Empirical investigation of industry-based artificial intelligence (AI) moderation tools for online bullying and harassment

Kanishk Verma[a,b], Dr Tijana Milosevic[a,b], Dr Brian Davis[b], Prof. James O'Higgins Norman[a]
[a]Anti Bullying Centre, [b]ADAPT Centre, Dublin City University, Ireland

1. Introduction

- AI-powered content moderation for bullying and harassment is on the rise on Meta (Facebook, Instagram) and YouTube (See Figure 1), but does it really work?
- Since it is proprietary information, one telling metric of the efficacy of this moderation is the number of appeals[1] which has been increasing in the recent years (See Figure 2).
- In the interest of "transparency" Meta AI and Google-Jigsaw have released tools that aid in the identifying toxicity of texts, which subsequently assists in proactive moderation.
- We delve into uncharted territory by comparing the efficacy of Meta AI's OPT [1] and Google-Jigsaw's Perspective API [2] for toxicity detection as a precursor for cyberbullying identification.
- Task at hand?
    - To understand the efficacy of Meta AI's, OPT and Google-Jigsaw's Perspective API in detecting toxicity in real-life cyberbullying texts.

[1] Appeals are options for users to contest removals or restrictions in case they think social media platform made a mistake in taking down content.

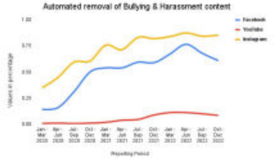


Figure 1. Automated removal of Bullying & Harassment content

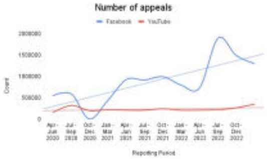


Figure 2. Number of appeals for automated removed content for bullying & harassment

Label	# sentences used in empirical investigation	Examples
Insult	1,641	'Ur not pretty at all' 'get it right, sillySL*T'
Curse or Exclusion	1,045	'off you, go away you waste of oxygen'
Attack	110	'nearly as wet as ur mum last night lmao..'
Threat or blackmail	180	'not if I ruin u first ;)'
Non-toxic	11,499	'nooo I was kiddin, sry, I assumed u knew I was kiddin'

Table 1. Details of labels and examples of sentences directly taken from [4], [5]

2. Hypothesis & Methods

- Hypothesis
    - Closely following established thresholds [1,3], we hypothesize that Perspective and both trained[2] and untrained OPT (zero-shot)[3] systems will detect toxicity in sentences labelled as `insult, curse, attack or threat`, with at least 70% probability or higher.
- Assessing Perspective & OPT:
    - We test the ability of both systems to identify toxic content using data from ASK.fm and WhatsApp collected by [4] and [5], respectively.
    - We leverage 3 types of systems, Perspective available via API, OPT as an un-trained system in its raw form (zero-shot), OPT further trained on the same data [6-12] as Perspective.
    - Specifically, we evaluate their performance on 14,475 sentences labelled for different forms of online harm associated with cyberbullying. (See Table 1 for details)
- Statistical significance test:
    - We leveraged Wilcoxon signed-rank test [13] to test the statistical significance of the difference between the threshold (70% probability) and the probability scores provided by Perspective, and both trained and untrained (zero-shot) OPT systems.
- Statistical error analysis:
    - We select 10% random samples of text from [4], [5] without explicit profanity but contain sarcasm, irony, and modification to words, then evaluate the system's efficacy in detecting them as toxic or non-toxic.

[2] Trained OPT involves augmenting OPT's knowledge through additional data for classification.
[3] Zero-shot OPT means testing OPT's existing knowledge by classifying without further training.

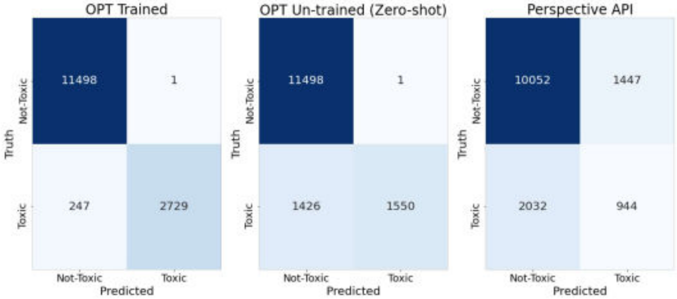
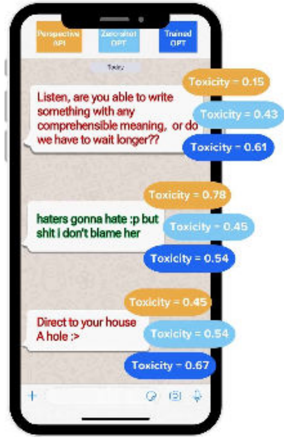


Figure 3. Performance of 3 systems when assessed on 14,475 sentences with toxic and non-toxic markers

3. Results & Discussions

- Statistical significance test:
    - All three systems Perspective and both trained and un-trained (zero-shot) OPT did identify toxicity with at least 70% probability (p<0.05). Thereby, confirming validity of our hypothesis.
- Sensitivity (true positive) & Specificity (true negative):
    - These are two key markers to validate how good a detection system is at identifying true positives (toxic sentences) and true negative (non-toxic sentences).
    - In Figure 3, as we move from left to right, the OPT system, with additional training, achieves 91.7% sensitivity and 99% specificity. In its untrained state (zero-shot), it exhibits 52% sensitivity and 99% specificity. Meanwhile, the Perspective API system shows 31.7% sensitivity and 70.5% specificity.
    - This indicates, OPT systems, especially after additional training, demonstrate reduced Type 1 and Type 2 errors compared to Perspective.
- Our investigation reveals that training Meta AI's OPT on toxic data enables it to outperform Perspective in `insults, curse, attack or threat` sentences labelled by [4-5].
- By making these tools available for researchers, Meta and Google etc., are taking the first steps towards transparency in online moderation.
- While they are getting better at detecting toxicity, they struggle with non-harmful profanity, and mean or offensive language without explicit profanity. (See Figure 4)



Listen, are you able to write something with any comprehensible meaning, or do we have to wait longer??

haters gonna hate :p but shit I don't blame her

Direct to your house A hole :>

Figure 4. Error Analysis

References

Paper & Project Details

Funding Acknowledgement

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn
Google