FONDAZIONE BRUNO KESSLER
MAGAZINE

Generating Counter Narratives against Online Hate Speech: Data and Strategies

September 13, 2020

How big is the hatespeech phenomenon online? What effects does it produce? How can we tackle it? What is the contribution of artificial intelligence? Marco Guerini and other researchers at FBK are pushing forward this knowledge frontier

Hate Speech (HS) refers to “expressions that attack or diminish, that incite violence or hate against groups, based on specific characteristics such as religion, ethnicity, sexual orientation, gender or other”.

With the rapid growth of social media platforms, abusive and offensive language can spread quickly and is difficult to monitor.

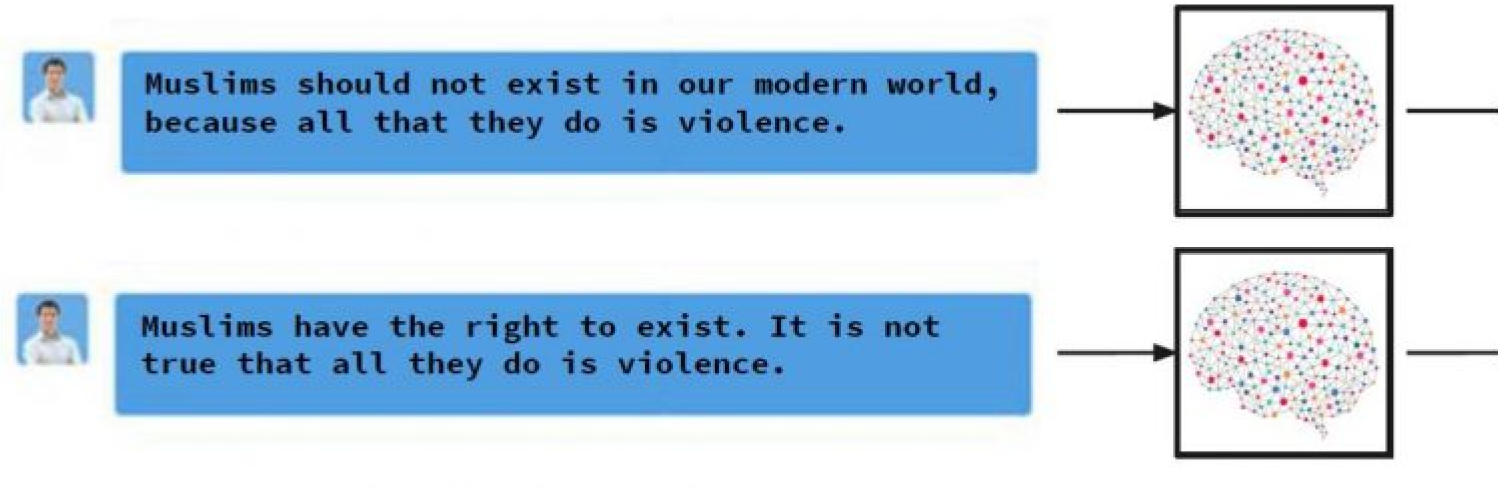
HS has real-life consequences: it can lead to depression or suicide, promote the use of violence, encourage discrimination, and increase societal divisions. An unprecedented effort to provide adequate responses in terms of laws and policies to hate content on social media platforms.

HS Countering

Standard policies are based on identify-and-sanction strategies.
? Content deletion (1st step: hate identification; 2nd step: content removal)
? User suspension
? Shadowbanning

Contribution of Artificial intelligence (AI)

Up to now, systems are able to classify content, telling if it is hateful or not.



Muslims should not exist in our modern world, because all that they do is violence.

Muslims have the right to exist. It is not true that all they do is violence.

These approaches have some limitations because they can be charged with censorship and overblocking.

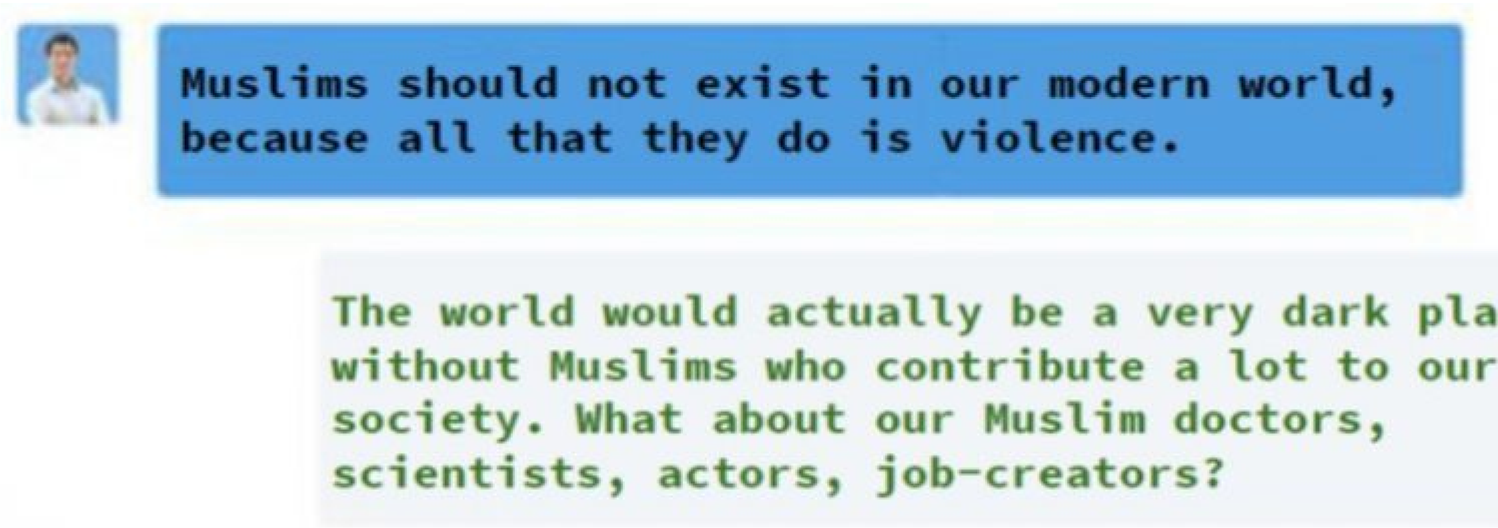
They can hinder Freedom of Speech. They cannot be applied to dangerous speech – in other words content that is stirring up hatred and divisiveness but does not fall into a formal definition of HS.

An alternative strategy

We can oppose hate content with **counter-narratives (CNs)**: direct intervention in the discussion to withstand hate messages using a non-aggressive textual response that offers feedback through fact-bound arguments.

This approach has many **advantages**:

- Preserve the right to freedom of speech
- Counter stereotypes with credible evidence
- Alter the viewpoints of haters and bystanders
- Encourage mutual understanding
- Help de-escalating the conversation



Muslims should not exist in our modern world, because all that they do is violence.

The world would actually be a very dark pla without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, actors, job-creators?

Counter Narrative example

There are also important **limitations**:

- Manual intervention alone is not scalable.
- Sheer amount of hate generated on a daily basis is simply too much.

There are NGOs training volunteers/operators to intervene with CNs.

Hence Computer Supported – Human activity can help them thanks to:

- Tools that assist NGO operators in fighting HS.
- Partial Automation of CN writing through Natural Language Generation (NLG) suggestions.
- Drastic reduction in time needed to compose a CN.

Algorithms for CN generation:
1. Information Retrieval
2. Generation using Neural Networks
3. Generation using pre-trained Generative Language Models
4. Generation using External Knowledge

A glimpse into the future

Multilingual aspects

- Each language/culture has its own nuances and prejudices.
- We might not have expert for each language
- Cross-language learning can help porting knowledge

Multi-Target aspects

- There are many target of hate that can benefit from our approach.
- Cross-target learning can help porting knowledge

Fake News and Hate

- What is the relation between Fake News and Hate?
- Can we use our tools and theoretical framework to fight Fake News as well?

Find out more: TENURE TRACK FINAL PRESENTATION

PERMALINK

https://magazine.fbk.eu/en/news/generating-counter-narratives-against-online-hate-speech-data-and-strategies/

TAGS

- #artificial intelligence
- #artificialintelligence
- #counter narratives
- #fake news
- #hatespeech
- #human language technology
- #natural language processing
- #research

RELATED MEDIA

- Tenure Track Discussion : https://magazine.fbk.eu/wp-content/uploads/2020/07/TENURE_FINAL_PRESENTATION.pdf

AUTHORS

- Giancarlo Sciascia