Watching the Watchers: A Comparative Fairness Audit of Cloud-based Content Moderation Services

David Hartmann[1,2], Amin Oueslati[3] and Dimitri Staufer[1]

[1]*Faculty of Electrical Engineering and Computer Science, TU Berlin*
[2]*Weizenbaum Institute for the Networked Society*
[3]*Hertie School Berlin*

Abstract

Online platforms face the challenge of moderating an ever-increasing volume of content, including harmful hate speech. In the absence of clear legal definitions and a lack of transparency regarding the role of algorithms in shaping decisions on content moderation, there is a critical need for external accountability. Our study contributes to filling this gap by systematically evaluating four leading cloud-based content moderation services through a *third-party audit*, highlighting issues such as biases against minorities and vulnerable groups that may arise through over-reliance on these services. Using a black-box audit approach and four benchmark data sets, we measure performance in explicit and implicit hate speech detection as well as counterfactual fairness through perturbation sensitivity analysis and present disparities in performance for certain target identity groups and data sets. Our analysis reveals that all services had difficulties detecting implicit hate speech, which relies on more subtle and codified messages. Moreover, our results point to the need to remove group-specific bias. It seems that biases towards some groups, such as *Women*, have been mostly rectified, while biases towards other groups, such as *LGBTQ+* and *PoC* remain.

Keywords

Content moderation as a service, hate speech detection, third-party audit, NLP fairness

1. Introduction

Hate speech has real-world effects, being the suppression of voices, exclusion, discrimination, and violence against minorities [1, 2]. It is all the more concerning that with the rise of online content in the digital age, more pernicious and unwanted content, such as hate speech and discriminatory content, is being proliferated [3]. Online platforms responded to the online hate speech proliferation by adopting extensive content moderation regimes [4] and assessing potential hateful content against so-called community guidelines by human moderators, who are assisted by algorithms [5]. Absent a translation of hate speech operationalizations into practice, private companies are given substantial autonomy in their moderation practices, effectively making them the judges of public speech [6, 7]. The largest technology firms, such as Google, Microsoft, Amazon, and OpenAI, additionally offer content moderation as a service via cloud-based API access. While most organisations do not report the extent to which algorithms

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

d.hartmann@tu-berlin.de (D. Hartmann); amin.m.oueslati@gmail.com (A. Oueslati); staufer@tu-berlin.de (D. Staufer)

CEUR Workshop Proceedings (CEUR-WS.org)

shape content moderation, the sheer amount of online speech makes reliance on algorithmic moderation inevitable [8].

The risks associated with hate speech are not limited to its lack of regulation or moderation. Over-moderation and under-moderation of specific groups and the non-functionality of automated hate speech classification can lead to serious harm. If content moderation algorithms malfunction, some users are wrongfully censored, while others are insufficiently protected [9]. Open-source content moderation algorithms have continuously displayed biases against minorities and target groups [10, 11, 12, 13, 14, 15].

Nonetheless, no systematic evaluation of cloud-based content moderation services exists, meaning an alarming absence of public scrutiny. This paper's contribution is twofold. Firstly, it offers the first comprehensive fairness assessment of four major cloud-based content moderation algorithms. Not only are these algorithms likely in use through the SaaS model. Secondly, our auditing strategy may inform future bias audits of (cloud-based) content moderation algorithms. Importantly, our proposed approach solely assumes limited black-box access [16] and offers guidance on reinforced sampling strategies to achieve maximal scrutiny with limited resources Noting the realities of unsolicited audits from civil society organisations and academia[17, 18, 19].

2. Data and Method

We gained researcher access to the Google Moderate Text API, Amazon Comprehend, Microsoft Azure Content Moderation, and the Open AI Content Moderation API. These services generate a hate speech score per text sequence, often split across several sub-categories, as well as a binary flag. Our study uses the MegaSpeech, Jigsaw, HateXplain, and ToxiGen datasets [20, 21, 22, 14]. The selected datasets capture various forms of hate speech, with ToxiGen containing implicit and adversarial hate speech constructed around indirect messages [23], while MegaSpeech and ToxiGen use generative AI to diversify speech corpora [20, 14]. Jigsaw and HateXplain contain human-written examples labeled by annotators, with MegaSpeech containing more hate speech corpora but no target group labels. MegaSpeech, HateXplain, and ToxiGen provide shorter text sequences, with on average 17.7, 23.3, and 18.1 words respectively, while Jigsaw is made up by longer sequences, 48.3 words on average.

We evaluate all cloud-based moderation algorithms across all datasets on a set of threshold-variant and threshold-invariant performance metrics [24, 25] at an aggregate level and also specifically for vulnerable groups. We ensure consistency across datasets by mapping these onto seven vulnerable groups (*Women*, *LGBTQ+*, *PoC*, *Muslim*, *Asian*, *Jewish*, *Latinx*). Since MegaSpeech comes without labels, we train a Bi-LSTM model with the collected data set by Yoder et al. [26] (preliminary evaluation accuracy 78 %) for target identity classification. At the group-level, we compute the pinned ROC AUC, a metric proposed by Dixon et al. [9], designed to provide a more robust measure for scale-invariant performance comparison across sub-groups.While this approach comes with its pitfalls, as the authors themselves note in a subsequent paper, it is the best scale-invariant metric to date when presented with group-level variation in biases [25].

Perturbation Sensitivity Analysis (PSA) offers an additional, arguably more robust evaluation of group-level biases by using counterfactual fairness evaluation[27]. We follow prior research

Dataset	Moderation Service	ROC AUC	F1	FPR	FNR	Dataset	Moderation Service	ROC AUC	F1	FPR	FNR
ToxiGen	Amazon	70.4%	68.9%	7.2%	52.0%	MegaSpeech	Amazon	72.8 %	72.0 %	10.4 %	43.9 %
	Google	62.7%	62.7%	39.1%	35.5%		Google	73.3 %	72.3 %	41.3 %	12.0 %
	OpenAI	70.3%	68.1%	33.2%	56.0%		OpenAI	77.1 %	76.7 %	8.4 %	37.3 %
	Microsoft	59.8%	57.4%	16.4%	64.0%		Microsoft	70.6 %	70.1 %	16.9 %	41.9 %
Jigsaw	Amazon	92.2%	92.2%	7.5%	8.1%	HateXplain	Amazon	66.8%	66.25%	46.3 %	20 %
	Google	69.9%	67.2%	58.4%	1.8%		Google	52.2 %	58.9 %	78.2 %	4 %
	OpenAI	78.6%	78.6%	17.1%	25.6%		OpenAI	72.9 %	76.7 %	45.4 %	8.86 %
	Microsoft	75.8%	75.7%	20.4%	28.1%		Microsoft	63.1 %	60.2 %	63.6 %	10.3 %

Table 1
Performance metrics by moderation service and dataset. Blue shading signals the best performance, while red shading indicates the worst performance. ToxiGen includes 7,800 observations and HateXplain 14,000, while Jigsaw and MegaSpeech each contain 50,000. All datasets are balanced on toxic and non-toxic phrases.

in defining an anchor group against which other groups are compared [27]. Using the dominant majority group as baseline, Counterfactual Token Fairness (CFT) scores are computed as the difference in toxicity between the baseline and the corresponding minority group.

PSA makes two assumptions: First, counterfactual pairs should convey the same or neutral meaning, avoiding any implicit biases or derogatory connotations. While constructing toxic counterfactuals is theoretically possible, it is methodologically demanding and exceeds the scope of this project. Instead, we construct 34 *neutral* counterfactual pairs. Importantly, each minority group is represented by multiple tokens, reflecting its different semantic representations. For instance, the minority group *female* also manifests as *woman* and *women.* Second, there should be no unique interactions between a particular minority token and the context of the sentence that would skew the analysis. This is challenging in real-world applications, as certain combinations might evoke stereotypes or specific cultural connotations. Thus, the project uses data consisting largely of short and explicit statements.

Furthermore, CFT scores are calculated separately for toxic and non-toxic statements, with the latter generally supporting the assumption of counterfactual symmetry more consistently. PSA experiments are conducted using two distinct data sets. First, the synthetic *Identity Phrase Templates* from Dixon et al. [9] are used. The set contains 77,000 synthetic examples of which 50% are toxic. These avoid stereotypes and complex sentence structures by design, which ensures that the symmetric counterfactual assumption is met. Mapping the dataset, which contains a broader set of identities, to the 34 minority token relevant to this study, results in 25,738 sentence pairs. Second, by applying the same logic, 9,190 sentence pairs are derived from the MegaSpeech dataset.

3. Results

Table 1 shows aggregated performance results for chosen benchmark data sets. Our results indicate notable disparities between moderation APIs. OpenAI's content moderation algorithm performs best for Megaspeech and Amazon Text Moderation on Jigsaw and ToxiGen, generalising well across data sets. On Jigsaw, Amazon Comprehend performs best. However, its near-optimal performance (92.2 % ROC AUC) suggests that the Jigsaw data was likely included in Amazon Comprehend API's training process. Overall, Google's API shows the worst

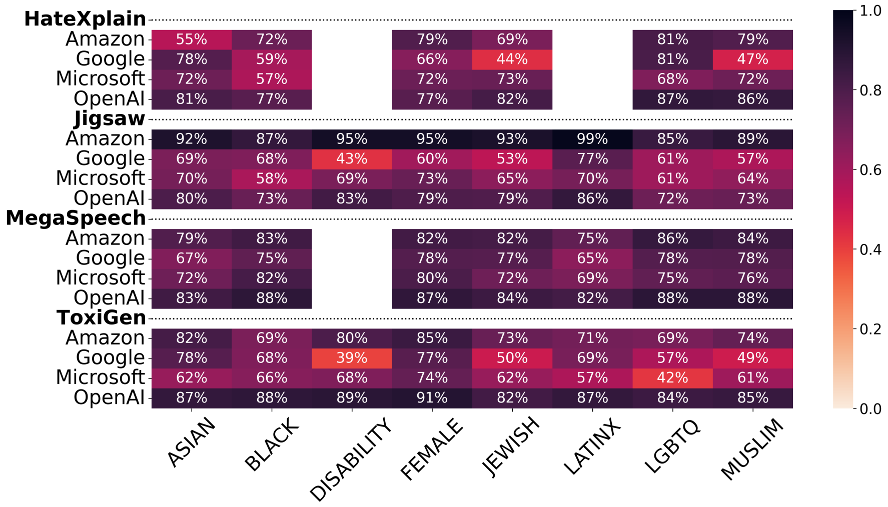

Figure 1: Pinned ROC AUC is presented by moderation service, dataset and minority group. ToxiGen includes 4,268 observations, HateXplain includes 1,748, Jigsaw consists of 19,228 observations and MegaSpeech is comprised of 33,886.

performance across data sets. Its poor performance seems driven by a comparably high FPR, which suggests that the algorithm tends to overmoderate. In contrast, Microsoft Azure Content Moderation is associated with a high FNR, suggesting it often misses hate speech.

Furthermore, all services struggle to detect implicit hate speech, reflected in their high False Positive Rates on ToxiGen. To this end, commercial moderation services do not fare much better than their open-source counterparts [14]. One likely cause is the limited availability of implicit hate speech datasets for training purposes.

The comparative fairness evaluation of the identity group is presented via group-level pinned ROC AUC scores in Figure 1. Due to space constraints, we only present one metric (ROC AUC). Future work includes a comprehensive analysis. We find that all services tend to overmoderate speech concerning groups *PoC* and *LGBTQ+*. This is somewhat surprising as extensive prior research uncovered biases in open-source content moderation algorithms in relation to these groups [28]. Commonly, such overmoderation occurs as toxic speech concerning these groups is overrepresented in the training data, and subsequently learned by the model. Most services fail to reliably detect hate speech aimed at groups *Disability*, *Asian*, and *Latinx*. Lastly, the tendency of Google Text Moderation to overmoderate is puzzling but also alarming. While we cannot entirely rule out an error on our end, this observation is robust to different configurations of API sub-categories. Figure 1 (right) displays the PSA results. We find (1) differences in toxicity scores by and large are more pronounced on non-toxic than toxic data. Intuitively this makes sense, as scores are generated non-linearly with a definite upper bound. Thus, when other elements in a sentence induce a high toxicity score, the marginal effect from identity tokens is comparably lower. We further find that (2) greater variation in the mean CFT scores in non-synthetic than in synthetic data. This was to be expected, as the sentences from MegaSpeech contain more contextual information that interacts with the tokens. Overall, the results suggest that most

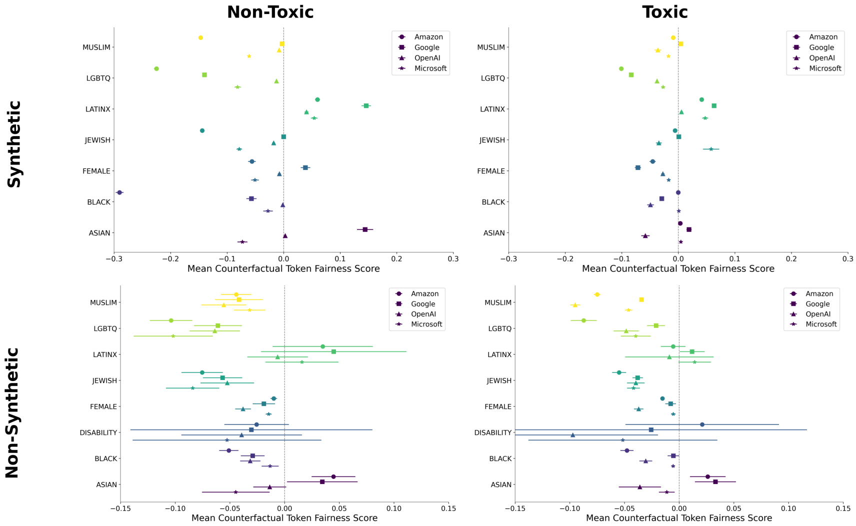

Figure 2: CFT scores are visualized. They are computed through PSA on synthetic data from the Identity Phrase Templates in Dixon et al. [9] and non-synthetic data from MegaSpeech, averaged per group and service, reported separately for non-toxic and toxic examples. Besides a point estimate, the figure also includes a 95% confidence interval assuming a student-t distribution.

minorities are associated with higher levels of toxicity than dominant majorities, although these effects appear relatively small, and vary across groups and services. Group *LGBTQ+* seems associated with the strongest negative bias, occurring for all samples and services. We observe limited negative bias against groups *Latinx* and *Asian*.

4. Conclusion

Summarizing, we uncovered both aggregate-level performance issues and group-level biases in major commercial cloud-based content moderation services. Importantly, while some shortcomings extend to all services, such as difficulty in detecting implicit hate speech or biases against group *LGBTQ+*, others are confined to a particular service.

Over the years, a lot of research has been done that shows the biases and limitations of automated hate speech detection classifiers. Nevertheless, these limitations persist in current content moderation APIs. We demonstrated that all five tested content moderation APIs show disparities in performance for specific target groups, for implicit hate speech, over moderate target groups which are strongly associated with hate speech online and penalize counter speech as well as reappropriation.

Challenges we encountered, such as the inherent subjectivity of hate speech moderation and data limitations, should not deter but encourage future work. Without public scrutiny, the subjectivity does not vanish, but it remains entirely to the discretion of private companies to make these subjective choices.

References

[1] M. J. Matsuda, C. R. L. III, R. Delgado, K. W. Crenshaw, Words That Wound: Critical Race Theory, Assaultive Speech, and The First Amendment, Faculty Books, 1993. URL: https://scholarship.law.columbia.edu/books/287, accessed: date-of-access.

[2] T. Marques, The expression of hate in hate speech, Journal of Applied Philosophy 40 (2023) 769–787. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/japp.12608. doi:https://doi.org/10.1111/japp.12608. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/japp.12608.

[3] C. Bakalis, Regulating hate crime in the digital age, Oxford University Press, 2016.

[4] G. De Gregorio, Democratising online content moderation: A constitutional framework, Computer Law & Security Review 36 (2020) 105376.

[5] R. Gorwa, R. Binns, C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, Big Data & Society 7 (2020) 205395171989794. URL: http://journals.sagepub.com/doi/10.1177/2053951719897945.

[6] J. Seering, Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation, Proceedings of the ACM on Human-Computer Interaction 4 (2020) 1–23.

[7] S. A. Einwiller, S. Kim, How online content providers moderate user-generated content to prevent harmful online communication: An analysis of policies and their implementation, Policy & Internet 12 (2020) 184–206. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.239. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.239.

[8] C. Schluger, J. P. Chang, C. Danescu-Niculescu-Mizil, K. E. C. Levy, Proactive moderation of online discussions: Existing practices and the potential for algorithmic support, Proceedings of the ACM on Human-Computer Interaction 6 (2022) 1 – 27. URL: https://api.semanticscholar.org/CorpusID:253460203.

[9] L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and Mitigating Unintended Bias in Text Classification, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 67–73.

[10] T. Garg, S. Masud, T. Suresh, T. Chakraborty, Handling Bias in Toxic Speech Detection: A Survey, CoRR abs/2202.00126 (2022). URL: https://arxiv.org/abs/2202.00126, arXiv: 2202.00126.

[11] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social Bias Frames: Reasoning about Social and Power Implications of Language, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5477–5490. URL: https://aclanthology.org/2020.acl-main.486.

[12] P. Fortuna, J. Soler, L. Wanner, Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille,

France, 2020, pp. 6786–6794. URL: https://aclanthology.org/2020.lrec-1.838.

[13] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of Abusive Language: the Problem of Biased Datasets, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:174799974.

[14] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, E. Kamar, ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection, in: Annual Meeting of the Association for Computational Linguistics, 2022. URL: https://api.semanticscholar.org/CorpusID:247519233.

[15] E. Sheng, K.-W. Chang, P. Natarajan, N. Peng, The Woman Worked as a Babysitter: On Biases in Language Generation, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3407–3412. URL: https://aclanthology.org/D19-1339. doi:10.18653/v1/D19-1339.

[16] S. Casper, C. Ezell, C. Siegmann, N. Kolt, T. L. Curtis, B. Bucknall, A. Haupt, K. Wei, J. Scheurer, M. Hobbhahn, L. Sharkey, S. Krishna, M. V. Hagen, S. Alberti, A. Chan, Q. Sun, M. Gerovitch, D. Bau, M. Tegmark, D. Krueger, D. Hadfield-Menell, Black-box access is insufficient for rigorous ai audits, 2024. arXiv:2401.14446.

[17] A. Birhane, R. Steed, V. Ojewale, B. Vecchione, I. D. Raji, Ai auditing: The broken bus on the road to ai accountability, ArXiv abs/2401.14462 (2024). URL: https://api.semanticscholar.org/CorpusID:267301287.

[18] A. Kak, S. M. West, Algorithmic Accountability: Moving Beyond Audits, AI Now Institute (2023). URL: https://ainowinstitute.org/publication/algorithmic-accountability.

[19] I. D. Raji, P. Xu, C. Honigsberg, D. E. Ho, Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance, 2022. URL: http://arxiv.org/abs/2206.04737, arXiv:2206.04737 [cs].

[20] S. Pendzel, T. Wullach, A. Adler, E. Minkov, Generative AI for Hate Speech Detection: Evaluation and Findings, 2023. URL: http://arxiv.org/abs/2311.09993, arXiv:2311.09993 [cs].

[21] Jigsaw, Jigsaw toxic comment classifi- cation challenge., 2019. URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[22] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 14867–14875. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17745, number: 17.

[23] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. D. Choudhury, D. Yang, Latent Hatred: A Benchmark for Understanding Implicit Hate Speech, CoRR abs/2109.05322 (2021). URL: https://arxiv.org/abs/2109.05322, arXiv: 2109.05322.

[24] F. Elsafoury, S. Katsigiannis, N. Ramzan, On Bias and Fairness in NLP: How to have a fairer text classification?, 2023. URL: http://arxiv.org/abs/2305.12829, arXiv:2305.12829 [cs].

[25] D. Borkan, L. Dixon, J. Sorensen, N. Thain, L. Vasserman, Nuanced metrics for measuring unintended bias with real data for text classification, in: Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 491–500. URL: https://doi.org/10.1145/3308560.3317593.

[26] M. M. Yoder, L. H. X. Ng, D. W. Brown, K. M. Carley, How hate speech varies by target

identity: A computational analysis, arXiv preprint arXiv:2210.10839 (2022).

[27] V. Prabhakaran, B. Hutchinson, M. Mitchell, Perturbation sensitivity analysis to detect unintended model biases, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5740–5745. URL: https://aclanthology.org/D19-1578.

[28] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, A. Beutel, Counterfactual Fairness in Text Classification through Robustness, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, ACM, Honolulu HI USA, 2019, pp. 219–226. URL: https://dl.acm.org/doi/10.1145/3306618.3317950.