Popping the Social Bubble: Using AI to Increase Transparency in Harmful Content Moderation

Argentina Anna Rescigno[1,2][0009-0000-3653-8492], Beatrice Melis[1,3][1111-2222-3333-4444], Francesco Di Cursi[1,4][0009-0005-9782-478X], Gianluca De Ninno[1,3][2222--3333-4444-5555], Gianmarco Pastore[1,5][2222--3333-4444-5555], and Paolo De Biase[1][2222--3333-4444-5555]

[1] Università di Pisa, Pisa, PI, 56126, Italy
{argentina.rescigno, beatrice.melis, francesco.dicursi, gianluca.deninno, gianmarco.pastore, paolo.debiase}@phd.unipi.it
[2] Università degli Studi di Napoli Orientale, Napoli, NA, 80121, Italy
[3] Gran Sasso Science Institute, L'Aquila, AQ, 67100, Italy
[4] Istituto di Informatica e Telematica - CNR, Pisa, PI, 56124, Italy
[5] Università Cattolica del Sacro Cuore, Milano, MI, 20123, Italy

Keywords: Social media moderation, Toxicity detection, Harmful contents, Content moderation, Large language models

1 Long Abstract

Content moderation in Online Social Networks (OSNs) is an increasingly complex task due to the growth of generated content resulting from the massive increase in the user base of such services. The crucial aspect of the problem has led big companies to hire specialised teams in order to make the service (i.e., the Online Social Network) accessible to potentially anyone; furthermore, in some cases, this task is performed under a nondisclosure agreement due to the very disturbing nature of such contents[1][2]).

The moderation can happen essentially in two ways: (1) preventive censorship through automated attempts in case of potentially disturbing content, or (2) moderation upon users' reports. In the former case, the challenge of identifying genuinely problematic contents (e.g., semantic reappropriation of slurs by minorities, satire, ...) could result in the excessive temporary removal of posts. Consequently, an OSN might display only a small *politically correct* fraction of content, giving the user the impression of minimally active OSN platforms. On the other hand, relying on users to report problematic content implies that some of them need to stumble upon it before it can be reported, causing potential harm to the user[3].

While the mental health of professional moderators cannot be spared and must remain a priority, a way to prevent users from encountering such harmful content while maintaining the quantity of the generated content involves temporarily rephrasing possibly problematic content. This rephrased version would be shown to users and reported to both the user and the professional moderator, allowing the latter to manually evaluate and take appropriate action on the content. In

other words, automation is used to ease the experience for both the users - by shielding them from exposure to harmful language while preserving the content's meaning - and the professional moderators - by giving them more time to perform their tasks, complying with the principle of the *man in the loop.*

We downloaded a Kaggle dataset consisting of highly problematic tweets[4]. A prompt is built so that, upon receiving the original text as input, it: (i) rephrases the text to make it non-toxic while trying to preserve the original meaning; (ii) assigns a toxicity score ranging from 0 to 2 where 0 is non-toxic, 1 is toxic and 2 is very toxic; (iii) clarifies which part of the text makes the model evaluate its toxicity and (iv) to which stereotype it is ascribable; (v) specifies the keywords that lead the model to detect the stereotype, and finally (vi) indicates the type of conveyed meaning (e.g., literal, metaphorical, sarcastic, ...).

However, our first attempt has shown some limitations. A noteworthy one is that the model is not able to handle highly toxic content (i.e., toxicity = 2). In other words, the model can successfully rephrase a text removing toxicity from the style while preserving it in the content only in cases of mild toxicity (Fig. 1). On the other hand, in cases of severe toxicity, it just replaces bad words with more general terms, failing at finding a *more acceptable way* to express a harsh thought (Fig. 2). Another important weakness concerns information loss, for instance, in cases of the semantic reappropriation of slurs (i.e., failing to understand the socio-cultural context in which the slur is used). Moreover, even if this approach manages to avoid excessive preventive shadowing on a platform and the user-based report system, there is no way to exempt the professional moderator from manually evaluating toxic content. To conclude, we noticed that ChatGPT4 is not the best choice to deal with toxicity, discrimination, hate speech and harmful stereotypes, possibly due to its training policies.

Our work, summarized in Figure 3, is thus an attempt to combine the avoidance of excessive politically correct censorship while mitigating unnecessary toxicity, particularly w.r.t. the form in which the content is conveyed, preserving the underlying meaning. In this way we establish a soft censorship “from below", by considering users' perspective towards the service, and not vice-versa, empowering them to have control over their experience. Furthermore, this approach could ensure explainability and transparency, as opposed to the current black-box policy of several big companies; such transparency would enable users to understand the extent and nature of toxicity present on a specific platform.

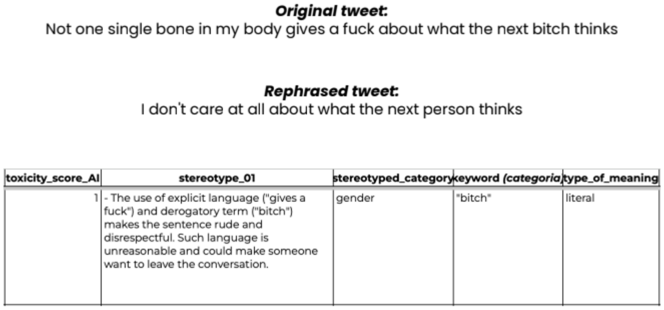
In future efforts, we aim to utilize an *ad-hoc* tool (such as Perspective API[5] for text) to enhance the accuracy of toxicity detection, rather than relying on a broad bibliographic definition, where toxicity is defined as content that prompts individuals to leave a conversation [6]; moreover, considering a general-purpose LLM like ChatGPT-4, we may still utilize it for aspects of the prompt focused on explainability (specifically, from points iii to vi).

Another meaningful advancement will involve dealing with any type of content, such as photos, videos, and audio. Most importantly, this initiative will be far from remaining just a conceptual study, but it will be applied through a web extension to any given platform, allowing users to personalise their service experience. In other words, users will be able to establish a set of filters in order to decide several key factors such as toxicity thresholds (e.g, content requiring rephrasing versus direct removal), the style of rephrasing (e.g., methods for cover-up images in case of NSFW photos), a list of unwanted words and stereotypes and so on. The user will also have the possibility to create several types of such customized filters to set them accordingly if necessary (e.g., in case of a computer shared between different users, such as parents and children).

Lastly, and most significantly, users will be provided with a set of statistics (e.g., number of rephrases, instances of content shadowing, and the relative words and stereotypes of flagged content) in the form of a leaderboard. This transparency aims to help users understand the nature of the actual content on the platform, despite being shielded from undesirable material. Therefore, our core purpose is to foster a safe yet representative space that facilitates genuine democratic debate within communities.

As final considerations, while acknowledging that the proposed approach is not fully automated, we aim to utilise tools such as PerspectiveAPI and ChatGPT-4 (with ongoing investigations into more advanced LLMs to handle toxicity optimally) to identify problematic content. This content can then be reshaped to preserve its meaning whenever possible (i.e., in case of civilized thoughts expressed uncivilly), thus minimising users' exposure to unnecessary toxicity while maintaining free access to information.

Nevertheless, professional moderators remain essential and cannot be excluded from the loop due to the ambiguous nature of toxicity; this approach can reduce their constant exposure to harmful content, benefiting their mental health. As a matter of fact, a common challenge in moderation is the need for swift removal of potentially harmful content to protect users from stumbling upon it, which requires moderators to be constantly exposed to NSFL material. The core idea of the project is to avoid *hard censorship*, ensure freedom of speech according to ToS, and find a balanced solution through content reshaping for the sake of ethical and civilized debate. In this scenario, users will ultimately decide whether to rephrase the content or not, via a dedicated browser extension.

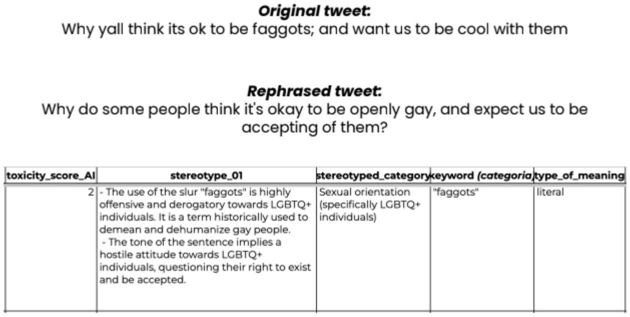


Original tweet:
Not one single bone in my body gives a fuck about what the next bitch thinks

Rephrased tweet:
I don't care at all about what the next person thinks

toxicity_score_AI	stereotype_01	stereotyped_category	keyword *(categoria*	type_of_meaning
1	- The use of explicit language ("gives a fuck") and derogatory term ("bitch") makes the sentence rude and disrespectful. Such language is unreasonable and could make someone want to leave the conversation.	gender	"bitch"	literal

Fig. 1. Succesfull rephrasing



***Original tweet*:**
Why yall think its ok to be faggots; and want us to be cool with them

***Rephrased tweet*:**
Why do some people think it's okay to be openly gay, and expect us to be accepting of them?

toxicity_score_AI	stereotype_01	stereotyped_category	keyword *(categoria*	type_of_meaning
2	- The use of the slur "faggots" is highly offensive and derogatory towards LGBTQ+ individuals. It is a term historically used to demean and dehumanize gay people. - The tone of the sentence implies a hostile attitude towards LGBTQ+ individuals, questioning their right to exist and be accepted.	Sexual orientation (specifically LGBTQ+ individuals)	"faggots"	literal

Fig. 2. Failed rephrasing

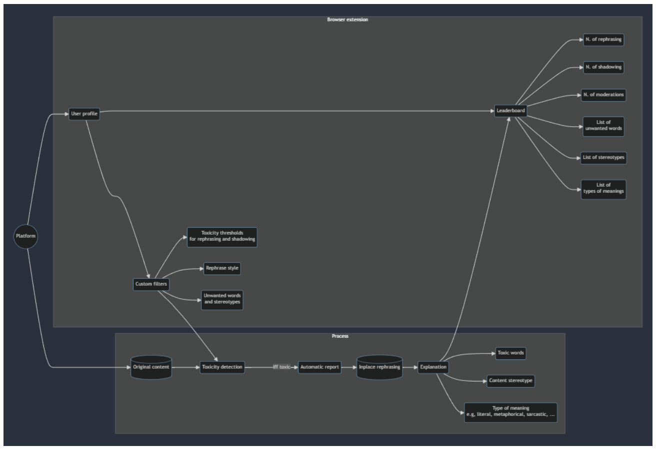


Fig. 3. Project pipeline

References

1. Alex Hern, *Facebook moderators 'told not to discuss working conditions'*, The Guardian, 2021, url: https://www.ft.com/content/afeb56f2-9ba5-4103-890d-91291aea4caa, accessed 1 July 2024
2. VICE, *The Horrors of Being a Facebook Moderator,* Youtube, 2022, url: https://www.youtube.com/watch?v=cHGbWn6iwHw, accessed 1 July 2024
3. Ulvi O, Karamehic-Muratovic A, Baghbanzadeh M, Bashir A, Smith J, Haque U. (2022). Social Media Use and Mental Health: A Global Analysis. Epidemiologia; 3(1):11-25. url:https://doi.org/10.3390/epidemiologia3010002, accessed 1 July 2024

4. Andrii Samoshyn, *Hate Speech and Offensive Language Dataset*, 2020, url:https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset, accessed 1 July 2024
5. Google Jigsaw, 2017, Perspective API, https://www.perspectiveapi.com/, accessed 1 July 2024
6. Borkan, D., Dixon, L., Sorensen, J., Thain, N. and Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification, WWW '19: Companion Proceedings of the 2019 World Wide Web Conference, 491–500. doi: https://doi.org/10.1145/3308560.3317593, accessed 1 July 2024