FONDAZIONE BRUNO KESSLER
MAGAZINE

Counter-narratives: paradigm shift in the fight against hate and fake-news

December 13, 2023

For a long time, it has been thought that Artificial Intelligence could be useful in censoring hate comments and fake news; however, when thought of in a supportive function to humans, it can be an effective tool for generating counter-narratives, a more democratic and in its own way "rational" means of responding to these contemporary challenges

Social networks have had a profound impact in the world of information: the two most notable changes are in the volume and direction of information, whsoe amount has increased by leaps and bounds. In the age of User Generated Content (UGC), millions of pieces of content are produced every day on Google, YouTube, Instagram, Facebook… and an even greater number of interactions (likes, dislikes, comments, replies, mentions, reactions…) shape an extremely fluid and changing debate. Information circulates more horizontally and less unidirectionally: compared to a past when there were only traditional media (TV, radio, newspapers), controlled by specific centers of power, which vertically and transmissively conveyed information, social networks have introduced a – relatively – more decentralized, vast and varied information network, in which consumers have become active users.

However, these changes, which are certainly positive when viewed from the perspective of democratizing society, have another side of the coin: hate speech and fake news. The ability of UGC to be created out of thin air, to be reshared and to gain resonance creates fertile ground for misinformation. A piece of content, no matter how much it is refuted by one party, can still be received by another; and even when a piece of news is concluded to be false to the general public, people can still continue to give it credence. Working on people’s emotions, fostering their loyalty to one’s own narrative and constantly providing alternative answers makes them more impervious to rational confrontation and critical thinking.

A similar argument applies to hate speech: having support, through feedback, on opinions one would be afraid to express in other contexts, along with echo chambers fostered by algorithms, creates self-feeding hate circle spaces.

In short, technology, as is often the case, causes both benefits and social problems: in the case of generative AI, as developed by our Digital Society, it can also be a means of analyzing and solving the latter.

Machine Learning and Classification

For a long time, hate speech and misinformation have been fought through classification-based machine learning algorithms: the goal of this approach is to make information filtering processes increasingly scalable and automated.
Of course, such an approach raises practical and ethical issues: algorithms are always trained on a finite dataset, and the forms of hate speech and misinformation can evolve over time, making it more difficult to track what is new. Updating algorithms is a long, time-consuming, and laborious process. Moreover, countering these new forms can also be problematic: laws may not be updated to respond to them, and algorithms may at most flag the possibility of the presence of hate or falsehood in a piece of content, without censoring it: this makes it less automatic.
Along with practical problems arise ethical, philosophical ones: what is the line between biased but legitimate opinion and hate speech? How to sort content out when truth and fiction are mixed? Where is the balance point between censorship and free speech?
Delegating decisions on similar questions to machines, especially in more nuanced cases, can be problematic. That's why a revolutionary new approach is emerging.

"Human-centric" AI and counter-narratives.

Generative AI is being developed at FBK to support web users in fighting hate. For example, if the algorithm on the Facebook page of a migrant safeguard NGO detects a hateful phrase, a chatbot can promptly process a text in response to this comment. However, the response is not automatic: the text is processed as advice to an operator, who can construct its response based on that text. AI and human beings integrate seamlessly to construct effective counter-narratives to hate speech; responses suggested to AI are on average rational and respectful, consistent with an image line of the organization, aimed at making people think about the implications of what they say. The "attacked" opinion is not censored, but passed through the critical scrutiny of a counter-narrative.

To achieve these requirements, training data are constructed ad hoc by researchers and experts in the field: since algorithms have no real understanding of the text, but build their semantics based on a probabilistic distribution of words frequently associated with others in a given syntactic context , it would be risky to start from "real" data.

As for fake news, artificial intelligence systems, when given a post to disprove, have a related debugging article at their disposal, with which they construct counter-arguments. A recent line of research concerns the ability to give responses an emotional tone, obtained through sentiment analysis of comments to be responded to.

As an example, consider a news story reported by an online news outlet during the pandemic. The same news story, conveyed through social media, generated several comments from online users. The image below shows two different responses to the same comment. These responses were derived from counter-narrative suggestions automatically produced by artificial intelligence. Specifically, in the first case the response is "cold and rational," while in the second case it is "emotionally toned."

207

Mi piace

Scrivi un commento...

Mario Rossi

Mio zio si è seriamente ammalato 3-4 giorn
in terapia intensiva con Covid in stato avan
iniettano una versione indebolita del virus
qualcuno che vuole che sia così... io intanto

2 a Mi piace Rispondi

Fredda e razion

Il principio di immunizzazione dei va
versione inattiva o indebolita del viru
nel caso dei vaccini anti-covid Pfizer-
di RNA messaggero che fornisce istru
una proteina simile a quella presente
Questa proteina stimola una risposta
i quali sono proteine che possono ric
Pertanto è scorretto sostenere che i v

There is still a long way to go for human-AI integration; however, this case of collaboration between algorithmic efficiency and human sentience is an excellent example of how the two dimensions can go hand in hand, shaping the more general "trustworthy & human-centric AI" that the Horizon project, in its latest update dated 2021, defines as the future research horizon.

PERMALINK

https://magazine.fbk.eu/en/news/counter-narratives-paradigm-shift-in-the-fight-against-hate-and-fake-news/

TAGS

- #fake news
- #fbkdictionary
- #Intelligenza artificiale
- #società digitale
- #societàdigitale

AUTHORS

- Lorenzo Perin