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

Kanishk Verma, Tijana Milosevic, Brian Davis, Dr James O'Higgins Norman

Introduction:

Online social networks (OSNs) have terms of service that prohibit cyberbullying, cyber harassment, and cyber aggression on their platforms. Most OSNs use a combination of user flagging[1] and artificial intelligence (AI) to detect and moderate such content. There is increasing reliance on AI enabled proactive detection, which has the benefit of alleviating user harm through reduced impressions of abusive content. Meta (Facebook, Instagram), YouTube, and TikTok all report an increase in reliance on AI to moderate online bullying content [1]–[3]. Despite these efforts, little is known about the efficacy of such proactive moderation of bullying content on social media. In the interest of collaborative research with independent researchers, Meta AI[2] and Google Jigsaw[3] have released tools that aid in the classification of texts, which subsequently assist in proactive moderation. To classify text as a potential form of cyberbullying, harassment or aggression, it is important to identify if the linguistic content is offensive i.e. toxic, threatening, sexually explicit, defamatory, or implies blackmail, as a precursor to identifying cyberbullying.

This is the first study which empirically investigates and compares the efficacy of two industry-led AI language models[4] for detecting toxicity in text: Google & Jigsaw's Perspective API[5] [4], and Meta's OPT[6] [5] (released in July 2022). Additionally, to define texts as toxic we adopt Jigsaw's definition and classify text as toxic if it is rude, disrespectful or unreasonable. [4] We evaluate the efficacy of both systems on real-world English language datasets containing cyberbullying texts [6]–[8]. In this study, we use toxicity as a precursor for online bullying detection to identify if industry systems can identify toxicity in texts labelled as `insult`, `curse`, `attack`, and `threat` by human annotators. Our analysis aims to enable greater transparency and accountability of the social media industry's efforts at cyberbullying detection, which is currently lacking.

[1] User flagging is when user reports a content on OSN as per reporting policies of the platforms.

[2] Meta AI is the research wing for Meta (Facebook, Instagram, WhatsApp) social media platforms: https://ai.facebook.com/

[3] A unit within Google that devises technological solutions for threats to society: https://jigsaw.google.com/

[4] Language models are AI systems that can predict the probability of pattern or sequence of words occurring in a sentence.

[5] API is called an application programming interface that allows users to access softwares via the internet without requiring to download or install the software.

[6] OPT is a language model with many parameters made publicly available by Meta AI https://github.com/facebookresearch/metaseq/tree/main/projects/OPT

Method:

When given a text, Perspective API using machine learning provides a probability score indicating the probability that someone will perceive the text as toxic. [4], [9]. Meta AI's OPT is a pre-trained language model, i.e., an AI system which has already learnt how to represent words contextually and semantically by using statistical probabilistic techniques and neural networks. Authors of OPT [5] claim[7] an accuracy of 67% to 81% in identifying hate speech-related text and suggest the performance improves when OPT is further trained on additional hate speech datasets.

We evaluate OPT as two systems, i) in its raw form as shared by [5] and ii) by further training[8] OPT on datasets shared publicly by Perspective [13]–[19]. Leveraging the comment analyser tool [10]–[12], we assess Perspective's effectiveness in identifying toxicity in texts. [6]–[8] devised a novel cyberbullying text dataset from two different social media platforms, Ask. fm[9] and WhatsApp[10]. The text in these datasets is categorised by domain experts as personal attacks, threat or blackmail, defamation, discrimination or insult, as described in Table 1 (See Appendix). In line with thresholds discussed by Perspective [20], we hypothesised that both systems for every message labelled as an insult, curse, attack, and threat will identify textual toxicity with a probability of 70% or more.

To that effect, we assess the efficacy of both Perspective and OPT in detecting textual toxicity by recording the toxicity probability scores for 1,641 messages labelled as an insult, 1,045 messages labelled as a curse, 110 messages labelled as attacking, and 180 messages labelled as a threat by [6], [8]. Moreover, for statistical error analysis of the systems we select 10% random samples of text messages from [6], [7] that do not have any markers for any type of online harm and assess the system's efficacy in detecting such messages as non-toxic.

Results:

A Wilcoxon signed-rank test [21] was used to test the statistical significance of the difference between the threshold (70% probability) and the probability scores provided by Perspective API, and the two OPT systems for every text. Results in Table 2 (See Appendix) show both systems did identify textual toxicity for most messages labelled as *insult, curse, attack*, and *threat* with more than 70% probability ($p < .05$).

[7] Table 3, pg 6 in [5]

[8] By training, we mean making the language model learn new textual and semantic representations by showing it data it hasn't seen before.

[9] https://ask.fm : A social media platform, where users can be anonymous and share, post, and chat online.

[10] https://www.whatsapp.com : A direct-messaging social media platform.

*Sensitivity a*nd *specificity are* two key markers to validate how good a detection system is in identifying true positives (toxic sentences) and true negatives (non-toxic sentences), respectively. [22], [23] Results depicted in Table 3 (See Appendix) show that Meta AI's OPT language model when further trained does outperform Perspective API. On closer examination of texts in the datasets leveraged for this experiment, we find that the Perspective API does not detect toxic messages lacking profanity or curse words as effectively as messages with profanity. (See Table 4 & Figures 8-11 in Appendix). For instance, one of the sentences containing irony or sarcasm, such as: "`Listen, are you able to write something with any comprehensible meaning or do we have to wait much longer?`" is mistakenly labelled by Perspective as non-toxic. Despite outperforming Perspective, OPT identifies non-toxic sentences such as *"`haters gonna hate :p but shit I don't blame her`"*, as toxic with a probability of 73%. This shows that even better-performing OPT fails when profanity exists in non-toxic sentences.

Moreover, in line with results by independent researchers [24], we found that Perspective API performed poorly for sentences when profanity was represented using a mix of textual modifications to words. For example, both OPT and Perspective fail to recognise toxicity in the text labelled insult: "`Direct to your house A hole :>`". Additionally, for sentences with profanity but not toxic, for example, "`Hahahaha omfg, really why? And I am too f**king cringe`" or "`Sushi tastes like shit`", both OPT and Perspective fail to recognise them as non-toxic.

Conclusion:

Our investigations and comparisons reveal that the OPT system in its raw form is poor in toxicity detection compared to Perspective API. However, this is expected behaviour as the OPT model shared in its raw form needs to be tuned for toxicity or other detection tasks. Overall, results denoted in Figures 1-4 (See Appendix) indicate that further training Meta AI's OPT on toxicity datasets yields better results in recognising toxicity for text messages labelled as an `insult`, `curse`, `attacking`, and `threat` than the Perspective API. Moreover, Figures 5-7 (See Appendix) show that further trained OPT systems misinterpret non-toxicity but are better at detecting toxicity than the Perspective API.

By making these text classification tools publicly available, social media companies are taking the first step towards transparency in online moderation. Our empirical investigations reveal while the tools are good at identifying toxicity, they fail a) when profanity or curse words have subtle modifications i.e., mixed use of capitalization, symbols, or numbers, and b) when profanity is used in the text but means no harm, c) when texts lack profanity but are mean, insulting or nasty.

Motivated by [25]–[27], such investigations coupled with thorough analysis, and stress testing of industry-led initiatives designed for online moderation of bullying and harassment help us understand their efficacy and pitfalls on real-world online bullying texts. This work has demonstrated failure points for both systems, which speaks to the importance of public scrutiny, which in turn should help social media platforms improve their internal moderation systems for a safer online experience. Planned future research will incorporate Meta AI's WPIE and RIO tools [28], [29] to analyse its performance on real-life cyberbullying multimodal (image, text, video) datasets collected by [30], [31] from Instagram and now-defunct Vine.

References:

[1] 'Community Standards Enforcement | Transparency Center'. https://transparency.fb.com/data/community-standards-enforcement/bullying-and-harassment/facebook (accessed Feb. 28, 2022).

[2] 'Community Guidelines Enforcement Report Apr - Jun 2022 | TikTok'. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2022-2/ (accessed Dec. 07, 2022).

[3] 'YouTube Community Guidelines enforcement – Google Transparency Report'. https://transparencyreport.google.com/youtube-policy/removals (accessed Dec. 07, 2022).

[4] 'Perspective API'. https://perspectiveapi.com/ (accessed Dec. 07, 2022).

[5] S. Zhang *et al.*, 'OPT: Open Pre-trained Transformer Language Models', May 2022, doi: 10.48550/arXiv.2205.01068.

[6] C. V. Hee *et al.*, 'Automatic detection of cyberbullying in social media text', *PLOS ONE*, vol. 13, no. 10, p. e0203794, Oct. 2018, doi: 10.1371/journal.pone.0203794.

[7] 'Leveraging machine translation for cross-lingual fine-grained cyberbullying classification amongst pre-adolescents | Natural Language Engineering | Cambridge Core'. https://www.cambridge.org/core/journals/natural-language-engineering/article/leveraging-machine-translation-for-crosslingual-finegrained-cyberbullying-classification-amongst-preadolescents/AC24445BCC1EBA67E9E9A92A247D8123 (accessed Dec. 08, 2022).

[8] R. Sprugnoli, S. Menini, S. Tonelli, F. Oncini, and E. Piras, 'Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying', in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, Oct. 2018, pp. 51–59. doi: 10.18653/v1/W18-5107.

[9] 'Perspective API - Case Studies'. 3 (accessed Dec. 08, 2022).

[10] 'About the API - Attributes and Languages'. https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages (accessed Feb. 28, 2022).

[11] 'About the API - Methods'. https://developers.perspectiveapi.com/s/about-the-api-methods?language=en_US (accessed Dec. 08, 2022).

[12] 'About the API - Key Concepts'. https://developers.perspectiveapi.com/s/about-the-api-key-concepts?language=en_US (accessed Dec. 07, 2022).

[13] 'Wikipedia Talk Corpus'. figshare, Jan. 17, 2017. doi: 10.6084/m9.figshare.4264973.v3.

[14] 'Wikipedia Talk Labels: Personal Attacks'. figshare, Feb. 22, 2017. doi: 10.6084/m9.figshare.4054689.v6.

[15] ipavlopoulos, 'Toxicity detection w/ and w/o context'. Nov. 30, 2022. Accessed: Dec. 07, 2022. [Online]. Available: https://github.com/ipavlopoulos/context_toxicity

[16] 'The Unhealthy Comments Corpus (UCC)'. Conversation AI, Nov. 30, 2022. Accessed: Dec. 07, 2022. [Online]. Available:

https://github.com/conversationai/unhealthy-conversations

[17] ‘Jigsaw Unintended Bias in Toxicity Classification’. https://kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification (accessed Feb. 28, 2022).

[18] ‘Jigsaw Multilingual Toxic Comment Classification’. https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification (accessed Dec. 07, 2022).

[19] V. Kolhatkar, N. Thain, J. Sorensen, L. Dixon, and M. Taboada, ‘Classifying Constructive Comments’, p. 24.

[20] ‘About the API’. https://developers.perspectiveapi.com/s/about-the-api-score?language=en_US (accessed Dec. 08, 2022).

[21] D. W. Zimmerman and B. D. Zumbo, ‘Relative Power of the Wilcoxon Test, the Friedman Test, and Repeated-Measures ANOVA on Ranks’, *J. Exp. Educ.*, vol. 62, no. 1, pp. 75–86, Jul. 1993, doi: 10.1080/00220973.1993.9943832.

[22] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas, ‘Understanding and using sensitivity, specificity and predictive values’, *Indian J. Ophthalmol.*, vol. 56, no. 1, pp. 45–50, 2008.

[23] M. Taggart *et al.*, ‘Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients’, *JAMA Netw. Open*, vol. 1, no. 6, p. e183451, Oct. 2018, doi: 10.1001/jamanetworkopen.2018.3451.

[24] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, ‘Deceiving Google’s Perspective API Built for Detecting Toxic Comments’, *ArXiv170208138 Cs*, Feb. 2017, Accessed: Feb. 28, 2022. [Online]. Available: http://arxiv.org/abs/1702.08138

[25] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, 2018. doi: 10.12987/9780300235029.

[26] ‘Content moderation, AI, and the question of scale - Tarleton Gillespie, 2020’. https://journals.sagepub.com/doi/full/10.1177/2053951720943234 (accessed Oct. 01, 2022).

[27] T. Gillespie *et al.*, ‘Expanding the debate about content moderation: scholarly research agendas for the coming policy debates’, *Internet Policy Rev.*, vol. 9, no. 4, Oct. 2020, Accessed: Feb. 25, 2022. [Online]. Available: https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy

[28] ‘How AI is getting better at detecting hate speech’. https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech/ (accessed Feb. 28, 2022).

[29] ‘Training AI to detect hate speech in the real world’. https://ai.facebook.com/blog/training-ai-to-detect-hate-speech-in-the-real-world/ (accessed Dec. 08, 2022).

[30] R. I. Rafiq, H. Hosseinmardi, S. A. Mattson, R. Han, Q. Lv, and S. Mishra, ‘Analysis and detection of labelled cyberbullying instances in Vine, a video-based social network’, *Soc. Netw. Anal. Min.*, vol. 6, no. 1, p. 88, Sep. 2016, doi: 10.1007/s13278-016-0398-x.

[31] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, ‘Detection of Cyberbullying Incidents on the Instagram Social Network’, *ArXiv150303909 Cs*, Mar. 2015, Accessed: Feb. 28, 2022. [Online]. Available: http://arxiv.org/abs/1503.03909

Appendix

Tables 1 - 4

Label	# sentences in the original dataset		# sentences used in this empirical investigation	Examples
	By [6]	By [8]		
Insult	2,456	779	1,641	'ur not pretty at all' "get it right 'SiLlySl*t'",
Curse or Exclusion	685	323	1,045	"Just being near her makes me hate her and the fact she's emo and a cutter just makes me hate her even more.", 'Of you. Go away you waste of oxygen.',
Attack	116	54	110	‘Nearly as wet as ur mum last night lmao..’ "Ask your sister if I can't do anything else"
Threat or blackmail	92	146	180	'Not if I ruin you first;)' "You'd better understand it clearly or I'll show you how it's clear to you"
Total	**3,349**	**1,302**	**2,976**	

Table 1. Labels in datasets by [6], [8] with examples and count

	ASK.fm dataset by [6]	**WhatsApp English dataset by [7]**
Perspective API	0.003	0.0*
OPT with additional training	0.0*	0.0*
OPT without additional training	0.0*	0.0*

Table 2. P-values for the Wilcoxon sign-rank test between the toxicity probability scores and the threshold of a 70% probability of toxicity.

(* indicates the values have been rounded up, as the p-value is infinitesimally small and close to 0.0)

	ASK.fm dataset by [6]		WhatsApp English dataset by [7]	
	Sensitivity	Specificity	Sensitivity	Specificity
Perspective API	0.55	0.87	0.49	0.80
OPT with additional training	0.92	0.9	0.92	0.93
OPT without additional training	0.52	0.9	0.50	0.83

Table 3. Sensitivity and Specificity results for the industry systems on real-life cyberbullying datasets.

Label	Example Sentences without any profanity or curse words	Perspective API Score
Insult	'Hahahaha whatever ever major loser'	58.41%
	'I love all, even the Muslim Fags.'	63.08%
	"It's okay one makes mistakes, just look at your parents"	36.09%
	"Hey little loser What's up?"	63.7%
Curse or Exclusion	“No one really wants to see you, not even now”	42.98%
	“Cry, no one cares”	36.08%
	“Jump off a cliff”	53.58%
	“Shut up you stoner”	68..63%
Attack	‘….UR MUMZ BUM HAHAHA'	50.98%
	'Nearly as wet as your mum last night lmao...'	62.13%
Threat or blackmail	'I will cut you'	66.11%
	"WALLAHI MATE I'M GOING PAKISTAN HAAHAHAHAHHA|I WILL BOMB YOUR LITTLE BIN HOUSE FROM THERE MATE CALL ME ON SKYPE FOR THE LIVE PREVIEW OF YOUR BIN BLOWING UP’	68.52%

Table 4. Perspective API returns a probability score less than the threshold (70%) when profanity is not used in sentences that are insult, curse, threat or blackmail.

Figures 1 - 11

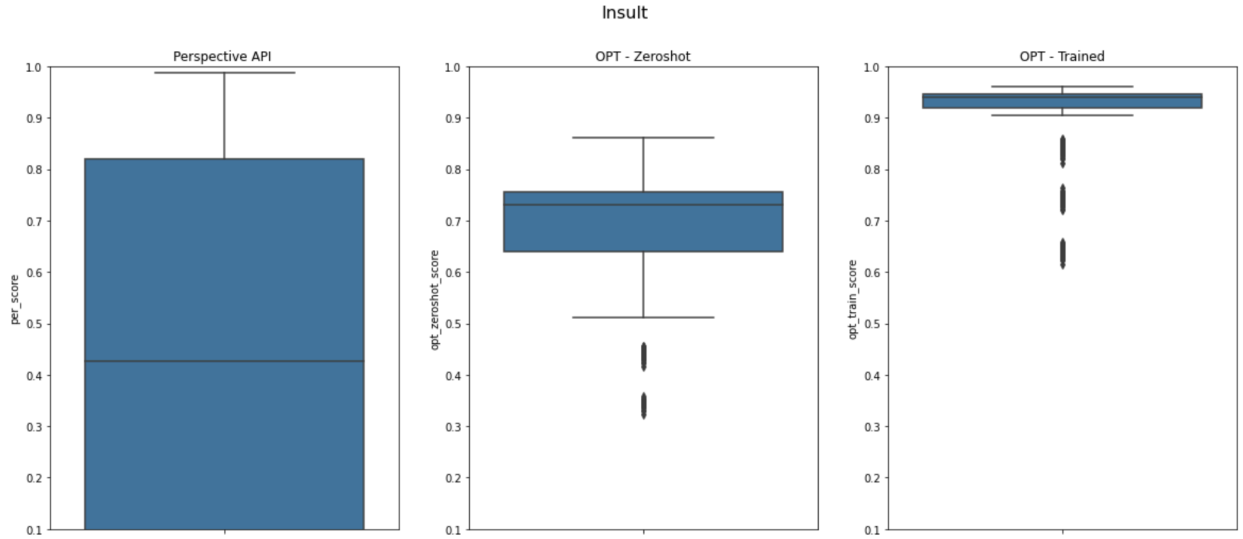

Figure 1. Distribution of probabilities of three systems Perspective API, OPT zero-shot, and OPT further trained on text messages labelled as Insult by [9] and [10]

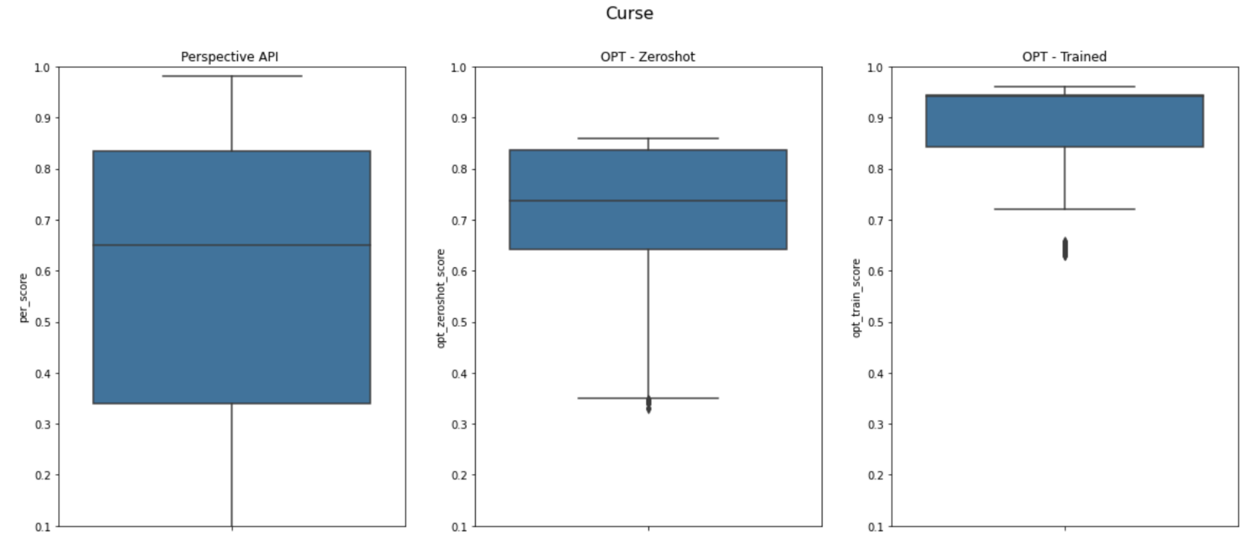

Figure 2. Distribution of probabilities of three systems Perspective API, OPT zero-shot, and OPT further trained on text messages labelled as Curse by [9] and [10]

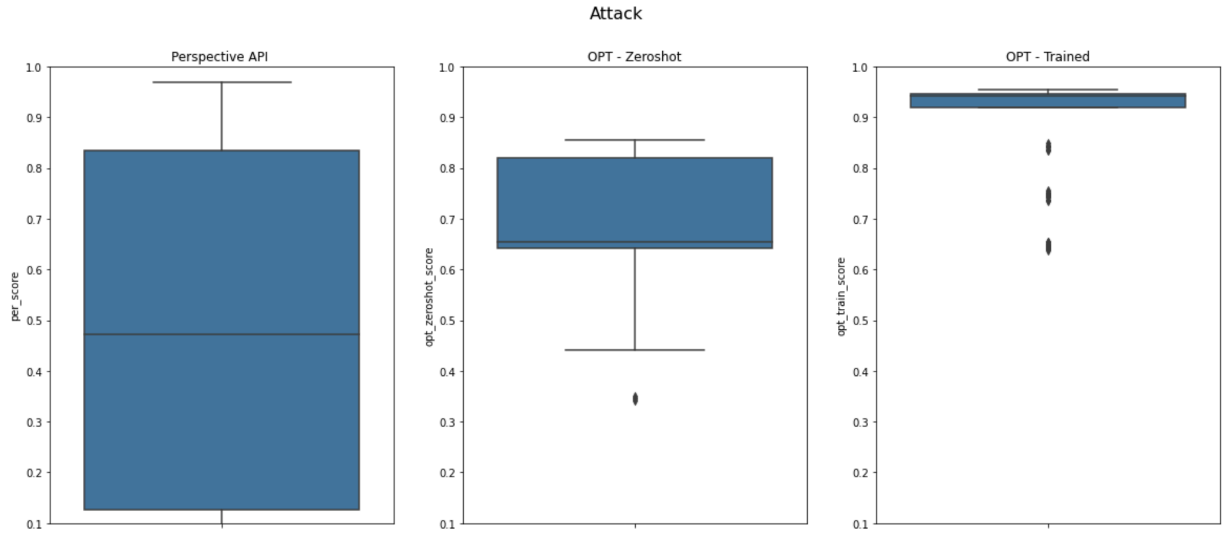

Figure 3. Distribution of probabilities of three systems Perspective API, OPT zero-shot, and OPT further trained on text messages labelled as Attack by [9] and [10]

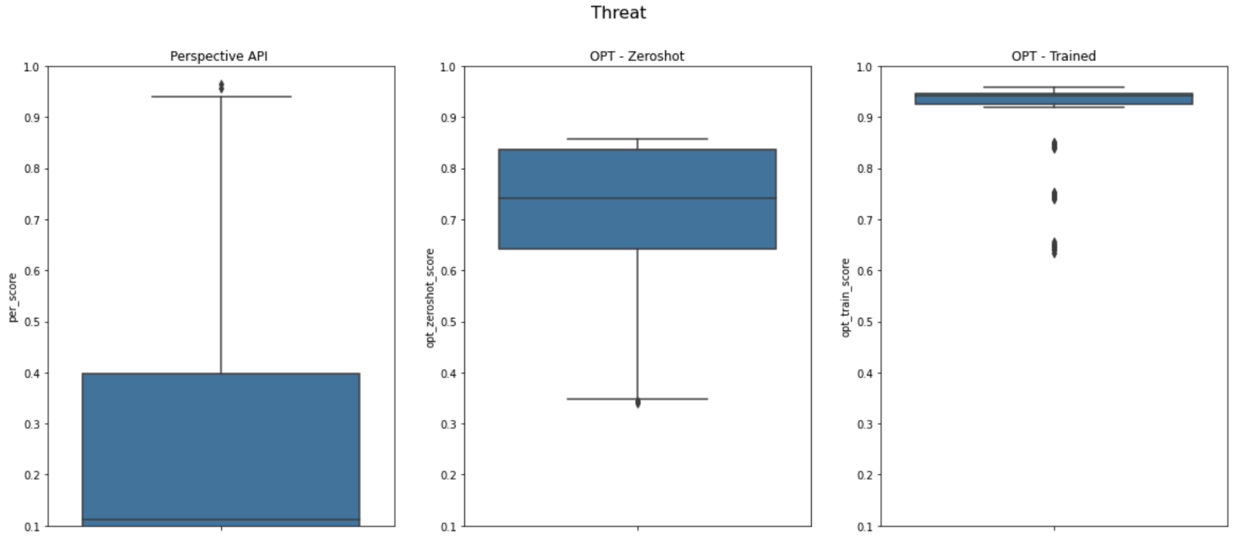

Figure 4. Distribution of probabilities of three systems Perspective API, OPT zero-shot, and OPT further trained on text messages labelled as Threat by [9] and [10]

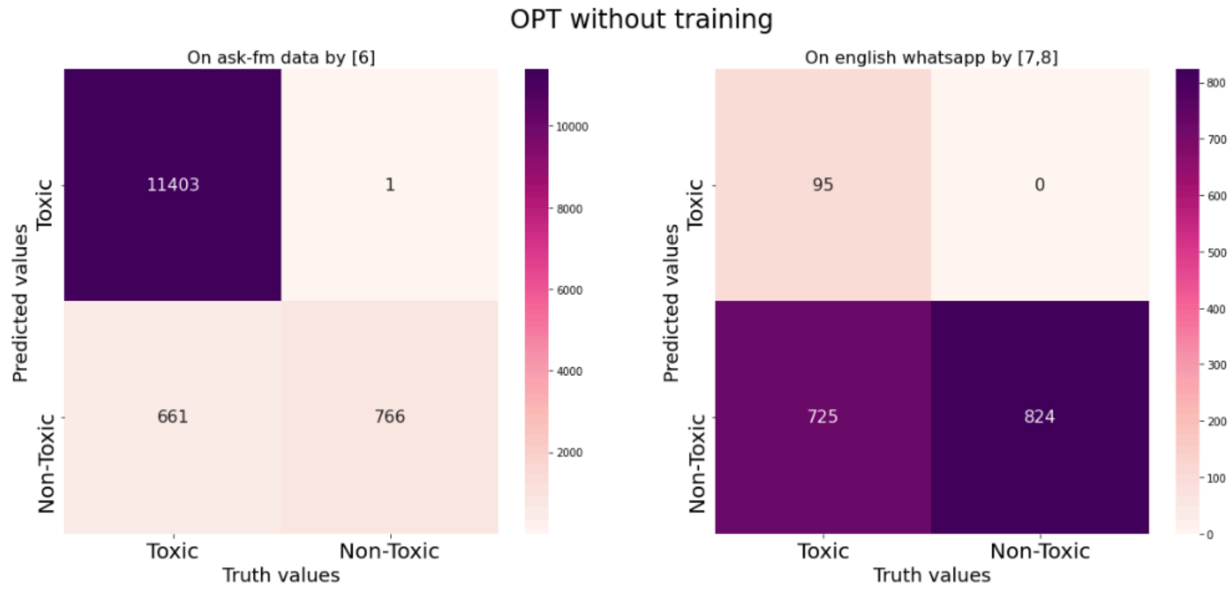

Figure 5. Confusion matrix or contingency table for OPT without training in toxicity detection on datasets by [6]–[8]

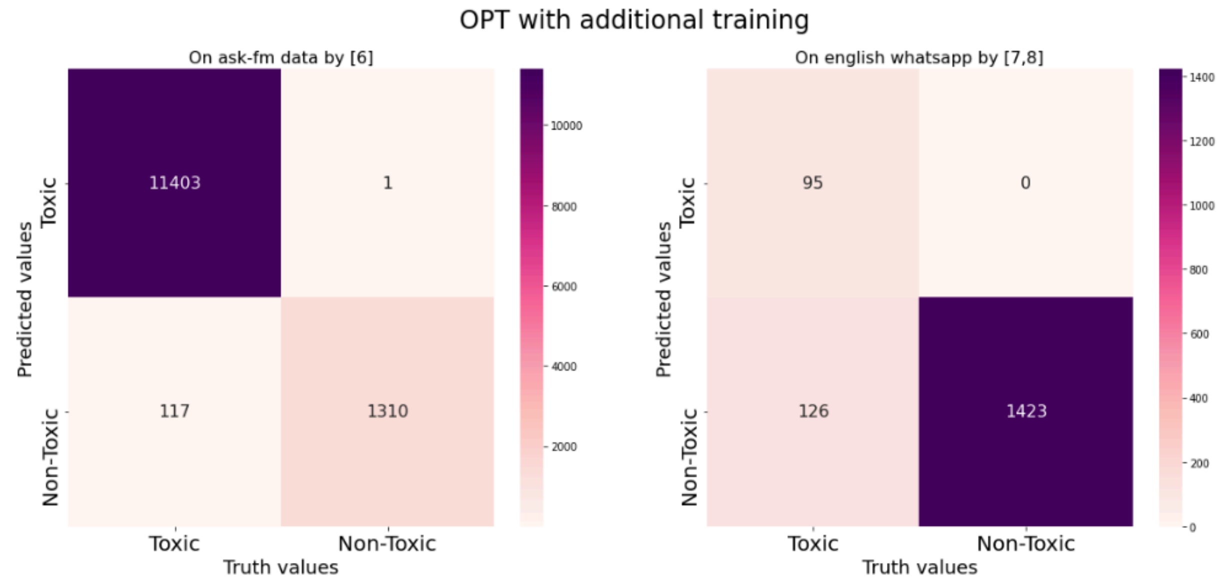

Figure 6. Confusion matrix or contingency table for OPT with additional training in toxicity detection on datasets by [6]–[8]

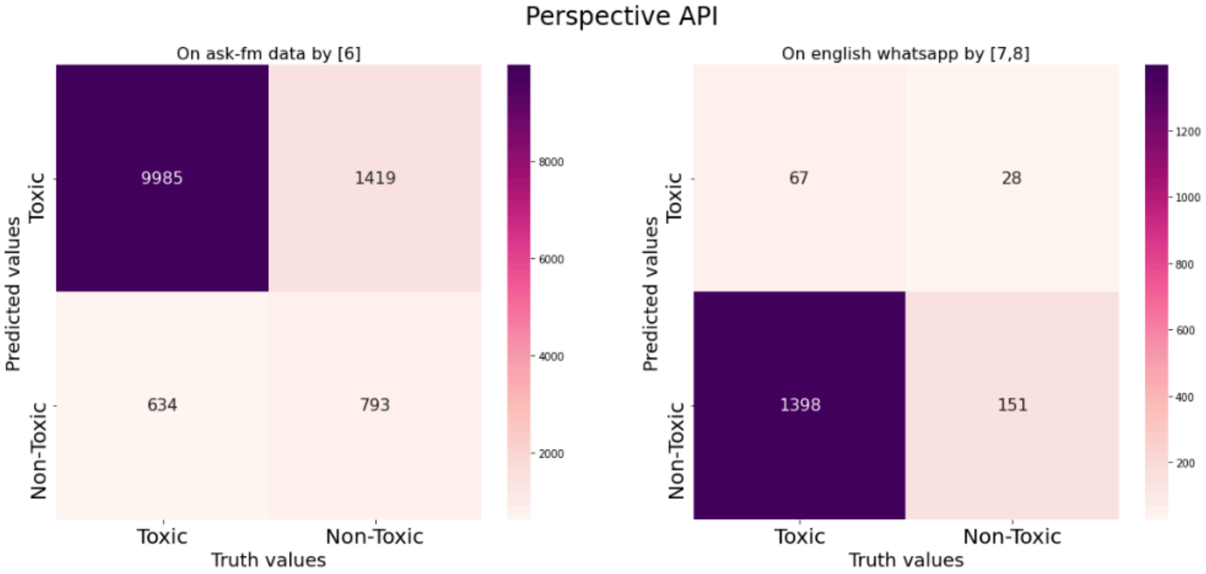

Figure 7. Confusion matrix or contingency table for Perspective API in toxicity detection on datasets by [6]–[8]

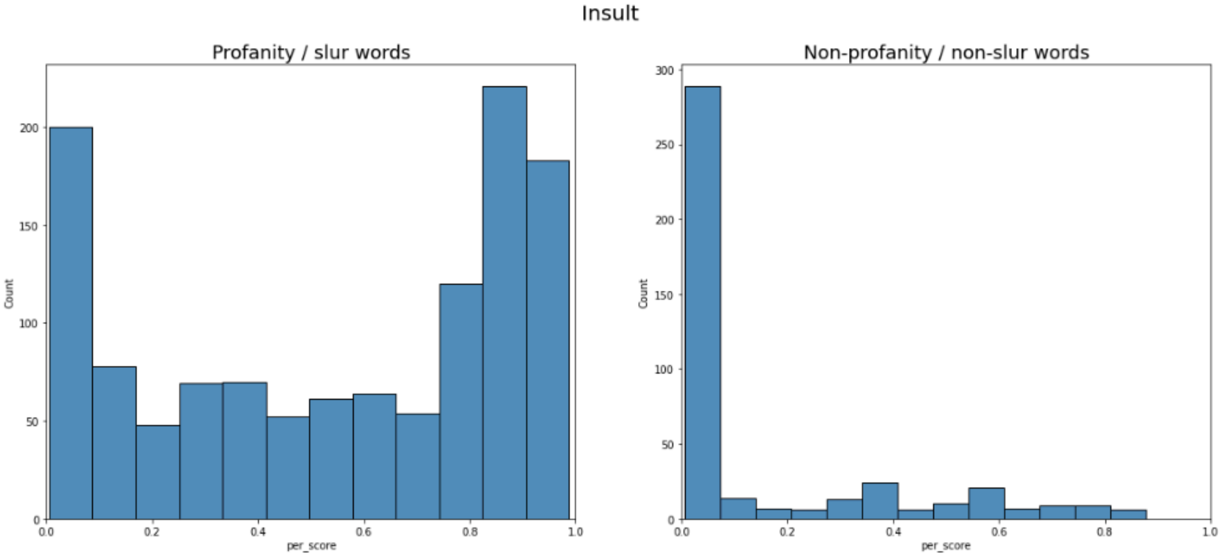

Figure 8. Perspective API toxicity scores for sentences labelled as insult with and without profanity.

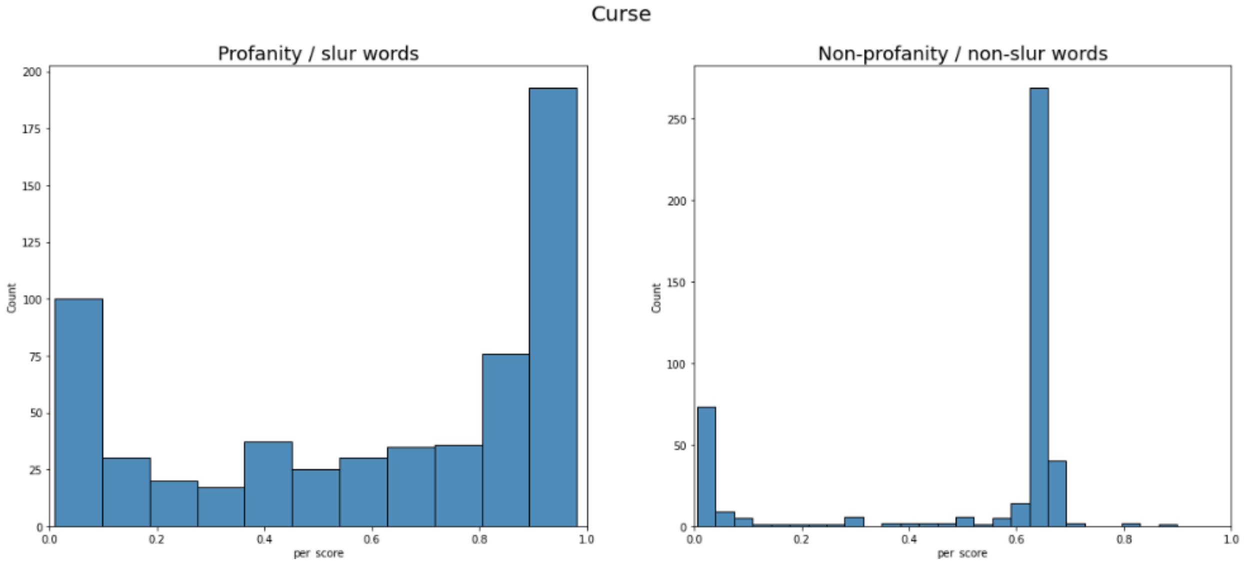

Figure 9. Perspective API toxicity scores for sentences labelled as a curse with and without profanity.

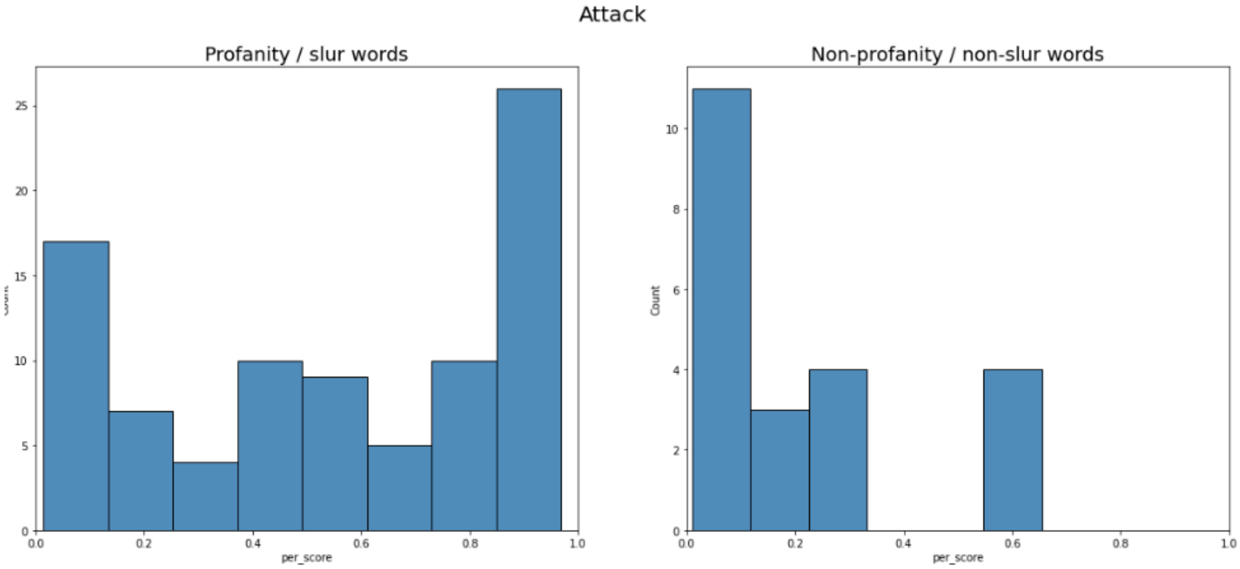

Figure 10. Perspective API toxicity scores for sentences labelled as attack with and without profanity.

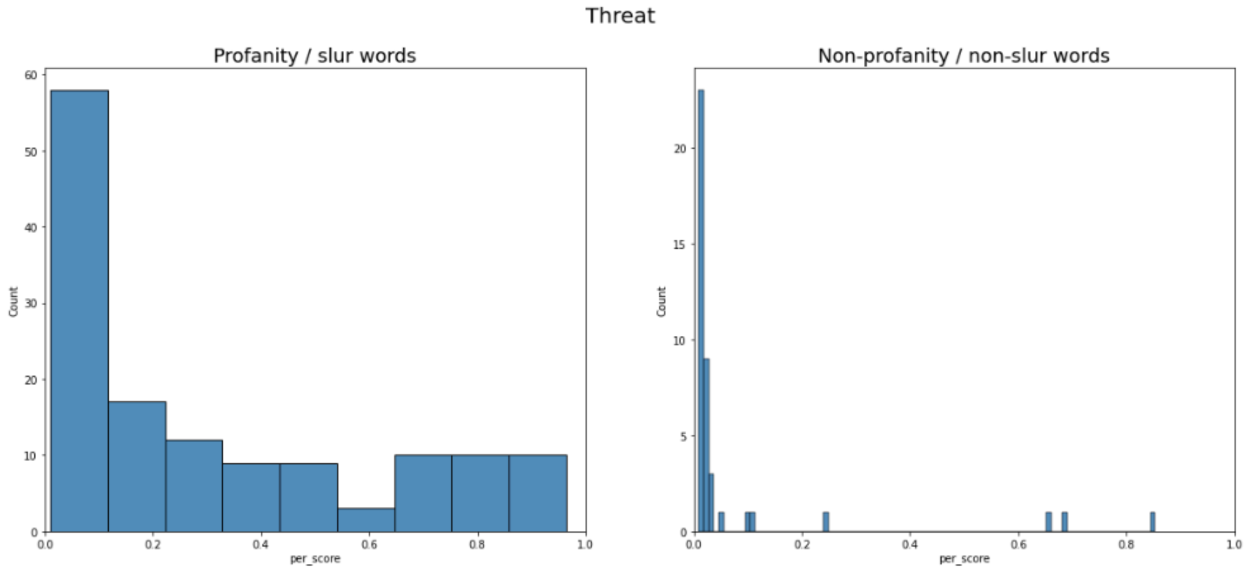

Figure 11. Perspective API toxicity scores for sentences labelled as a threat with and without profanity.