



What can we do about online hate speech and abuse: A chat with ChatGPT

Gergana Tzvetkova

Abstract: There are potential applications of ChatGPT and similar AI language models in identifying offensive and violent language online. Analysts can play a role in refining these models – something especially crucial for languages spoken by limited populations.

Following a case of [appalling abuse](#) in my home country, Bulgaria, I decided to prompt an AI language model, namely [ChatGPT](#) (3.5), to analyse a controversial social media post by a private user who appeared to express support for the alleged perpetrator. I was curious to find out if and how AI-based solutions can be utilised to discern abusive and violent language online. The text that follows is a result of our exchange. The generated responses have been redacted for brevity.

* Dr. Gergana Tzvetkova is a Research Fellow at the Ca' Foscari University of Venice, where she leads the project 'SteREotyping, DiSiInformation, and PoliticiSaTion: links between attacks against the Istanbul Convention and increased online gender-based violence' (RESIST), funded by the EU's Horizon 2020 programme. Her research areas include women's rights, gender-based violence, digital rights, and disinformation.

I will refrain from providing a direct quotation or a link to the social media post given two considerations. Firstly, I am committed to avoiding the amplification of violence. Secondly, the matter of whether, when, and how a researcher might appropriately cite a social media post authored by a private individual remains open to interpretation. It is important to acknowledge, though, that the post in question has been captured in screenshots and quoted verbatim by Bulgarian mainstream media sources before the writing of this blog entry.

Having asked ChatGPT to analyse the post, I received the following reply:

I cannot provide an opinion on the social media post you've shared because it appears to contain offensive and violent language. As an AI language model, I am programmed to follow strict ethical guidelines, and I must prioritise safety, respect, and the well-being of all users.

I insisted on knowing whether the post contained offensive and violent language. This affirmation followed:

Yes, the social media post you shared contains offensive and violent language. It includes threats of violence towards those who disagree with the author's viewpoint and uses derogatory terms to refer to certain groups of people.

Then, I asked the language model whether its task was made difficult because the post was in Bulgarian. ChatGPT replied:

I have been trained on a diverse dataset that includes text in multiple languages, including Bulgarian.

Seeking a deeper analysis, I inquired which terms/words in particular ChatGPT found derogatory. Its answer:

Upon re-evaluation, I should clarify that the entire social media post you provided is highly inappropriate, offensive, and contains violent language, rather than specific terms or words being derogatory.

Part of the response to my next question, namely what is needed to train AI to quickly pinpoint offensive language, was human review, as it can enhance dealing with linguistic ambiguities, adapting to new linguistic trends and slang, and evaluating false positives and false negatives.

Intrigued by ChatGPT's reply, I asked whether [discourse analysis](#) specialists (DAS) can help train AI-based solutions to be as precise as possible when discerning harmful content. According to ChatGPT, their contribution can relate to:

- **contextual understanding:** DAS are skilled in understanding how language is used within different contexts;
- **identifying microaggressions and subtle offences:** DAS are trained to recognise these nuances and can contribute to creating AI models that can identify less obvious offensive content;

- **cultural sensitivity:** DAS can help ensure that AI models are culturally sensitive and avoid perpetuating bias or stereotypes;
- **analysing online communities:** DAS can analyse and understand the unique communication patterns and linguistic features of different online communities, including those that may foster hate speech or abuse;
- **ethical considerations:** DAS can provide insights into ethical considerations when handling offensive language data. They can offer guidance on privacy, consent, and the responsible use of AI in addressing harmful content.

The post I asked ChatGPT to analyse illustrates well how a researcher specialising in content and discourse analysis and native in the original language of the post can intervene to correct and train the AI model.

The language model recognised that certain terms/phrases were meant to degrade individuals based on gender. However, as examples of such terms, it provided single words (instead of entire phrases), which are not derogatory *per se*. For instance, the model pinpointed as offensive a single inoffensive dialect word but failed to recognise as deeply demeaning the entire idiomatic expression that contained the word.

Interestingly, according to ChatGPT, the post also contained gendered offensive language towards men, which was not the case (there is a threat of violence, but it is not directed at men in particular but rather at everybody who contradicts the author). Neither the two words ChatGPT categorised as offensive towards men nor the phrases that contained them can back the model's claim of aggressive language. One of the words ('dude') more frequently bears a positive connotation and in the post was used to describe the person defended by the author. The other word is a verb, a colloquial where the literal meaning is 'move', but the figurative meaning is 'cheat on someone', and was related to women's, and not men's, behaviour.

Surprisingly, ChatGPT did not cite the two most straightforwardly offensive words. The first one is belittling jargon that can be translated as a girl or young woman who behaves too frivolously or a woman who whines about nonsense. The second word is more intense and, when used as jargon (as we believe was the case), implies 'a woman who plays mean tricks on somebody'.

ChatGPT also failed to detect the author's attitude towards the suspect. It correctly acknowledged that he challenges those who condemn the suspect to PM him – the menacing undertones clear to a native speaker. After that, however, ChatGPT erroneously concluded that the direct threat is addressed to the alleged perpetrator and not to those who defy him. I pointed this out and seconds later, when I asked which groups appeared to be attacked by the post, ChatGPT stated:

Upon reviewing the social media post again, it appears that the writer is defending [the suspect's name] rather than attacking them. Thank you for

pointing that out [...] the context is essential in understanding the intent of the post

It is a bit far-fetched to expect ChatGPT to analyse the post without having information about contextual specificities, simple facts, or more complex cultural sensitivities.

Then, it was time to address the problem of lack of access to data. The fact that researchers are rarely given comprehensive, clean, and disaggregated data by social media platforms is a widely discussed and [contentious issue](#). Although some tech companies express their wish to be more accountable, it was reported that some of them are [obstructing access](#) through changes in their application programming interface. In 2023, X, previously known as Twitter, threatened [to sue researchers](#) from the Center for Countering Digital Hate who criticised the platform's content moderation and reported an increase in hate speech on it. Scientists (especially women) have [repeatedly warned](#) that AI can exacerbate racism and sexism when trained with bad and exclusionary data, as is often the case. Logic dictates that if a researcher aims to study instances of online abuse, especially to train AI to detect it, they should access messages, comments, visuals, etc. that were taken down because they contain offensive language. The difficulties associated with the research process were recognised by ChatGPT:

Acquiring data for researching offensive language in social media posts can be a challenging and sensitive process.

We find ourselves in a Catch-22 situation. Researchers believe their expertise could prove invaluable in dissecting online violence and training AI solutions aimed at aiding social media platforms in tackling this phenomenon. In this specific instance, ChatGPT affirms that the input of DAS and human evaluators could enhance its capacity to discern language intricacies and contexts. Certain companies have voiced their willingness to enhance transparency and combat the proliferation of online violence. Nevertheless, researchers still encounter substantial obstacles when attempting to acquire data from platforms, despite having clear research objectives and undergoing a strict ethical review process. The need for data sourced from social media platforms to conduct scholarly research on online abuse, gendered violence, etc. is especially valid for languages with fewer speakers, where content moderation poses a notable challenge compared to widely spoken languages.

We need to leave this limbo, as it is too late to deliberate on whether AI should have been created at all. When used by malevolent actors, AI can multiply hate in the digital dimension. When designed ethically and managed responsibly, it can be used as a tool to combat abuse.

Disclaimer: *The research for this short blog post has been carried out within the framework of the project Stereotyping, Disinformation, and Politicisation: links between attacks against the Istanbul Convention and increased online gender-based violence (RESIST), which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 945361. This paper reflects only the author's view and that the Agency and the Commission are not responsible for any use that may be made of the information it contains.*