Considerations for Successful Counterspeech

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

[1]Berkman Klein Center for Internet & Society, Harvard University, Massachusetts
[2]School of Computer Science, McGill University, Montreal
[3]Department of Communication, Wittenberg University, Ohio
[4]Dangerous Speech Project, Washington D.C.

Contact authors:
Susan Benesch - sbenesch@cyber.law.harvard.edu
Derek Ruths - derek.ruths@mcgill.ca

This guide was written for the Kanishka Project of Public Safety Canada, as part of "Evaluating Methods to Diminish Expressions of Hatred and Extremism Online," a research effort that we conducted from May 2014 to March 2016. It draws on "Counterspeech on Twitter: A Field Study," a paper produced as part of the same research project. For further details on the ideas and methods outlined here, please see that paper.

Introduction

It may sometimes seem that the Internet is sullied by a relentless tide of hatred, vitriol, and extremist content, and that not much can be done to respond effectively. Such content cannot all be deleted, after all, since even if a statement, image, or user is deleted from one platform, there is always somewhere else to go.

We have been pleasantly surprised, however, that our study of Twitter turned up numerous cases of effective counterspeech, which we define as a direct response to hateful or dangerous speech.[1] Based on this first, qualitative study of counterspeech as it is practiced spontaneously on Twitter, we offer some preliminary suggestions on which strategies may help to make counterspeech successful.

We define *successful* counterspeech in two ways. The first is speech (text or visual media) that has a favorable impact on the original (hateful) Twitter user, shifting his or her discourse if not also his or her beliefs. This is usually indicated by an apology or recanting, or the deletion of the original tweet or account. The second type of success is to positively affect the discourse norms of the 'audience' of a counterspeech conversation: all of the other Twitter users or 'cyberbystanders' who read one or more of the relevant exchange of tweets. This impact is difficult to assess when studying counterspeech "in the wild" as we have, but it may be indicated by long conversations that remain civil, and by counterspeech that leads to others counterspeaking. It may also be evident where large numbers of Twitter users join in a campaign of counterspeech. In this project we focused primarily on the first form of success since it is easier to study - and was our stated goal. Effects on audience may be more significant, however, since they are likely to be larger in scale and may be more common.

Our list of strategies is neither exhaustive nor definitive. In addition, our research for this project was conducted entirely on Twitter and may be somewhat idiosyncratic to that platform. Finally, we acknowledge gratefully - and refer readers to - the work of our colleagues in this nascent field. A list of further reading appears at the end.

[1] Hateful speech is speech which contains an expression of hatred on the part of the speaker/author, against a person or people, based on their group identity. Dangerous speech is which we have defined in previous work as speech that can inspire or catalyze intergroup violence. For more on this, see http://dangerousspeech.org/.

Recommended Strategies

The strategies described below are often used in counterspeech that seems to have a favorable impact on users who tweet hateful or inflammatory content. In general, they are most likely to succeed in response to those who are not deeply committed to hatred or extremism. Those who are strongly committed, or those who enjoy eliciting a reaction by using outrageous speech (often called trolls), or those who are both (e.g. a racist troll) are more difficult to sway using counterspeech. However, even exchanges with those who cling to hateful or extremist positions may shift norms of discourse and belief among the audience.

While we present the strategies individually, multiple strategies are often used together, even in a single tweet.

Warning of Consequences

Counterspeakers often warn of the possible consequences of speaking hatefully on a public platform like Twitter, and in many cases this seems to have been effective at getting the speaker to delete the hateful tweet. Such counterspeakers can:

- Remind the speaker of the harm that hateful or dangerous speech may do to the target group, since words can catalyze action.
- Remind the speaker how many people in his or her offline world (including employers, friends, family, and future employers) can see what is online, and note that offline consequences can include losing one's job and relationships.
- Remind the speaker of the permanence of online communication.
- Remind the speaker of the possible online consequences of hateful or dangerous speech, such as blocking, reporting, and suspended accounts.

We do not know when or whether this method is effective at changing behavior in the medium or long terms. Speakers may continue speaking hatefully on a less public or more obscure platform or on Twitter under a pseudonym. They may even revert back to speaking hatefully on the same account in the future.

Example: Alumni of the University of Illinois at Champaign-Urbana warned students that current and future employers would be able to see their tweets using the hashtag #FuckPhyllis, created to attack Phyllis Wise, the then-chancellor of their university, with misogynist and racist content. (They were angry that she had decided not to grant a snow day on a very cold Monday.) Many of the students responded by deleting the hateful tweets.

Shaming and Labeling

We have observed successful counterspeech which labels tweets as hateful, racist, bigoted, misogynist, etc. Since these words carry such a shameful stigma for many people in contemporary North America, speakers who do not perceive themselves as racists, for example, are often quick to alter such tweets. With this strategy, counterspeakers can also:

- Denounce the speech as hateful or dangerous, to others. This can help cyberbystanders identify, interpret, and respond to it.
- Explain to the original speaker why their statement is hateful or dangerous. In addition to eliciting the favorable reaction that often comes from labeling, this can also help to educate the speaker so he or she will repeat the mistake.

Example: In 2013, a Twitter user tweeted his outrage that Nina Davuluri (whom he erroneously identified as an Arab) had been crowned as Miss America. Other users called him a racist and corrected his assumption that Davuluri was Arab. Two days later he apologized with a specific reference to the racist label, saying, "*@MissAmerica sorry for being rude and 'racist' and calling you a Arab please tweet back so everyone will know its real.*"

Empathy and Affiliation

Changing the tone of a hateful conversation is an effective way of ending the exchange. While we have scant evidence that it will change behavior in the long term, it may prevent the escalation of the hateful rhetoric being used in the present moment. Counterspeakers can consider:

- Using a friendly, empathetic, or peaceful tone in responses to messages with a hostile, hateful, or violent tone.
- Affliating with the original speaker to establish a connection (e.g. I am also a conservative, but…).
- Affliating with the group targeted by the hateful speech to generate empathy (e.g. What you said was hurtful to me as an Asian…).

When counterspeakers are able to transform hateful exchanges into civil, sustained conversations, the contact hypothesis[2] suggests that those involved may become less hateful or extreme over time. We therefore identify this as a successful outcome even when the original speaker does not recant or apologize or change his or her views.

Example: In an extended one-on-one exchange with a user intent on using racial slurs on Martin Luther King Day in the United States, another Twitter user repeatedly countered with empathy, "*still wishing you love*," "*try as you might, I'm never going to wish you anything but love*," and "*I would never laugh at your murder. I would never taunt your grieving. I would never mock your fight for equality.*"

The first user at first responded with tweet after tweet of racist vitriol, but eventually apologized.

[2]See Allport, G. W. (1954). *The Nature of Prejudice*. Cambridge, MA: Perseus Books. Also see Dekker, R., Belabas, W., & Scholten, P. (2015). Interethnic Contact Online: Contextualising the Implications of Social Media Use by Second-Generation Migrant Youth. *Journal of Intercultural Studies*, *36*(4), 450-467.

Humor

We have observed that humorous counterspeech can shift the dynamics of communication, de-escalate conflict, and draw much more attention to a message than it would otherwise garner. It comes in many forms, of course, including caricature and sarcasm, and can vary immensely in tone, from conciliatory to provocative or even aggressive. We advise that counterspeakers using humor consider doing the following:

- Neutralize hateful and dangerous speech that is viewed as powerful or intimidating (cf the rubber duck ISIS images below).
- Attract the attention of a larger audience to the counterspeech.
- Use humor to soften a message that would otherwise be harsh or aggressive.

Example: In November 2015, Internet users began sharing images of ISIS members edited so that they had rubber duck heads in place of their own (see image below).

Similarly, when the human rights activist Iyad El-Baghdadi tweeted part of a call to action from ISIS in 2015, Muslim Twitter users responded with reasons why they wouldn't be joining. As the tweet below illustrates, they bundled acerbic criticism of ISIS with amusing references to other activities.

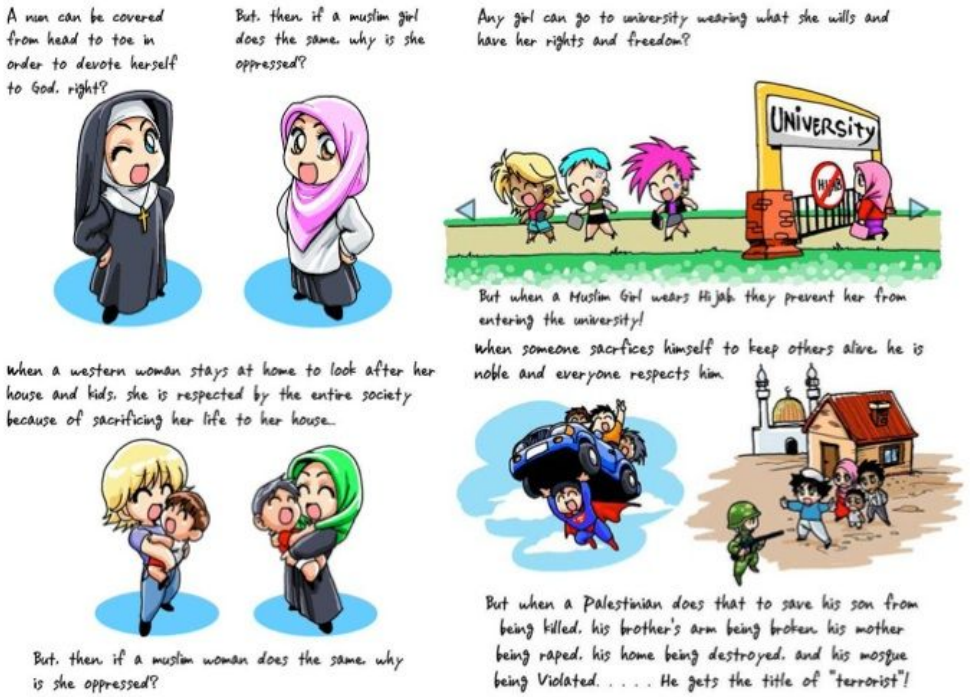
Images

Images are often more persuasive than text alone, and counterspeakers commonly incorporate them in the form of memes, graphics, photographs, animated gifs, and videos, in their tweets in response to hateful or dangerous speech. Images can transcend cultural and linguistic boundaries, which can allow counterspeech to spread virally, almost without geographic or normative boundaries. Visuals can also "send people along emotive pathways where textual/verbal material leaves them in a more rational, logical and linear pathway of thought."[3]

Counterspeakers should consider using images:

- To further a point beyond the text or the textual limitations of Twitter.
- In tandem with other strategies (e.g. humorous images).
- That are not inflammatory in nature and are from credible sources.

Example: In response to the trending of the hashtag #KillAllMuslims in January 2015, one Twitter user posted a short tweet saying, "*Not muslim but never thought about this b4 #CharlieHebdo #KillAllMuslims #Muslims*," with a link to an image (below.) The tweet was retweeted over 10,000 times. The image contains a more powerful message than could be expressed in text alone, especially with Twitter's 140 character limit.



[3] Joffe, H. (2008). The power of visual material: Persuasion, emotion and identification. *Diogenes*, *55*(1), 84-93.

Discouraged Strategies

We have observed that these strategies are often ineffective at favorably influencing the original speaker. In some cases, they may even be counterproductive or harmful.

Hostile or Aggressive Tone, Insults

Many counterspeakers respond to hateful speech with a hostile, aggressive tone, and insults. This includes but is not limited to the use of profanity, slurs, name-calling, and aspersions. We have observed that counterspeech which uses these strategies can:

- Cause a backfire effect[4] (e.g., stronger adherence to original speech).
- Cause an escalation of hateful rhetoric.
- Turn off other potential counterspeakers from joining any intervention.

The distinction between this strategy and shaming is important. While counterspeakers often use profanity and name calling in their tweets when shaming another user, this counterspeech will likely be more effective without the negative or hostile tone.

Fact-checking

Counterspeakers often react to hateful speech by correcting falsehoods or misperceptions contained in the speech. Unfortunately, this is usually not an effective method of influencing the original speaker. Especially when original speakers are entrenched in their views, they tend to find a way to fit the new facts presented to the conclusion to which they are already committed (social psychologists have named this process 'motivated reasoning') or find different evidence to support their position rather than concede. Corrections that insult or threaten an original speaker's worldview can lead him/her/them to dig in their heels.[5]

It is possible that fact checking can have a positive impact on the audience of a counterspeech exchange, especially when the audience is uninformed, but this causal link is difficult to prove.

Harassment and Silencing

Some strategies for responding to hateful speech cross the line between counterspeech and harassment. Counterspeech can be as vitriolic and hostile as the speech to which it responds, for example. Worse, some respond to speech they disapprove of with threats and harassment. We believe that these strategies should never be used even if well intentioned, since they can cause harm. Sometimes they even harm Internet users who merely share a

---

[4] Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303-330.

[5] Ibid.

name with the person who is the true target.[6] Avoiding this requires care and self-control on the part of counterspeakers. Examples of such strategies are:

- Silencing. This is also a tactic that is often used to limit the speech of women, minority groups, and dissenters. It is important to ensure that controversial views are not silenced.
- Dogpiling. Dogpiling is a term for ganging up online, or sending many hostile responses to an individual in a short period of time. It often occurs when a tweet is perceived as offensive or hateful and goes viral.
- Offline Punishment. Warnings of offline consequences can easily swell into threats and eventually harassment of online adversaries, sometimes by large numbers of people. This can lead to Internet users losing their jobs, for example, because a tweet, especially when users take it upon themselves to contact an employer.

It can be tempting to see such strategies as useful or even constructive, when used against those who are hateful. However like any serious punishments, they should not be meted out by self-appointed online police whose zeal often outstrips their good judgment. Such activity must not be confused with constructive counterspeech.

[6] Lopez, G. (2016, June 1). *The freakout over Harambe the gorilla shows the dangers of internet mob justice.* Vox.com. Retrieved from http://www.vox.com/2016/5/31/11818858/harambe-gorilla-michelle-gregg-mob-justice.

Further Reading

Bartlett, J., & Krasodomski-Jones, A. (2015). Counter-speech: Examining content that challenges extremism online. *Demos*. Retrived from http://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Brown, R. (2016). Defusing hate: A strategic communication guide to counteract dangerous speech. *US Holocaust Memorial Museum.* Retrieved from https://www.ushmm.org/m/pdfs/20160229-Defusing-Hate-Guide.pdf

Frenett, R., Dow, M. (2015). One to one online interventions: A pilot CVE methodology. *Institute for Strategic Dialogue.* Retrieved from http://www.strategicdialogue.org/wp-content/uploads/2016/04/One2One_Web_v9.pdf

Saltman, E.M., & Russell, J. (2014). The role of Prevent in countering online extremism. *The Quilliam Foundation.* Retrieved from https://www.quilliamfoundation.org/wp/wp-content/uploads/publications/free/white-paper-the-role-of-prevent-in-countering-online-extremism.pdf