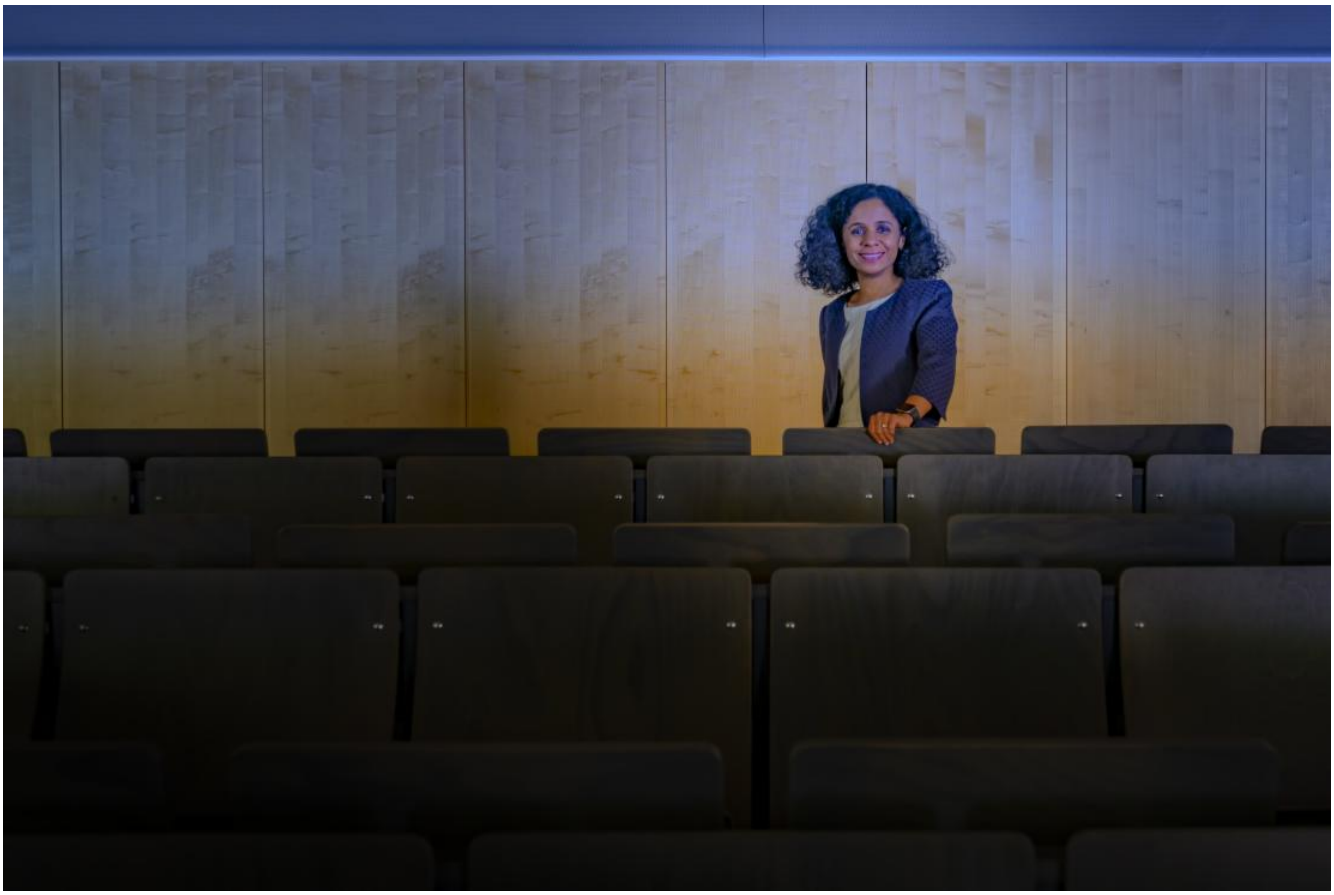
Horizon
The EU Research & Innovation Magazine

European Commission



Q&A: Why cultural nuance matters in the fight against online extreme speech

Artificial intelligence (AI) used by governments and the corporate sector to detect and extinguish online extreme speech often misses important cultural nuance, but bringing in independent factcheckers as intermediaries could help step up the fight against online vitriol, according to Sahana Udupa, professor of media anthropology at Ludwig Maximilian University of Munich, Germany.

18 January 2021 - By NATALIE GROVER

Factcheckers who operate independently of large media corporations or social media companies can shape and use AI to go beyond keywords to help locate context-specific patterns, according to Prof. Udupa. This is because they are trained to pick up disinformation — and extreme speech is a very close cousin of that, she says.

What motivated you to look into online abuse and extreme speech?

I was looking at online mediation of political cultures in India when I noticed the high prevalence of online abuse. Online abuse appears to be the language of today's politics.

Online speech has become such an influential way of experiencing politics today — not just participating in democratic processes like elections — but also the way we lead our daily democratic life through digital communication. If these sorts of irreverent and vitriolic exchanges are so prominent online, then we need to see how we can bring nuance to understanding them.

In some ways, online vitriol is presented as funny, but it could also lead to intimidation and shaming. We don't know exactly when the jokes stop, and the insults begin ... when the insults stop and when the intimidation starts. To understand this slippery slope, it's really important to understand the context.

How effective are extreme speech pushback mechanisms?

Increasingly, companies and governments are trying to deploy AI systems to combat the scale and speed of extreme speech. From my research, it is apparent extreme speech is heavily context dependent, and the datasets AI algorithms are based on are not.

Companies such as Facebook tend to look into instances of extreme speech when things get out of hand or when this context is extremely important, for instance, the US or Indian elections because it involves huge numbers of people.

However, corporate AI systems still lack the linguistic competence to detect problematic speech around the world — for example, in the northeastern Indian state of Assam (Facebook's) AI systems did not pick up an upsurge in rhetoric against religious and ethnic minorities in 2019. In contrast, the company beefed up resources and tried to recruit people who can speak the language to combat extreme speech in Myanmar, a region that grabbed international attention after the army cracked down on Rohingya Muslims (in 2017) sending thousands fleeing across the border into Bangladesh.

The way social media companies are tackling online extreme speech is therefore fragmented — both in terms of training datasets to train their AI models and implementing timely action during unfolding crises.

To address this unevenness we need to create collaborative frameworks that don’t just focus on English, Mandarin and Spanish, but include many different languages. We also have to go beyond a keyword-based approach and identify cultural and contextual markers.

The best way to do that is to mobilise and connect existing communities like factcheckers who can bring cultural nuance and contextual knowledge. They're trained to pick up disinformation, and extreme speech is a very close cousin of that.

How are you tackling extreme speech in your project AI4Dignity?

We are working on a community-based way to pick up problematic online extreme speech. The hope is to build a framework that can help factcheckers to flag this content without disrupting their core activities of reporting disinformation.

In the coming months we are requesting factcheckers to bring these annotated datasets and these will be the basis for training AI models. If everything goes well, in July we will bring factcheckers, academic researchers, and AI developers together. Afterwards, we hope to develop a tool based on multiple algorithms and integrate it with at least one particular platform or browser.

The model should be able to pick up some expressions that are problematic. It might not reach top-notch accuracy at the moment, but it's a stepping stone. The very dynamism of this process means it has to be repeated.

And we could replicate this process on a grander scale, with a wider network of factcheckers, bring in languages and dialects from more countries if more funding comes our way.

The other aim of the project is to also see if there are globally shared patterns of extreme speech. For example, criticism of legacy media has been well documented among different right-wing groups. But we want to actually investigate datasets and see if there are globally circulating tropes — if the Trump supporter, for instance, is actually providing discursive resources for the Hindu nationalist back in India, or if this anti-immigrant discourse is being picked up in Brazil and so on.

Projects like ours will help create critical knowledge that might not have applicability the very next day but will have long-term societal benefit. We are trying to create pushback mechanisms to regressive anti-immigrant and xenophobic discourses.

How stark are cultural differences in the expression of online extreme speech?

The gut feeling is that there's a lot of variation, but we also have documented this in our research. It's very clear that some expressions are very culturally rooted, as are target groups. For instance, it has been documented that people in northern Chile who are themselves marginalised try to peddle anti-immigrant discourse against people who come from places like Bolivia and Peru.

But when you look at a country like Denmark, their millionaires have supported far right-wing movements. So, there's this vast variation in who actually engages in extreme speech.

Complicating matters further, people who peddle derogatory and extreme speech practices engage in wordplay and adopt coded language to avoid detection.

And for us, that's important to understand how cultural variation is not just in the world of words, i.e. the kinds of expressions that people use, but also the actors who engage in them and the political structures that foster vitriolic exchanges.

Do you see patterns between extreme speech in homeland and diaspora communities?

Homeland communities and diaspora communities are closely connected because of internet channels so what we see is a sort of shared discourse. Whether you're in favor of a particular political ideology or not, expressions and tropes circle between the communities.

But there could still be cultural variation, I wouldn't rule it out. What is seen as extreme speech in one particular country might not be the case in a different country. For example, 'anti-national' could be a derogatory label in some countries — in others not so much. Labels themselves evolve within countries and regions.

Online extreme speech is often linked to offline violence. What does your research show?

Answering this question requires comprehensive case-study based field work and research like that has been undertaken in places like Kenya and South Africa suggesting that particular social media discourses could escalate conflict situations. There is also some data indicative of this association in North America and Europe. In India, I documented what was referred to a '"social media riot' because of the circulation of a video on WhatsApp before a planned protest rally. This was a complex social event since protesting Muslims were accused of being 'incited' by mashed up videos that claimed to depict violence in Myanmar and Northeast India.

But to actually pin down causality between online extreme speech and offline violence is very difficult. However, you can clearly identify trends and correlations. In certain cases, there will be a peak in extreme speech expressions prior to a violent episode.

This interview has been edited for length and clarity.

The research in this article was funded by the EU's European Research Council. If you liked this article, please consider sharing it on social media.

More info

ONLINERPOL

AI4Dignity