Vectors for Counterspeech on Twitter

Lucas Wright[4], Derek Ruths[2], Kelly P Dillon[3], Haji Mohammad Saleem[2], and Susan Benesch[1,4]

[1]Berkman Klein Center for Internet and Society, Harvard University, Massachusetts
sbenesch@cyber.law.harvard.edu

[2]School of Computer Science, McGill University, Montreal
[3]Department of Communication, Wittenberg University, Ohio
[4]Dangerous Speech Project, Washington DC

Abstract

A study of conversations on Twitter found that some arguments between strangers led to favorable change in discourse and even in attitudes. The authors propose that such exchanges can be usefully distinguished according to whether individuals or groups take part on each side, since the opportunity for a constructive exchange of views seems to vary accordingly.

1 Introduction

As abusive language proliferates online, researchers struggle to define it, to detect it reliably, and to find the best ways to diminish it. 'Counterspeech' is gaining currency as a grassroots alternative to takedown, for diminishing abuse and hatred online. Counterspeech - which we define as a direct response to hateful or harmful speech - can be practiced by almost anyone, requiring neither law nor institutions. In this paper, we report counterspeech that apparently had a favorable effect on people to whom it responded. We also offer distinctions that may be useful for more reliable detection of both counterspeech and of abusive language - and for designing more effective counterspeech.

Many authors observe, as we do, that counterspeech varies greatly, in tone and in communicative strategies, and several papers offer categories of counterspeech, providing useful frameworks for observation and further study (Bartlett & Krasodomski-Jones, 2015; Briggs & Feve, 2013; Saltman & Russell, 2014). Some authors use the term 'counterspeech' expansively, however, to refer to any content that counters or contradicts hateful or extremist content - not necessarily in response to any particular speech act. A much broader category than ours, this could include forms of education, propaganda, and public information.

Our findings on counterspeech are preliminary, yet novel. The idea that 'more speech' is a remedy for harmful speech has become widely accepted since U.S. Supreme Court Justice Louis Brandeis propounded it in 1927[1] – without supporting data. We found counterspeech on Twitter[2] that, to our surprise, was followed by apologies or other signs of favorable impact on the account to which the counterspeech responded. Our findings are qualitative, since reliable quantitative detection of hateful speech or counterspeech is a puzzle yet to be fully solved due to the great variations in language employed, though we have made some progress on detection (Saleem, Dillon, Benesch, & Ruths, 2016). It is even more difficult to detect automatically 'successful counterspeech,' or counterspeech that has a favorable impact on an interlocutor. Therefore, although we used automated collection methods, most of the cases reported here were found in news reports and other literature.

Here we focus on a central idea: that just as "abusive language" is a very broad category, so is counterspeech, and in both cases, the nature and impact of the language varies with the number of people involved: whether it is produced by an

[1] Justice Brandeis asserted in his concurring opinion in Whitney v California that to expose "falsehood and fallacies" and to "avert the evil," "the remedy is more speech, not enforced silence" (Whitney v California, 1927, U.S. Supreme Court, p. 377)

[2] We first observed successful counterspeech on Twitter in Kenya in 2013, during a project to study hateful and dangerous speech online during the months leading to a presidential election. See iHub Research (2013). Subsequently, we worked with Twitter staff to find other examples of successful counterspeech, including in response to the selection of Nina Davuluri as Miss America 2014, and in response to homophobia on Twitter in France.

Proceedings of the First Workshop on Abusive Language Online, pages 57–62,
Vancouver, Canada, July 30 - August 4, 2017.

individual or a group, and whether it is directed at an individual or a group. Thus there are four "vectors" in each of which counterspeech functions quite differently, as abusive speech also does: one-to-one exchanges, many-to-one, one-to-many, and many-to-many. We also extrapolated a set of counterspeech communicative strategies from our data; those will be reported separately.

Hate speech and abusive speech online have been studied by multiple authors[3] yet they are still contested terms (Benesch, 2014; Mendel, 2012). Since it can be difficult to know a speaker's state of mind or intent, especially from a tweet, we use the term "hateful speech" to identify, and focus on, an expression of hate.

2 Challenges to detection of counterspeech

Computational approaches are required in order to study and engage counterspeech efforts at scale. The most fundamental computational capability we sought is automated detection of counterspeech (and the original posts to which the counterspeech responded).

To our knowledge, virtually no work has been published on the detection of counterspeech. Despite being entirely open, the typology outlined here offers several insights into the complexity of the detection problem and promising ways of understanding the relative hardness of different sub problems. Specifically, we have identified that counterspeech can involve a broad range of audience sizes - from single counterspeakers to whole communities. Further, we find that a single counterspeech act can exhibit a number of different communicative strategies including humor, emotional appeals, multi-stage dialog, and overt verbal attack itself. These two factors have implications for the difficulty of the detection task.

2.1 Forms of counterspeech acts

Counterspeech acts can assume many forms. Crucially, in our review of known counterspeech acts, we have observed no indication that these forms are templated - meaning that any two arbitrary counterspeech acts will not share language, syntax, or style. This contrasts, for example, with event references, conversations, mentions of politicians, and other tweet acts that carry more regular structure. The implication of this is that before automated methods can be developed, we require a better understanding of the (potentially quite subtle) structures that counterspeech acts have in common. Notably, a viable alternative to this would be using deep learning techniques, which would learn the relevant structures themselves. To use such approaches, we require very large datasets of known counterspeech acts in order to train a classifier.

2.2 Number of speakers

Given the sheer number of tweets generated each day, detection of specific tweet sets can become harder as the size of that set shrinks. This is particular true of tweet sets that lack easy-to-identify structural indicators (e.g., the use of a shared hashtag). As a result, counterspeech involving only one or a few counterspeakers is quite hard to identify: not only will there be few tweets in an entire 'conversation,' but the tweets may lack a strong signal that a classifier can use. On the other hand, counterspeech acts involving many users may adopt Twitter conventions such as mentions, retweets, and hashtags that could act as strong signals for a classifier.

Ultimately, it seems that some counterspeech acts and events will be easier to detect than others. While focusing on these easier sub problems presents a promising direction for future work, we - as a community - must remain aware that these classifiers will offer an incomplete picture of the broader counterspeech phenomenon on Twitter. Future studies should appropriately contextualize their findings and advances by also exploring the kinds of counterspeech their classifiers *cannot* detect.

3 Methods

As mentioned earlier, automatic detection of counterspeech is currently an unsolved problem. This made collecting data for our analysis a non-trivial task. Primarily, we closely followed developing news stories on controversial topics searched Twitter for discussion of such topics, and carried out informal surveys, searching for what we nicknamed "golden conversations" - three step exchanges between at least two accounts, in which hateful speech was met by counterspeech, followed by a sign of favorable impact on the first account or accounts. The last step could be an

[3] Cyberbullying also has an extensive literature, which is outside the scope of this project.

apology, a recanting, or a deleted tweet or account (the latter two were ambiguous signals, however).

We also collected our own sample of Twitter data with trending and relevant hashtags on the controversial topics, using Twitter streaming and search APIs to aggregate public tweets while the selected hashtags were still being used, sometimes for conversations. We then qualitatively analyzed the collected tweets and coded them as hateful speech or counter speech. In some cases, we used metadata from the collected tweets to find specific conversations on Twitter, to gain a better understanding of how a hashtag was used in context.

4 Vectors

We observed significant distinctions in counterspeech conversations, according to the number of participants in each stage or side. Harassment of an individual by a group of people, for example, is very different in nature and likely consequences, from hatred directed by one person against an entire racial or ethnic group.

Likewise, responses to an individual can be received very differently than responses to groups. Identifying numerous models for responses will help the chances of successful attempts, especially in media, including online (Pajares, Prestin, Chen, & Nabi, 2009, p. 293-297). These vectors can be helpful for individuals who witness abusive language, but are unsure how to respond. For example, the threshold to assume the responsibility to respond one-to-many may be too high, and thus a one-to-one response can be a more attractive or feasible counterspeech strategy.

4.1 One-to-one

Some of the most striking cases in which counterspeech seems to convince a person to change discourse are conversations between (only) two people. Where someone seems firmly committed not only to hateful ideology but to declaring it publicly, we would not expect counterspeech to sway that person. Yet in some cases, it apparently has – and has even helped to bring about lasting change in beliefs, not only speech. In these cases, we observe counterspeech strategies including: an empathic and/or kind tone, use of images, and use of humor. This counterspeech usually labels the content as hateful or racist, not its author.

A conversation in which nearly all of these strategies were used took place on January 19, 2015 – Martin Luther King (MLK) Jr. Day in the United States. It began with this tweet:

"In honor of MLK day today, I'm taking a vow to use the word "nigger" as many times as possible and in the most inappropriate times"

A writer and activist[4] discovered the tweet and responded with anti-hatred quotes from King, one after another. The first tweeter replied with a torrent of racist messages. The activist made an empathetic reference to the mother of the first tweeter. After several more exchanges, he abruptly wrote to the woman he had been attacking viciously, "you're so nice and I'm so sorry." (Payne, 2015).

Another striking example of one-to-one counterspeech is the case of Megan Phelps-Roper, who was fully convinced of the extreme homophobic tenets of the Westboro Baptist Church, in which she was raised - until she started a Twitter account to spread the views of the church. On Twitter she encountered people who challenged her views and engaged her in other ways, including humor and suggestions for music she might enjoy. Extended online conversations with two of them completely changed Phelps-Roper's views, by her own account. She ended up leaving the church. This case is described in detail by Adrian Chen (2015).

It is no surprise that deep and/or lasting change in discourse and beliefs - difficult to achieve by any means, online or offline - can take many tweets. Another distinguishing feature of one-to-one conversations is that, even on Twitter, they are not always public, since a message sent through Twitter's "direct message" feature is visible only to the sender and the receiver. In Megan Phelps-Roper's case, she and her new interlocutors also used one-to-one messaging apps other than Twitter.

In a less public online context, people may feel less guarded and therefore more open to dissenting views. On the other hand, if their conversations are invisible to the larger 'audience,' the audience can neither join in nor be favorably influenced by the conversation, except in rare cases when it is described elsewhere, as in Chen's article (2015).

[4] We've erred on the side of not revealing the identities of people in the cases we describe in order to protect them and to preserve their privacy. We've made exceptions, however, for public figures and/or those who have already chosen to discuss the case publicly.

4.2 One-to-Many

Some Twitter users have taken it upon themselves to try to change way in which others express themselves publicly on Twitter, by searching for the use of certain terms or phrases and rebuking those who use them. This sort of activist effort can be described as one-to-many counterspeech, though we note that it can also be understood as many one-to-one exchanges.

In one example, Dawud Walid, an African-American Muslim man, searched for variations of the word 'abeed' which means 'slave' and was used in tweets to refer to black people. He sent an op-ed he had written, entitled "Fellow humans are not abeed," to Twitter users who had tweeted the term. He received a variety of responses, from apologies and promises not to use the word again, to a tweet that repeated the word as many times as possible in 140 characters (Walid, 2013). Other similar efforts are the accounts @YesYoureRacist and @YesYoureSexist, which retweeted examples of racist and sexist content (often e.g. beginning with the phrase "I'm not racist, but…").

In each of these cases, counterspeech is met with a range of responses, from apologies to angry argument. In another example of one-to-many counterspeech, some users deliberately tweet on a hashtag with which they disagree, such as #stopislam, to reach people who agree with it.

4.3 Many-to-One

In some cases, news of an objectionable tweet (or hashtag) goes viral, and many Twitter users – sometimes thousands – join in counterspeech. This can be salutary where it catches enough of the attention of the original speaker to be successful but not harassing, as in the case of a user who tweeted his outrage that Nina Davuluri (whom he erroneously identified as an Arab) had been chosen as Miss America 2014. After receiving tweets that variously corrected his error and called him a racist, he first responded "I didn't realize it would explode like that #unreal" and then tweeted at Davuluri, apologizing. The furor died down quickly, and the user is still on Twitter, at this writing.

In other cases, however, huge numbers of angry Twitter users have overwhelmed others, rising to the level of harassment. Original speakers hastily delete tweets or even their accounts, but even that can be an insufficient refuge in the face of, for example, counterspeakers who contact their employers, demanding that they be fired for tweets or posts. This has indeed led to firing in several cases (Ronson, 2015).

The blog "Racists Getting Fired" made a practice of punishing people who posted racist content by contacting their employers and, similarly, demanding that they be fired (McDonald, 2014). Such responses are no doubt successful at changing the online speech of their targets, but may only harden the hateful convictions of those targets, and constitute online mob justice.

4.4 Many-to-Many

Conversations among large numbers of people online are of interest, not least because of the impressive scale on which they often take place. We observed counterspeech surging when strangers met and argued online, often because they were interested in the same offline event. On Twitter, such conversations generally form around hashtags.

Hashtags can themselves constitute hateful and abusive language – or counterspeech – and they often gather or inspire 'many-to-many' conversations. The use of "a hashtag can be seen as an explicit attempt to address an imagined community of users… as each user participating in a hashtag conversation acts potentially as a bridge between the hashtag community and members of their own follower network" (Bruns & Burgess, 2012, p. 804). Often, one hashtag represents one general view or normative group, such as #BlackLivesMatter, with others represent opposing or dissenting views, such as #BlueLivesMatter (which refers to police for their blue uniforms), or #AllLivesMatter.

One of the most vitriolic hashtags we found, #KillAllMuslims, trended in the immediate aftermath of the Charlie Hebdo massacre of January 2015 - and then was quickly taken over by counterspeakers expressing their dismay that it existed. One counterspeech tweet that uses the hashtag was retweeted more 10,000 times: "Not muslim but never thought about this b4 #CharlieHebdo #KillAllMuslims #Muslims pic.twitter.com/LL1pkPk6uk." The link was to an image of visual similarities among religious traditions, e.g. a Catholic nun in a habit and a Muslim woman in hijab.

Notably, trending hashtags can be more widely and quickly disseminated than any tweet. When #KillAllMuslims trended, for example, thousands

of people on Twitter could not help but notice two things: the hashtag called for mass murder or genocide, and thousands of people had typed it and sent it.

The fact that a hashtag is trending can also have a major impact on how Twitter users perceive norms on the platform. It is dismaying when hateful hashtags trend, and reassuring when counter-speech does. The hashtag #YouAintNoMuslimBruv, for example, trended after a bystander yelled the same phrase at a would-be attacker in London in December 2015.

5 Further Research

A worthy topic for further study would be the norm-influencing capacity of hashtags around public events and controversies, for two reasons: they draw large numbers of people, and those people are often of strikingly different views.

Without such a catalyst, people of very different convictions are less likely to exchange them since they spend most of their time in like-minded silos, reading content with which they mainly agree (Anderson & Raine, 2010, p. 18; Conover et al., 2011; Lewandowsky et al., 2012, p. 111; Zuckerman, 2013). Certain 'places' online, including Twitter accounts that draw devoted fans and ardent critics, also draw strikingly different readers or audiences, who are thus exposed to one another's ideas. This famously leads to conflict; however in some cases there are constructive exchanges which are worth finding and studying.

Acknowledgments

We thank Public Safety Canada's Kanishka Project for funding the research described here, and the John D. and Catherine T. MacArthur Foundation for supporting the Dangerous Speech Project.

References

Janna Quitney Anderson and Lee Rainie. 2010, July 2. The future of social relations. *Pew Research Center's Internet & American Life Project*, Washington, DC. http://www.pewinternet.org/files/old-media/Files/Reports/2010/PIP_Future_of_Internet_%202010_social_relations.pdf.

Jamie Bartlett and Alex Krasodomski-Jones. 2015. Counter-speech: Examining content that challenges extremism online. *Demos*. https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Susan Benesch. 2014. Defining and diminishing hate speech. In *Freedom from hate: State of the world's minorities and indigenous peoples 2014*. Minority Rights International, pages 18-25. http://minorityrights.org/wp-content/uploads/old-site-downloads/mrg-state-of-the-worlds-minorities-2014-chapter02.pdf.

Rachel Briggs and Sebastien Feve. 2013. Review of programs to counter narratives of violent extremism. *Institute of Strategic Dialogue*. https://www.counterextremism.org/download_file/117/134/444/

Axel Bruns and Jean E. Burgess. 2011. The use of Twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, University of Iceland, Reykjavik. http://eprints.qut.edu.au/46515/

Adrian Chen. 2015, November 23. Unfollow: How a prized daughter of the Westboro Baptist Church came to question its beliefs. *New Yorker*. http://www.newyorker.com/magazine/2015/11/23/conversion-via-twitter-westboro-baptist-church-megan-phelps-roper

Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media,* pages 89-96. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847

iHub Research. 2013. Umati Final Report, Sept. 2012 –May 2013. http://dangerousspeech.org/umati-final-report/

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13(3*), 106-131. https://doi.org/10.1177/1529100612451018

Soraya Nadia McDonald. 2014, December 2. 'Racists Getting Fired' exposes weaknesses of Internet vigilantism, no matter how well-intentioned. *The Washington Post.* https://www.washingtonpost.com/news/morning-mix/wp/2014/12/02/racists-getting-fired-exposes-weaknesses-of-internet-vigilantism-no-matter-how-well-intentioned/.

Toby Mendel. 2014. Does international law provide for consistent rules on hate speech? *International Journal of Constitutional Law*, 12(3):417-429. https://doi.org/10.1093/icon/mou053.

Steven Payne. 2015, January 20. An amazing woman fields a troll on MLK Day and it was nothing short of inspirational. http://www.dailykos.com/story/2015/1/20/1359055/-An-amazing-woman-feeds-a-

troll-on-MLK-Day-and-it-was-nothing-short-of-inspirational.

Jon Ronson. 2015, February 12. How one stupid tweet blew up Justine Sacco's life. *The New York Times*. http://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html.

Haji Mohammed Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2016. A Web of Hate: Tackling hateful speech in online social spaces. In *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety*. European Language Resource Association. http://www.tacos.org/node/17.

Erin Saltman and Jonathan Russell. 2014. The role of Prevent in countering online extremism. *The Quilliam Foundation.* http://preventviolentextremism.info/sites/default/files/White%20Paper%20–%20The%20Role%20of%20Prevent%20in%20Countering%20Online%20Extremism%20.pdf

Dawud Walid. 2013, November 24. Responses to my calling out the term 'abeed'. https:// dawud-walid.wordpress.com/2013/11/24/responses-to-my-calling-out-the-term-abeed/.

Ethan Zuckerman. 2013. Digital cosmopolitans: Why we think the Internet connects us, why it doesn't, and how to rewire it. WW Norton & Company.