain types of abuse and facilitating automatic or computer-assisted content moderation. Current technology has primarily focused on overt forms of abusive language and hate speech, without considering both (i) the success and failure of technology beyond getting the classification correct, and (ii) the myriad forms that abuse can take. As Figure 1 shows, a large spectrum of abusive behavior exists—some with life-threatening consequences—much of which is currently unaddressed by language technologies. Explicitly hateful speech is just one tool of hate, and related tactics such as rape threats, gaslighting, First Amendment panic, and veiled insults are effectively employed both off- and online to silence, scare, and exclude participants from what should be inclusive, productive discussions (Filipovic, 2007).

In this position paper, we argue that to promote healthy online communities, three changes are needed. First, the NLP community needs to rethink and expand what constitutes abuse. Second, current methods are almost entirely *reactive* to abuse, entailing that harm occurs. Instead, the community needs to develop proactive technologies that assist authors, moderators, and platform owners in preventing abuse before it occurs. Finally, we argue that both of these threads point to a need for a broad re-aligning of our community goals towards *justice*, rather than simply the elim-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3658–3666
Florence, Italy, July 28 - August 2, 2019.

ination of abusive behavior. In arguing for these changes, we outline how each effort offers new challenging NLP tasks that have concrete benefits.

2 Rethinking What Constitutes Abuse

The classifications we adopt and computationally enforce have real and lasting consequences by defining both what is and what is not abuse (Bowker and Star, 2000). Abusive behavior is an omnibus term that often includes harassment, threats, racial slurs, sexism, unexpected pornographic content, and insults—all of which can be directed at other users or at whole communities (Davidson et al., 2017; Nobata et al., 2016). However, NLP has largely considered a far narrower scope of what constitutes abuse through its selection of which types of behavior to recognize (Waseem et al., 2017; Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). We argue that NLP needs to expand its computational efforts to recognize two additional general types of abuse: (a) infrequent and physically dangerous abuse, and (b) more common but subtle abuse. Additionally, we need to develop methods that respect community norms in classification decisions. These categories of abuse and the importance of community norms have been noted elsewhere (Liu et al., 2018; Guberman and Hemphill, 2017; Salminen et al., 2018; Blackwell et al., 2017) but have not yet received the same level of attention in NLP.

Who has a right to speak and in what manner are subjective decisions that are guided by social relationships (Foucault, 1972; Noble, 2018), and the specific choices our algorithms make about what speech to allow and what to silence have powerful effects. For instance, rejecting behavior as not being abusive because it is outside the scope of our classification can cause substantial harm to victims (Blackwell et al., 2017), tacitly involving the NLP community in algorithmic bias that sanctions certain forms of abuse. Thus, categorization is particularly thorny: a broad categorization is likely too computationally inefficient, yet a narrow categorization risks further marginalizing affected community members and can lead to lasting harm. Following, we outline three key directions for the community to expand its definitions.

2.1 Physically Threatening Online Abuse

We outline three computational challenges related to infrequent but overt physically-manifesting abuse that NLP could be applied to solve. First, such behaviors do not necessarily adopt the language of hate speech or more common forms of hate speech and may in some contexts appear innocuous but are clearly dangerous in others. For example, posting a phone number to call could be acceptable if one is encouraging others to call their political representative, yet would be a serious breach of privacy (doxxing) if posted as part of a public harassment campaign. Similarly, declarations of "keep up the weight loss!" may be positive in a dieting community, yet reinforce dangerous behavior in a pro-anorexia community. Speech that in isolation appears offensive, such as impoliteness or racial slurs, may serve pro-social functions such as promoting intimacy (Culpeper, 1996) or showing camaraderie (Allan, 2015).

Second, behaviors such as swatting, human trafficking, or pedophilia have all occurred on public social media platforms (Jaffe, 2016; Latonero, 2011; Holt et al., 2010). However, methods have yet to be developed for recognizing when users are engaging in these behaviors, which may involve coded language, and require recognizing these alternative forms. Current approaches for learning new explicitly-hateful symbols could be adapted to this task (e.g., Roy, 2016; Gao et al., 2017).

Third, online platforms have been used to incite mobs of people to violence (Siegel, 2015). These efforts often use incendiary fake news that plays upon factional rivalries (Samory and Mitra, 2018). Abusive language detection methods can build upon recent advances at detecting fake news to identify content-sharing likely to lead to violence (McLaughlin, 2018; Oshikawa et al., 2018).

2.2 Subtle Abuse

Many forms of abusive behavior are linguistically subtle and implicit. Behaviors such as condescension, minimization (e.g., "your situation isn't that bad"), benevolent stereotyping, and microagressions are frequently experienced by members of minority social groups (Sue et al., 2007; Glick and Fiske, 2001). While subtle, such abuse can still be as emotionally harmful as overt abuse to some individuals (Sue, 2010; Nadal et al., 2014). The NLP community has two clear paths for growth into this area.

First, although recognized within the larger NLP abuse typology (Waseem et al., 2017), only a handful of approaches have attempted these problems, such as identifying benevolent sexism (Jha and Mamidi, 2017), and new methods must be developed to identify the implicit signals. Successful approaches will likely require advances in natural language understanding, as the abuse requires reasoning about the implications of the propositions. A notable example of such an approach is Dinakar et al. (2012) who extract implicit assumptions in statements and use common sense reasoning to identify social norm violations that would be considered insults.

Second, new methods should identify *disparity* in treatment of social groups. For example, in a study of the respectfulness of police language, Voigt et al. (2017) found that officers were consistently less likely to use respectful language with black community members than with white community members—a disparity in a *positive* social dimension. As NLP solutions have been developed for other social dimensions of language such as politeness (Danescu-Niculescu-Mizil et al., 2013; Munkova et al., 2013; Chhaya et al., 2018) and formality (Brooke et al., 2010; Sheikha and Inkpen, 2011; Pavlick and Tetreault, 2016), these methods could be readily adapted for identifying such systematic bias for additional social categories and settings.

2.3 Community Norms Need to be Respected

Social norms are rules and standards that are understood by members of a group, and that guide and constrain social behavior without the force of laws (Triandis, 1994; Cialdini and Trost, 1998). Norms can be nested, in that they can be adopted from the general social context (e.g., use of pejorative adjectives are rude), and more general internet comment etiquette (e.g., using all caps is equivalent to shouting). Yet, norms for what is considered acceptable can vary significantly from one community to another, making it challenging to build one abuse detection system that works for all communities (Chandrasekharan et al., 2018).

Current NLP methods are largely context- and norm-agnostic, which leads to situations where content is removed unnecessarily when deemed inappropriate (i.e., false positives), eroding community trust in the use of computational tools to assist in moderation. A common failure mode for sociotechnical interventions like automated moderation is failing to understand the online community where they are being deployed (Krishna, 2018). Such community-specific norms and context are important to take into account, as NLP researchers are doubling down on context-sensitive approaches to define (e.g., Chandrasekharan and Gilbert, 2019) and detect abuse (e.g., Gao and Huang, 2017).

However, not all community norms are socially acceptable within the broader world. Even behavior considered harmful in one community might be celebrated in another, e.g., Reddit's r/fatpeoplehate (Chandrasekharan et al., 2017), and Something Awful Forums (Pater et al., 2014). The existence of problematic normative behaviors within certain atypical online communities poses a challenge to abuse detection systems. Fraser (1990) notes that when a public space is governed by a dominant group, its norms about participation end up perpetuating inequalities. One approach to address this challenge would be to work closely with the different stakeholders involved in online governance, like platform administrators, policy makers, users and moderators. This will enable the development of solutions that cater to a wider range of expectations around moderating abusive behaviors on the platform, especially when dealing with deviant communities.

2.4 Challenges for Creating New NLP Shared Tasks on Abusive Behavior

Shared tasks have long been an NLP tradition for establishing evaluating metrics, defining data guidelines, and, more broadly, bringing together researchers. The broad nature of abusive behavior creates significant challenges for the shared task paradigm. Here, we outline three opportunities for new shared tasks in this area. First, new NLP shared tasks should develop annotation guidelines accurately define what constitutes abusive behavior in the target community. Recent works have begun to make progress in this area by modeling the context in which a comment is made through user and community-level features (Qian et al., 2018; Mishra et al., 2018; Ribeiro et al., 2018), yet often the norms in these settings are implicit making it difficult to transfer the techniques and models to other settings. As one potential solution, Chandrasekharan et al. (2018) studied community norms on Reddit in a large-scale, data-driven manner, and released a dataset of over 40K removed comments from Reddit labeled according to the specific type of *norm* being violated (Chandrasekharan and Gilbert, 2019).

Second, new NLP shared tasks must address the data scarcity faced by abuse detection research while minimizing harm caused by the data. Constant exposure to abusive content has been found to negatively and substantially affect the mental health of moderators and users (Roberts, 2014; Gillespie, 2018; Saha et al., 2019). However, labeled ground truth data for building and evaluating classifiers is hard to obtain because platforms typically do not share moderated content due to privacy, ethical and public relations concerns. One possibility for significant progress is to work with platform administrators and stakeholders to make proprietary data available as private test sets on platforms like Codalab, thereby keeping annotations in line with community norms and still allowing researchers to evaluate on real behavior.

Third, tasks must clearly define *who* is the end-user of the classification labels. For example, will moderators use the system to triage abusive content, or is the goal to automatically remove abusive content? Current solutions are often trained and evaluated in a *static* manner, only using pre-existing data; whether these solutions are effective upon deployment remains relatively unexplored. Evaluation must go beyond just traditional measures of performance like precision and recall, and instead begin optimizing for metrics like reduction in moderator effort, speed of response, targeted recall for severe types of abuse, moderator trust and fairness in predictions.

3 Proactive Approaches for Abuse

Existing computational approaches to handle abusive language are primarily reactive and intervene only after abuse has occurred. A complementary approach is developing *proactive* technologies that prevent the harm from occurring in the first place, and we motivate three proactive computational approaches to prevent abuse here.

First, bystanders can have a profound effect on the course of an interaction by steering the direction of the conversation away from abuse (Markey, 2000; Dillon and Bushman, 2015). Prior work has used experimenter-based intervention but a substantial opportunity exists to operationalize these interventions through computational means. Munger (2017) developed a simple, but effective, computational intervention for the use of toxic language (the n-word), where a human-looking bot account would reply with a fixed comment about the harm such language caused and an appeal to empathy, leading to long-term behavior change in the offenders. Identifying how to best respond to abusive behavior—or whether to respond at all—are important computational next steps for this NLP strategy and one that likely needs to be done in collaboration with researchers from fields such as Psychology. Prior work has shown counter speech to be effective for limiting the effects of hate speech (Schieb and Preuss, 2016; Mathew et al., 2018; Stroud and Cox, 2018). Wright et al. (2017) notes that real-world examples of bystanders intervening can be found online, thereby providing a potential source of training data but methods are needed to reliably identify such counter speech examples.

Second, interventions that occur after a point of escalation may have little positive effect in some circumstances. For example, when two individuals have already begun insulting one another, both have already become upset and must lose face to reconcile (Rubin et al., 1994). At this point, de-escalation may prevent further abuse but does little for restoring the situation to a constructive dialog (Gottman, 1999). However, interventions that occur *before* the point of abuse can serve to shift the conversation. Recent work has shown that it is possible to predict whether a conversation will become toxic on Wikipedia (Zhang et al., 2018) and whether bullying will occur on Instagram (Liu et al., 2018). These predictable abuse trajectories open the door to developing new models for pre-emptive interventions that directly mitigate harm.

Third, messages that are not intended as offensive create opportunities to nudge authors towards correcting their text if the offense is pointed out. This strategy builds upon recent work on explainable ML for identifying which parts of a message are offensive (Carton et al., 2018; Noever, 2018), and work on paraphrase and style transfer for suggesting an appropriate inoffensive alternative (Santos et al., 2018; Prabhumoye et al., 2018). For example, parts of a message could be paraphrased to adjust the level of politeness in order to minimize any cumulative disparity towards one social group (Sennrich et al., 2016).

4 Justice Frameworks for NLP

Martin Luther King Jr. wrote that the biggest obstacle to Black freedom is the "white moderate, who is more devoted to 'order' than to justice, who prefers a negative peace which is the absence of tension to a positive peace which is the presence of justice" (King, 1963). Analogously, by focusing only on classifying individual unacceptable speech acts, NLP risks being the same kind of obstacle as the white moderate: Instead of seeking the absence of certain types of speech, we should seek the presence of equitable participation. We argue that NLP should consider supporting three types of justice—social justice, restorative justice, and procedural justice—that describe (i) what actions are allowed and encouraged, (ii) how wrongdoing should be handled, and (iii) what procedures should be followed.

First, the *capabilities approach to social justice* focuses on what actions people can do within a social setting (Sen, 2011; Nussbaum, 2003) and provides a useful framework for thinking about what justice online could look like. Nussbaum (2003) provides a set of 10 fundamental capabilities for a just society, such as the ability to express emotion and to have an affiliation. These capabilities provide a blueprint for articulating the values and opportunities an online community provides: Instead of a negative articulation—an ever-growing list of prohibited behaviors—we should use a positive phrasing (e.g., "you will be able to") of capabilities in an online community. Such effort naturally extends our proposal for detecting community-specific abuse to one of promoting community norms. Accordingly, NLP technologies can be developed to identify positive behaviors and ensure individuals are able to fulfill these capabilities. Several recent works have made strides in this direction by examining positive behaviors such as how constructive conversations are (Kolhatkar and Taboada, 2017; Napoles et al., 2017), whether dialog on contentious topics can exist without devolving into squabbling (Tan et al., 2016), or the level of support given between community members (Wang and Jurgens, 2018).

Second, once we have adequately articulated what people in a community should be able to do, we must address how the community handles transgressions. The notion of *restorative justice* is a useful theoretical tool for thinking about how wrongdoing should be handled. Restorative justice theory emphasizes repair and uses a process in which stakeholders, including victims and transgressors, decide together on consequences. A restorative process may produce a punishment, such as banning, but can include consequences such as apology and reconciliation (Braithwaite, 2002). Just responses consider the emotions of both perpetrators and victims in designing the right response (Sherman, 2003). A key problem here is identifying which community norm is violated and NLP technologies can be introduced to aid this process of elucidating violations through classification or use of explainable ML techniques. Here, NLP can aid all parties (platforms, victims, and transgressors) in identifying appropriate avenues for restorative actions.

Third, just communities also require just means of addressing wrongdoing. The notion of *procedural justice* explains that people are more likely to comply with a community's rules if they believe the authorities are legitimate (Tyler and Huo, 2002; Sherman, 2003). For NLP, it means that our systems for detecting non-compliance must be transparent and fair. People will comply only if they accept the legitimacy of both the platform and the algorithms it employs. Therefore, abuse detection methods are needed to justify why a particular act was a violation to build legitimacy; a natural starting point for NLP in building legitimacy is recent work from explainable ML (Ribeiro et al., 2016; Lei et al., 2016; Carton et al., 2018).

5 Conclusion

Abusive behavior online affects a substantial amount of the population. The NLP community has proposed computational methods to help mitigate this problem, yet has also struggled to move beyond the most obvious tasks in abuse detection. Here, we propose a new strategy for NLP to tackling online abuse in three ways. First, expanding our purview for abuse detection to include both extreme behaviors and the more subtle—but still offensive—behaviors like microaggressions and condescension. Second, NLP must develop methods that go beyond reactive identify-and-delete strategies to one of proactivity that intervenes or nudges individuals to discourage harm *before it occurs*. Third, the community should contextualize its effort inside a broader framework of justice—explicit capabilities, restorative justice, and procedural justice—to directly support the end goal of productive online communities.

Acknowledgements

This material is based upon work supported by the Mozilla Research Grants program and by the National Science Foundation under Grant No. 1822228.

References

Keith Allan. 2015. When is a slur not a slur? the use of nigger in 'pulp fiction'. *Lang. Sci.*, 52:187–199.

Anti-Defamation League. 2019. Online hate and harassment: The american experience. https://www.adl.org/onlineharassment. Accessed: 2019-3-4.

Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):24.

Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

John Braithwaite. 2002. *Restorative Justice & Responsive Regulation*. Oxford University Press.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 90–98. Association for Computational Linguistics.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of EMNLP*.

Eshwar Chandrasekharan and Eric Gilbert. 2019. Hybrid approaches to detect comments violating macro norms on reddit. *arXiv preprint arXiv:1904.03596*.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):32.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media*, pages 76–86.

Robert B Cialdini and Melanie R Trost. 1998. Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, and G. Lindzey, editors, *The handbook of social psychology*, pages 151–192. McGraw-Hill.

Jonathan Culpeper. 1996. Towards an anatomy of impoliteness. *J. Pragmat.*, 25(3):349–367.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Kelly P Dillon and Brad J Bushman. 2015. Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior*, 45:144–150.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Bruce Drake. 2014. The darkest side of online harassment: Menacing behavior. *Pew Research Center, http://www.pewresearch.org/fact-tank/2015/06/01/the-darkest-side-of-online-harassment-menacing-behavior/*.

Maeve Duggan. 2017. Online harassment 2017.

Jill Filipovic. 2007. Blogging while female: How internet misogyny parallels "Real-World" harassment. *Yale J. Law Fem.*, 19(1).

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Michel Foucault. 1972. *The Archaeology of Knowledge & The Discourse on Language*. Pantheon Books, New York.

Nancy Fraser. 1990. Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text*, (25/26):56–80.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of RANLP*.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of ICJNLP*.

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

Peter Glick and Susan T Fiske. 2001. An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56(2):109.

John Mordechai Gottman. 1999. *The marriage clinic: A scientifically-based marital therapy*. WW Norton & Company.

Joshua Guberman and Libby Hemphill. 2017. Challenges in modifying existing scales for detecting harassment in individual tweets. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing" trolling" in a feminist forum. *The information society*, 18(5):371–384.

Thomas J Holt, Kristie R Blevins, and Natasha Burkert. 2010. Considering the pedophile subculture online. *Sexual Abuse*, 22(1):3–24.

Elizabeth M Jaffe. 2016. Swatting: the new cyberbullying frontier after elonis v. united states. *Drake L. Rev.*, 64:455.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Martin Luther King. 1963. Letter from a birmingham jail.

Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17.

Rachael Krishna. 2018. Tumblr launched an algorithm to flag porn and so far it's just caused chaos, dec 2018. *https://www.buzzfeednews.com/article/krishrach/tumblr-porn-algorithm-ban*.

Mark Latonero. 2011. Human trafficking online: The role of social networking sites and online classifieds. *Available at SSRN*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of EMNLP*.

Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Twelfth International AAAI Conference on Web and Social Media*.

Patrick M Markey. 2000. Bystander intervention in computer-mediated communication. *Computers in Human Behavior*, 16(2):183–188.

Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherje. 2018. Thou shalt not hate: Countering online hate speech. *arXiv preprint arXiv:1808.04409*.

Timothy McLaughlin. 2018. How whatsapp fuels fake news and violence in india. *https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/*.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1088–1098.

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

Dasa Munkova, Michal Munk, and Zuzana Fráterová. 2013. Identifying social and expressive factors in request texts using transaction/sequence model. In *Proceedings of RANLP*, pages 496–503.

Kevin L. Nadal, Katie E. Griffin, Yinglee Wong, Sahran Hamit, and Morgan Rasmus. 2014. The impact of racial microaggressions on mental health: Counseling implications for clients of color. *Journal of Counseling and Development*, 92(1):57–66.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

David Noever. 2018. Machine learning suites for online toxicity detection. *arXiv preprint arXiv:1810.01869*.

Martha Nussbaum. 2003. Capabilities as fundamental entitlements: Sen and social justice. *Feminist Economics*, 9(2-3):33–59.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Zizi Papacharissi. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283.

Jessica Annette Pater, Yacin Nadji, Elizabeth D Mynatt, and Amy S Bruckman. 2014. Just awful enough: the functional dysfunction of the something awful forums. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2407–2410. ACM.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics (TACL)*, 4(1):61–74.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of ACL*.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 118–123.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of KDD*, pages 1135–1144. ACM.

Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Jessica Roy. 2016. "'cuck,"snowflake,"masculinist': A guide to the language of the'alt-right'. *http://www.latimes.com/nation/la-na-pol-alt-right-terminology-20161115-story.html).LosAngelesTimes*.

Jeffrey Z Rubin, Dean G Pruitt, and Sung Hee Kim. 1994. *Social conflict: Escalation, stalemate, and settlement*. Mcgraw-Hill Book Company.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *WebSci*.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM)*.

Mattia Samory and Tanushree Mitra. 2018. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Twelfth International AAAI Conference on Web and Social Media*.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of ACL*.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *Proceedings of ICA*, pages 1–23.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Amartya Sen. 2011. *The Idea of Justice*, reprint edition edition. Belknap Press: An Imprint of Harvard University Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of NAACL*, pages 35–40.

Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 187–193. Association for Computational Linguistics.

Lawrence W Sherman. 2003. Reason for emotion: Reinventing justice with theories, innovations, and research—the american society of criminology 2002 presidential address. *Criminology*, 41(1):1–38.

Alexandra Siegel. 2015. *Sectarian Twitter Wars: Sunni-Shia Conflict and Cooperation in the Digital Age*, volume 20. Carnegie Endowment for International Peace.

Scott R Stroud and William Cox. 2018. The varieties of feminist counterspeech in the misogynistic online world. In *Mediating Misogyny*, pages 293–310. Springer.

Derald Wing Sue. 2010. *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. Wiley, Hoboken, NJ.

Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha M.B. B. Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist*, 62(4):271–286.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Harry Charalambos Triandis. 1994. *Culture and social behavior*. McGraw-Hill New York.

Tom R Tyler and Yuen Huo. 2002. *Trust in the Law: Encouraging Public Cooperation with the Police and Courts*. Russell Sage Foundation.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language*.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of ACL*.