FONDAZIONE BRUNO KESSLER
MAGAZINE

Combating hate speech, preserving freedom of expression

September 8, 2020

Novel techniques against hate speech are ready to be exploited to generate counter narratives about the many dimensions of hate speech: right-wing extremism, islamophobia, racism, cyber hate against LGBTI+ community, ecc. Interview with Marco Guerini

GS: Is there a way for hate speech (HS) to be regulated and prevented online? And actually – what is hate speech and what it is not on social media? **How can hate speech be contained without affecting the freedom of expression?**

MG: Actually there are many definitions of hate speech, in fact it is challenging to define it given the broadness and the nuances in cultures and languages. Usually I prefer to stick to a broad definition, such as: **Hate speech refers to "expressions that attacks or diminishes, that incites violence or hate against groups or single targets, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other."**

In this respect the medium is not so relevant: **wether online or in the real world, hate speech is always the same**. In fact, we must make clear that social media platforms are not hate generators as such; rather, they are a means by which expressions of hatred take on new lifeblood and new strength. Thanks to the combination of the way content selection algorithms work (i.e. those algorithms that decide what to display on user's News Feed), and how humans reason, hatred has the opportunity to spread like never before. In fact, people tend to react faster and with greater force to content that generates negative emotions such as fear and hatred, while algorithms tend to promote content that generates greater engagement. In this configuration, the hate comments have therefore found fertile ground to spread quickly and pervasively.

It must be acknowledged that online platforms, authorities, and Non-Governmental Organizations (NGOs) have put forward an unprecedented effort to prevent hate spreading. However, standard approaches, based on content moderation, such as account suspension, content deletion, shadow-banning may be charged with censorship, overblocking, and lastly with hindering Freedom of

Speech.

Another possible solution would be to intervene on how selection algorithms work, but this somehow presupposes to have already identify hatred, so we fall back into the previous situation.

One additional problem is that these identify-and-delete strategies cannot be easily applied to dangerous speech – i.e. content that is stirring up hatred and divisiveness but does not fall into a formal definition of HS.

“Since <group x> started arriving in our country there has been a spike in violent crimes.” which is based on a true – but spurious correlation.

GS: Do social media giants need to develop a special project dedicated to fighting hate speech and how it would look like? What other ways are there for combating hate speech on social media?

MG: To my knowledge they are actively trying to find new ways to fight online hate. Still, I believe that we must completely change our perspective: we must overstep reactive identify-and-delete strategies. We must focus on an alternative strategy that is used by some NGOs. These **NGOs are training operators to directly intervene in online hateful conversations by writing polite textual responses, called counter narratives, that are meant to oppose the hateful content with credible evidence and prevent it from further spreading.**

GS: How can people on social media help in the battle against hate speech?

MG: People are already helping social media platforms with their reporting activity, but there are two main problems:

1. it often happens that people report content they don’t like and makes them upset, that is not necessarily an HS.
2. some people are also using this reporting possibility to silence opponents. For this reason media platforms need a lot of manual intervention from moderators to check each and every report from users.

The alternative idea to manually writing responses to all hatred online requires a lot of expertise and time and ultimately is not a scalable task, it is really a toil of Sisyphus. In this respect we believe that in the future it will be crucial to mix the advantages of AI and deep learning with human expertise to overcome the proliferation and rapid spread of online hate speech. I look forward to a new Age of Enlightenment where reason replace hate and censorship, where differences can enrich our society by means of discussion.

GS: **What does the latest monitoring researchers of online hate speech show? How high is the risk of hate speech on the internet becoming real hate crimes?**

MG: The risk of hate speech becoming hate crime is always high: it can lead to depression or suicide of the targets, promote the use of violence, encourage discrimination, and increase societal divisions. It happened in the past and unfortunately will happen in the future.

One key aspect of online hate is that it allows people to see others use hateful words and slurs, making those things become normal. Social Norms have the power to influence people's behaviors. If you see a stream of slurs, that makes you feel like these things are acceptable and somehow "normal". That is why **we believe that we must enter the stream of hate and change the violent narratives with positive and inclusive ones. Maybe we won't be able to change the mind of haters but for sure we will prevent bystanders to fall into the hate trap.**

GS: **Online hate speech has many dimensions – right-wing extremism, islamophobia, racism, cyber hate against LGBTI+ community… What should be the different responses to them?**

MG: Talking about counter narratives, while there are generic principles at the root of these responses (for example, it is always the individual that is responsible for something, it is never the group), there are also "arguments" that are specific for each target – e.g. debunking all canards about jews from the "Protocols of the Elders of Zion". This is why **we need experts and education programs** as well **to win this fight**.

PERMALINK

https://magazine.fbk.eu/en/news/combating-hate-speech-preserving-freedom-of-expression/

TAGS

- #artificial intelligence
- #counter narratives
- #diversity
- #education
- #freedom
- #hatespeech
- #internet
- #social media

AUTHORS

- Giancarlo Sciascia