9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Data Augmentation for Improving Automated Harmful Language Detection: Potentials and Pitfalls

Keywords: sexism, hate speech, content moderation, data augmentation, NLP

Extended Abstract

As fully or semi-automated models are increasingly used for platform governance [3, 1] there are several questions about their performance and the implications of model errors [2, 5]. Language technologies underpinning these content moderation strategies, especially models for detecting problematic content like hate speech and sexism, need to be designed to ensure several complex desiderata, including robustness across domains of application as well as low misclassification rates. Indeed, misclassifications can have a range of repercussions from allowing problematic content to proliferate to sanctioning users who did nothing wrong, often minorities and activists [4].

To facilitate model robustness, several solutions encompass improving training data for these models, such as by training them on counterfactually augmented data (CAD). CAD, also called contrast sets, are obtained by making *minimal* changes to existing datapoints to flip their label for a particular NLP task [6]. Previous research has established that training on CAD increases out-of-domain generalizability [6, 7]. [7] explores characteristics of effective counterfactuals, finding that models trained on *construct-driven* CAD, or CAD obtained by directly perturbing manifestations of the construct, e.g., gendered words for sexism, lead to higher out-of-domain generalizability. Previous research also notes that gains from training on CAD can be attributed to learning more core features, rather than dataset artifacts [6]. However, it is unclear how learning such core features can affect model misclassifications, especially for cases where the effect of the core feature is modulated by context—e.g., how models trained on CAD classify non-sexist examples containing gendered words. Investigating this type of misclassification can help uncover *unintended false positive bias*. Unintended false positive bias can lead to wrongful moderation of those not engaging in hate speech, or even worse, those reporting or protesting it. Such type of bias is especially concerning in the use of social computing models for platform governance. Recent work has shown that AI-driven abusive language or toxicity detection models disproportionately flag and penalize content that contains markers of identity terms even though they are not toxic or abusive [4]. Over-moderation of this type, driven by unintended false positive bias, can hurt marginalized communities even more.

This work. We assess the interplay between CAD as training data and unintended bias in sexism and hate speech models. Grounding the measure of unintended bias as the prevalence of falsely attributing hate speech or sexism to posts which use identity words without being hateful, we assess if training on CAD leads to higher false positive rates (FPR). In line with past research, models trained on CAD show higher accuracy on out-of-domain data (higher model robustness) *but* also have higher FPR on non-hateful usage of identity terms.

Datasets and Methods. We use the same experimental setup and notation as [7], but instead only focus on sexism and hate speech as these are the NLP tasks widely used in text-based content moderation. For testing our models, we distill a subset of tweets that contain gender and identity terms, for sexism and hate speech respectively (based on lexica for gendered words and identity terms, calling this dataset the identity subgroup (ISG). We use two different families

const	data	model	mode	macro F1	FPR	FNR
hate speech	ISG	bert	CF	**0.65**	0.43	0.24
			nCF	0.60	0.36	0.44
		logreg	CF	**0.58**	0.31	0.53
			nCF	0.40	0.12	0.93
sexism	ISG	bert	CF	**0.65**	0.37	0.33
			nCF	0.57	0.16	0.64
		logreg	CF	**0.56**	0.29	0.56
			nCF	0.51	0.19	0.71

Table 1: **Macro F1 and FPR on the Identity Subgroup (ISG) for models trained on CAD (CF) vs those trained on original data (nCF).** While CF models improve in terms of F1, they tend to have a higher False Positive Rate than their nCF counterparts. This is especially pronounced for the BERT models.

of models: logistic regression (LR) with a TF-IDF bag-of-words representation, and finetuned-BERT. We train two types of binary text classification models of each model family on the in-domain data only—nCF models trained on original data, and CF models trained on both original data and CAD. The nCF models are trained on 100% original data.

Results. We use *false positive rate (FPR)* to measure unintended bias as our concept of unintended bias in harm detection systems entails misclassifications of non-harmful content containing identity-related terminology. We contrast measures of FPR with Macro F1 (as an overall performance metric) and false negative rate (FNR). We assess the overall performance of models trained on CAD as well as models trained on specific types of CAD. Table 1 shows the result of CF and nCF models on ISG. **Our results indicate that CF models do, indeed, have higher FPR compared to their nCF counterparts, for both sexism and hate speech, while having a lower FNR.** On all examples, CF models have higher F1, but higher FPR. Our analysis and results indicate that while training on CAD can lead to gains in model robustness by promoting core features, not taking into account the context surrounding these core features can lead to false positives, possibly due to the confounding relationship between identity terms and hate speech.

References

[1] T. Gillespie. *Custodians of the Internet*. Yale University Press, 2018.

[2] T. Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 2020.

[3] R. Gorwa. The platform governance triangle: Conceptualising the informal regulation of online content. *Internet Policy Review*, 8(2):1–22, 2019.

[4] K. L. Gray and K. Stein. We 'said her name'and got zucked": Black Women Calling-out the Carceral Logics of Digital Platforms. *Gender & Society*, 35(4):538–545, 2021.

[5] S. T. Roberts. *Behind the screen*. Yale University Press, 2019.

[6] M. Samory, I. Sen, J. Kohne, F. Flöck, and C. Wagner. Call me sexist, but...: Revisiting sexism detection using psychological scales and adversarial samples. 2021.

[7] I. Sen, M. Samory, F. Flöck, C. Wagner, and I. Augenstein. How does counterfactually augmented data impact models for social computing constructs? In *EMNLP*, 2021.