Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection

Indira Sen[1]   **Mattia Samory**[1]   **Claudia Wagner**[1,2]   **Isabelle Augenstein**[3]

[1]GESIS – Leibniz Institute for the Social Sciences   [2]RWTH Aachen University

[3]University of Copenhagen

{indira.sen, mattia.samory, claudia.wagner}gesis.org; augenstein@di.ku.dk

Abstract

Counterfactually Augmented Data (CAD) aims to improve out-of-domain generalizability, an indicator of model robustness. The improvement is credited to promoting core features of the construct over spurious artifacts that happen to correlate with it. Yet, over-relying on core features may lead to unintended model bias. Especially, *construct-driven* CAD—perturbations of core features—may induce models to ignore the context in which core features are used. Here, we test models for sexism and hate speech detection on challenging data: non-hateful and non-sexist usage of identity and gendered terms. On these hard cases, models trained on CAD, especially construct-driven CAD, show higher false positive rates than models trained on the original, unperturbed data. Using a diverse set of CAD—construct-driven and construct-agnostic—reduces such unintended bias.

1 Introduction

As fully or semi-automated models are increasingly used for platform governance (Gorwa, 2019; Gillespie, 2018; Nakov et al., 2021) there are several questions about their performance and the implications of model errors (Gorwa et al., 2020; Gillespie, 2020; Roberts, 2019). 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 (Gray and Stein, 2021; Haimson et al., 2021). Such misclassifications are a threat to model robustness and non-robust models can cause a great deal of collateral damage.

To facilitate model robustness, several solutions encompass improving training data for these models (Dinan et al., 2019; Vidgen et al., 2021), such as by training them on counterfactually augmented data (CAD). CAD, also called contrast sets (Gardner et al., 2020; Atanasova et al., 2022), are obtained by making *minimal* changes to existing datapoints to flip their label for a particular NLP task (Kaushik et al., 2019; Samory et al., 2021). Previous research has established that training on CAD increases out-of-domain generalizability (Kaushik et al., 2019; Samory et al., 2021; Sen et al., 2021). Sen et al. (2021) 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 (Kaushik et al., 2019; Samory et al., 2021). 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 (Gray and Stein, 2021; Haimson et al., 2021). Over-moderation of this type, facilitated by unintended false positive bias, can end up hurting

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, Logistic Regression and BERT 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. This effect is most prominent in models trained on *construct-driven* CAD. Our results uncover potential negative consequence of using CAD, and its different types, for augmenting training data. We release our code to facilitate future research here: `https://github.com/gesiscss/Uninteded_Bias_in_CAD`.

2 Background

For a given text with an associated label, say a sexist tweet, a counterfactual example is obtained by *making minimal changes to the text to flip its label*, i.e., into a non-sexist tweet. Counterfactual examples in text have the interesting property that, since they were generated with minimal changes, they allow one to focus on the manifestation of the construct; in our example, that would be what makes a text sexist. Previous research has exploited this property to nudge NLP models to look at the points of departure and thereby learn core features of the construct rather than dataset artifacts (Kaushik et al., 2019; Samory et al., 2021; Sen et al., 2021).

Types of CAD. Sen et al. (2021) used a causal inference-inspired typology to categorize different types of CAD and found that models trained on certain types of CAD are more robust. We follow the same typology and distinguish between— **Construct-driven CAD** obtained by making changes to an existing item by acting on the construct, e.g, on the gendered terms for sexism, and **Construct-agnostic CAD** obtained by making changes to general characteristics of an item, such as inserting negation. Following Sen et al. (2021), we use lexica to automatically characterize CAD.

3 Datasets and Methods

We use the same experimental setup and notation as Sen et al. (2021), but instead only focus on sexism and hate speech as these are the NLP tasks

widely used in text-based content moderation. Table 1 summarizes the datasets we use, training on an in-domain dataset and using two datasets for testing— Identity Subgroup (**ISG**) which is a subset of the out-of-domain dataset used by Sen et al. (2021) and Hatecheck (**HC**) (Röttger et al., 2021). The test sets are described in more detail in Section 4.1. All the in-domain datasets come with CAD, gathered by crowdworkers (Samory et al., 2021) or expert annotators (Vidgen et al., 2021) in previous research.

Since previous work has shown that models trained on CAD tend to perform well on counterfactual examples (Kaushik et al., 2019; Samory et al., 2021), we do not include CAD in any of the test sets. All datasets contain only English examples.

We use two different families of models: logistic regression (LR) with a TF-IDF bag-of-words representation, and finetuned-BERT (Devlin et al., 2019). 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, namely, the "Original" in the "Train" column in Table 1. The CF models for hate speech are trained on ~ 50% original data and ~ 50% CAD, sampled from the "Train" and "Counterfactual" columns in Table 1, respectively. Since only non-sexist CAD are provided for sexism classification, the sexism models are trained on 50% original sexist data, 25% original non-sexist data, and 25% counterfactual non-sexist data (Samory et al., 2021).

Based on Sen et al. (2021), to unpack the effect of different types of CAD on model performance, we further disaggregate the CAD training sets, and train models on different types of CAD: only construct-driven counterfactuals (CF_const), only construct-agnostic counterfactuals (CF_agn), and equal proportions of both (CF_mix).[1] Due to the lack of data and unequal distributions of different types of CAD, instead of training on 50% CAD, we train on 20% for these three types of models. Training details including model hyperparameters are described in the Appendix (Section 9).

---

[1] We use a slightly different terminology for the models trained on different types of CAD than Sen et al. to ease interpretability. Namely, CF_c is CF_const, CF_r is CF_agn, and CF_a is CF_mix in this work.

Construct	Train				Test			
Sexism	Original (Samory et al.)		Counterfactual (Samory et al.)		ISG (EXIST)		HC (women)	
	sexist	**non-sexist**	**sexist**	**non-sexist**	**sexist**	**non-sexist**	**sexist**	**non-sexist**
	1244	1610	-	912	1178	1046	373	136
Hate speech	Original (Vidgen et al.)		Counterfactual (Vidgen et al.)		ISG (Basile et al.)		HC	
	hate	**not-hate**	**hate**	**not-hate**	**hate**	**not-hate**	**hate**	**not-hate**
	6524	5767	5096	5852	625	865	2563	1165

Table 1: NLP tasks/constructs and datasets used in this work. Following a similar set up as Sen et al. (2021), we train models on the in-domain datasets, while out-of-domain datasets are used for testing. EXIST refers to the dataset from the shared task on sexism detection (Rodriguez-Sanchez et al., 2021)
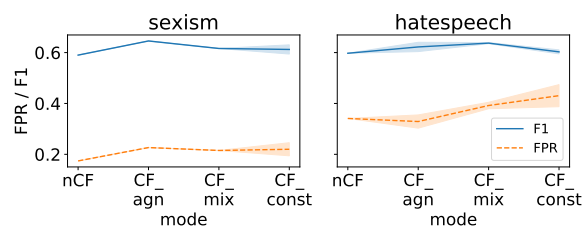
const	data	model	mode	macro F1	FPR	FNR
hate speech	ISG	bert	CF	**0.65**	<u>0.43</u>	0.24
			nCF	0.60	0.36	0.44
		logreg	CF	**0.58**	<u>0.31</u>	0.53
			nCF	0.40	0.12	0.93
sexism	ISG	bert	CF	**0.65**	<u>0.37</u>	0.33
			nCF	0.57	0.16	0.64
		logreg	CF	**0.56**	<u>0.29</u>	0.56
			nCF	0.51	0.19	0.71

Table 2: Macro F1 and FPR on the Identity Subgroup (ISG) for mod- els trained on CAD (CF) vs those trained on orig- inal data (nCF). While CF models improve in terms of F1, they tend to have a higher <u>False Positive Rate</u> than their nCF counterparts. This is especially pronounced for the BERT models.

4 Unintended Bias

Previous research has shown that training on CAD can improve model robustness, i.e, generalization to data beyond the training domain (Kaushik et al., 2019; Samory et al., 2021; Sen et al., 2021). Here, we take a closer look at one aspect of model robustness, i.e false positives, and conduct a focused error analysis inspired by real-world applications of social NLP systems – particularly the case of misclassification of content with identity terms, an example of unintended bias. Previous research has shown that CF models tend to promote core features, namely, gender words for sexism and identity terms for hate speech (Sen et al., 2021). One potential consequence of this promotion of identity features for detecting problematic content could be an increase in false positives, particularly in innocuous posts that contain identity terms.[2] This can be especially harmful if the misclassified posts happen to be reports or disclosures of facing hate

Figure 1: FPR and F1 for different types of CF BERT models, with the nCF model as a baseline on ISG. For hate speech, CF_const has the highest FPR.

speech.[3] Our work builds on recent literature that investigates the performance and misclassification rate of toxicity or hate speech detection models on test sets containing identity terms (Dixon et al., 2018; Borkan et al., 2019; Kennedy et al., 2020; Nozza et al., 2019; Calabrese et al., 2021). However, unlike previous work, we specifically focus on the behaviour of models trained on CAD on test cases with identity terms. We do so since previous work established that models trained on CAD tend to learn "core" features (Kaushik et al., 2019; Samory et al., 2021; Sen et al., 2021) which are often identity terms for sexism and hate speech detection — therefore, it is important to uncover how this increased focus on core features modulates misclassifications of instances where these terms are used in a non-hateful context.

4.1 Test Sets for Measuring Unintended Bias

To understand if CF models facilitate this type of unintended bias, we leverage two tests sets. First, we include a subset of the out-of-domain datasets used in Sen et al. (2021) which contains both sexist (hateful) and non-sexist (non-hateful) posts with gendered (identity) words, called the Identity Sub-

---

[2]Identity terms subsume gender words

[3]In a real world example, tweets by activists are often flagged and deleted by commercial content moderation systems (Gray and Stein, 2021).

const	data	model	mode	macro F1	FPR	FNR
hate speech	HC	bert	CF	**0.88**	0.08	0.13
			nCF	0.67	<u>0.24</u>	0.35
		logreg	CF	**0.61**	<u>0.50</u>	0.28
			nCF	0.52	0.45	0.47
sexism	HC	bert	CF	**0.53**	<u>0.49</u>	0.41
			nCF	0.49	0.32	0.57
		logreg	CF	**0.53**	<u>0.60</u>	0.32
			nCF	0.44	0.32	0.65

Table 3: **Macro F1 and FPR on Hatecheck for models trained on CAD (CF) vs those trained on original data (nCF).** All CF models have both higher F1 and higher FPR, with the exception of the hate speech BERT CF model which has a lower FPR.

group. Second, we include the HateCheck test suite (Röttger et al., 2021), designed to test hate speech models.

Identity Subgroup in Out-of-Domain Data (ISG). While Sen et al. (2021) used the entire out-of-domain dataset for testing, we distill a subset of tweets that contain gender and identity terms, for sexism and hate speech respectively (based on lexica for gendered words[4] and identity terms[5]), calling this dataset the identity subgroup. ISG is an effective dataset for investigating unintended bias since it specifically contains tweets with mentions of identities in both hateful and non-hateful contexts.

HateCheck (HC). HateCheck (Röttger et al., 2021) is a functional test suite for hate speech detection models that is inspired by software testing (Ribeiro et al., 2020). HateCheck provides 3728 test cases covering 29 different functionalities such as "Non-hateful homonyms of slurs", making it a challenging test set for models where many of the instances include identity terms used in a non-hateful context. To evaluate the hate speech models, we use all instances, while for the sexism models, we use a subset that targets women.[6] To find the subset of instances targeting women, we reuse the gender word lexica to exclude instances that do not have gendered words.

---

[4]obtained from: `https://github.com/uclanlp/gn_glove/tree/master/wordlist`

[5]We use a list of identity words and slurs curated by (Khani and Liang, 2021) and from Hatebase (`https://hatebase.org/`)

[6]Note that HateCheck does not cover sexism against men but does have a category for gay people. In the future, we hope to see how sexism models compare against hate speech models in detecting sexism for intersectional identities.
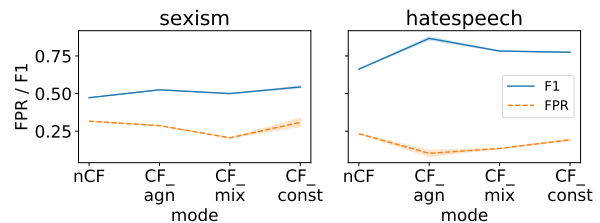
Figure 2: **F1 and FPR for different types of CF BERT models for HC, with the nCF model as a baseline.** Models trained on construct-driven CAD (CF_const) incur the highest FPR for both sexism and hate speech.

4.2 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 fale negative rate (FNR). We assess the overall performance of models trained on CAD as well as models trained on specific types of CAD.

Performance of CAD vs. Original Data. Table 2 shows the result of CF and nCF models on ISG while Table 3 shows the results for HateCheck. **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 (e.g., 88% vs. nCF model's 67% for hate speech in Table 3), but higher FPR. The CF BERT model for hate speech is the exception, having a lower FPR compared to its nCF counterpart. To unpack this discrepancy, we qualitatively assessed cases where nCF models had misclassifications but CF models did not. These included cases of positive sentiment towards minorities, e.g. "I absolutely adore Muslims". As Sen et al. (2021) note, the majority of CAD generated for hate speech changes affect words (55%), possibly explaining how the CF BERT model became proficient at correctly recognizing such instances. This adds to the evidence that a mixture of different types of CAD is ideal for aiding models in understanding the concept of hate speech in a holistic manner.

Performance of CAD Types. We repeat this analysis with models trained on different types of CAD, namely, construct-driven CAD (CF_const), construct-agnostic (CF_agn), and equal proportions of both (CF_mix) for ISG (Figure 1) and HC

(Figure 2). Figures 1 and 2 show the results for the BERT models while the results for the logistic regression models are included in the Appendix (Section 7). Overall, complementing the comparison between CF and nCF models, we find that models trained on either type of CAD have higher F1 than nCF models. However, the ranking between models trained on different types of CAD is not clear, especially for ISG. In ISG, For hate speech, we see that construct-driven (CF_const) models demonstrate high FPR. For HC, the high FPR of models trained on construct-driven CAD is more pronounced — for both sexism and hate speech, CF_const models incur a high FPR. Surprisingly, for hate speech, a higher FPR does not translate to higher F1, indicating that a combination of different types of CAD (CF_mix) reduces unintended bias without sacrificing F1 score.

Overall, we find that CF models have higher F1, especially models trained on construct-driven CAD (for e.g., the sexism CF_const model for HC). A potential reason for this is that construct-driven CAD is obtained by editing identity words; while identity words indeed co-occur with hate speech or sexism, they can also have a confounding impact, i.e., sexism manifests via attacks on gender identity, however mentioning gender is not always associated with sexism. Indeed, many minorities may disclose their experience and identity using such terms without being sexist or hateful. The confounding nature of identity terms makes automated methods all the more vulnerable to unintended false positive bias. Our analysis reveals that while construct-driven CAD has its uses and is often easier to generate (based on the distribution of different types of CAD (Sen et al., 2021)), we should use them judiciously. Specifically, we need more research on various characteristics of individual CAD such as minimality, semantic distance from the original instance, as well as corpus-level attributes like training data composition and diversity.

5 Conclusion

Text counterfactuals, drawing from and informing current developments for causal inference in NLP (Keith et al., 2021; Feder et al., 2021; Jin et al., 2021) can be used for training, testing, and explaining models (Wu et al., 2021). We build on research that explores the former — using counterfactually augmented data (CAD) as training examples, to in-

vestigate conditions of model robustness and unintended false positive bias. Performing experiments on challenging datasets for hate speech and sexism detection, we find that models trained on CAD have higher false positive rates compared to those that are not. In addition, models trained on construct-driven counterfactuals tend to have the highest false positive rate. 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. Future work includes unpacking the strengths and weaknesses of different types of CAD by studying their various characteristics, including but not limited to the exact changes made to derive a counterfactual and their impact on unintended bias.

6 Ethical Considerations

Constructs like sexism and hate speech detection are often depicted as neutral or objective, but they are deeply contextual, subjective, and ambiguous (Vidgen et al., 2019; Jurgens et al., 2019; Nakov et al., 2021). Promoting features like identity terms can increase the risk of misclassifying non-hateful content with such terms, such as disclosures or reports of facing hate speech, leading to unintended bias (Dixon et al., 2018) that can cause harm (Blackwell et al., 2017). Following up on this subjectivity and based on recommendations by Blodgett et al. (2020), we motivate our analyses of unintended bias on normative grounds, situated in the context of the harms wrought by misclassification of content containing identity terms despite being non-sexist or non-hateful. We acknowledge that we only study one type of unintended bias and there are other aspects that require further investigation (Blackwell et al., 2017).

Acknowledgements

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact Checking with Insufficient Evidence. *Transactions of the Association for Computational Linguistics (TACL), to appear*, 10.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Agostina Calabrese, Michele Bevilacqua, Björn Ross, Rocco Tripodi, and Roberto Navigli. 2021. Aaa: Fair evaluation for abuse detection systems wanted. In *13th ACM Web Science Conference 2021*, pages 243–252.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Tarleton Gillespie. 2018. *Custodians of the Internet*. Yale University Press.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

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

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Kishonna L Gray and Krysten Stein. 2021. We 'said her name'and got zucked": Black Women Callingout the Carceral Logics of Digital Platforms. *Gender & Society*, 35(4):538–545.

Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Katherine Keith, Douglas Rice, and Brendan O'Connor. 2021. Text as causal mediators: Research design for causal estimates of differential treatment of social groups via language aspects. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 21–32, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442.

Fereshte Khani and Percy Liang. 2021. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 196–205.

Preslav Nakov, Vibha Nayak, Kyle Dent, Ameya Bhatawdekar, Sheikh Muhammad Sarwar, Momchil Hardalov, Yoan Dinkov, Dimitrina Zlatkova, Guillaume Bouchard, and Isabelle Augenstein. 2021. Detecting abusive language on online platforms: A critical analysis. *arXiv preprint arXiv:2103.00153*.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended Bias in Misogyny Detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.

Francisco Rodriguez-Sanchez, Jorge Carrillo de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0).

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. Call me sexist, but...: Revisiting sexism detection using psychological scales and adversarial samples.

Indira Sen, Mattia Samory, Fabian Flöck, Claudia Wagner, and Isabelle Augenstein. 2021. How does counterfactually augmented data impact models for social computing constructs? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 325–344, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving

models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Appendix

This is the appendix for the paper, "Counterfactu-ally Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection". The appendix contains results of the logistic regression models trained on different types of CAD (7), per-formance on in-domain data (8), and details for facilitating reproducibility (9).

7   Performance of LogReg Models trained on different types of CAD

In Figures 3 and 4, we present th results for logis-tic regression models trained on different types of CAD. We note that, similar to results for the CF and nCF models, logistic regression models have lower performance than BERT models. Further-more, we also note that, like the BERT models, the logistic regression models trained on construct-driven CAD (CF_const) have high false positive rates compared to models trained on other types of CAD.

8   In-domain Performance

In Table 4, we report the performance of the CF and nCF models on the in-domain datasets. To ensure fair comparison with the results in 2, instead of computing results on the entire test set, we subset it in a manner similar to ISG; i.e., we retain only those instances which have identity words for hate speech and gender word for sexism. The results are in line with what Sen et al. reported — nCF models perform better in the in-domain datasets. We note that even though CF models have lower F1 score, they have a higher FPR even in-domain.

We report the results of models trained on differ-ent types of CAD in Table 5. Notably, the CF_const models have the highest FPR similar to results on ISG and HC, but also have the lowest F1 score in the in-domain subset.

9   Reproducibility

9.1   Compute Infrastructure

For the logistic regression models we used the scikit learn package (Pedregosa et al., 2011) and for finetuning BERT, we used the Transformers library from HuggingFace (Wolf et al., 2020). All mod-els were trained or finetuned on a 40 core Intel(R) Xeon(R) CPU E5-2690 (without GPU).
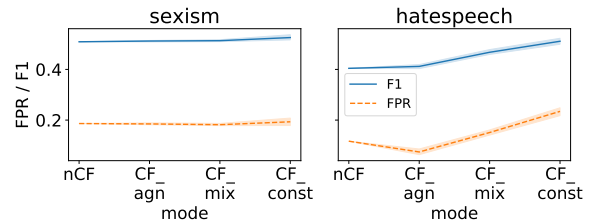
Figure 3: FPR and F1 for different types of CF Lo-gistic Regression models, with the nCF model as a baseline on ISG. For hate speech, CF_const has the highest FPR and the highest F1, while the models per-form equally well for sexism.
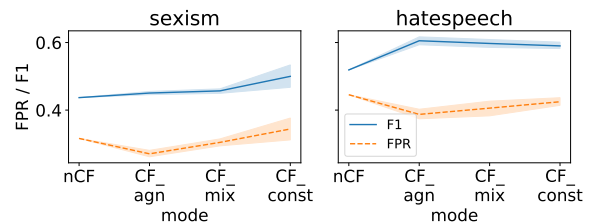
Figure 4: F1 and FPR for different types of CF logis-tic regression models for HC, with the nCF model as a baseline. Models trained on construct-driven CAD (CF_const) incur the high FPR, especially for sexism.

9.2   Model Training Details: Hyperparameters and Time Taken

We preprocess all the data by removing social me-dia features such as hashtags and mentions. The hyperparameter bounds for LR models are:

1. stopwords: English, none, English without nega-tion words
2. norm: ('l1', 'l2')
3. C: (0.01, 0.1, 1)
4. penalty: ('l2', 'l1')

while for BERT we use:

1. epochs:[4, 5]
2. learning rate: 2e-5, 3e-5, 5e-5

For LR, we have 36 combinations over 5 fold cross-validation, leading to 180 fits, while for BERT, we have 6 combinations also over 5 fold CV, leading to 30 fits.

We use gridsearch for determining hyperparam-eter, where the metric for selection was macro F1. Run times and hyperparameter configuartions for the best performance for all CF (with randomly sampled 50% data) and nCF models (RQ1) are in-cluded in Table 6. The hyperparameters and run times for the CF models trained on different types of CAD (RQ2) are in Table 7.

# In-domain Subset	pos class prop	construct	model	mode	macro F1	FPR	FNR
105	65	hate speech	bert	CF	0.92	<u>0.10</u>	0.06
				nCF	**0.99**	0.00	0.02
			logreg	CF	0.77	<u>0.20</u>	0.23
				nCF	**0.90**	0.10	0.09
886	501	sexism	bert	CF	0.76	<u>0.41</u>	0.09
				nCF	**0.81**	0.21	0.17
			logreg	CF	0.65	<u>0.44</u>	0.26
				nCF	**0.75**	0.24	0.26

Table 4: Performance of CF and nCF models on the subset of the in-domain dataset containing identity terms. As reported in (Sen et al., 2021) and in contrast to out-of-domain performance, nCF models have higher F1 scores in the in-domain dataset, while CF models still have higher FPR as they do in ISG.

construct	model	mode	macro F1	FPR	FNR
hate speech	bert	CF_mix	0.98	0.02	0.02
		CF_const	0.96	<u>0.08</u>	0.02
		CF_agn	0.99	0.00	0.02
	logreg	CF_mix	0.90	0.12	0.08
		CF_const	0.88	<u>0.15</u>	0.09
		CF_agn	0.95	0.02	0.06
sexism	bert	CF_mix	0.80	0.25	0.16
		CF_const	0.79	<u>0.29</u>	0.14
		CF_agn	0.80	0.27	0.13
	logreg	CF_mix	0.74	0.25	0.27
		CF_const	0.68	<u>0.35</u>	0.28
		CF_agn	0.74	0.23	0.28

Table 5: Performance of different types of CF models on the subset of the in-domain dataset containing identity terms. Models trained on construct-driven CAD (CF_const) have the lowest F1 score while having the highest FPR.

9.3 Metrics

The evaluation metrics used in this paper are macro average F1, False Positive Rate (FPR) and False Negative Rate (FNR). We used the sklearn implementation of the macro F1 score: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html. The code for computing FPR and FNR is included in our code (uploaded with the submission)

9.4 Model Parameters

Model parameters are included in Table 8.

construct	model	best model hyperparameters	time to train (one run)
sexism	CF LR	english, l2, 0.01, l2	5.42s
	CF BERT	epochs: 5, learning rate: 2e-5	3h42m20s
	nCF LR	none, l2, 0.01, l2	4.87s
	nCF BERT	epochs: 5, learning rate: 2e-5	3h38m57s
hate speech	CF LR	english without negation, l2, 0.01, l2	26.27s
	CF BERT	epochs: 4, learning rate: 5e-5	17h54m03s
	nCF LR	english without negation, l2, 0.01, l2	26.67s
	nCF BERT	epochs: 5, learning rate: 5e-5	17h39m29s

Table 6: Hyperparameters for CF (trained on 50% CAD) and nCF models.

construct	model	best model hyperparams	time to train (one run)
sexism	CF_c LR	english, l1, 1, l1	5.91s
	CF_a LR	english without negation, l1, 1, l1	6.15s
	CF_r LR	english, l2, 0.1, l2	5.27s
	CF_c BERT	epochs: 5, learning rate: 5e-5	3h42m20s
	CF_a BERT	epochs: 5, learning rate: 3e-5	3h34m36s
	CF_r BERT	epochs: 5, learning rate: 2e-5	3h50m18s
hate speech	CF_c LR	english without negation, l1, 1, l1	33.35s
	CF_a LR	english without negation, l1, 0.1, l1	30.08s
	CF_r LR	none, l1, 0.1, l1	32.67s
	CF_c BERT	epochs: 5, learning rate: 3e-5	18h09m11s
	CF_a BERT	epochs: 5, learning rate: 3e-5	17h58m33s
	CF_r BERT	epochs: 5, learning rate: 2e-5	17h49m46s

Table 7: CF models trained on different types of CAD.

construct	model	#params
Sexism	CF LR	4750
	nCF LR	5505
	CF BERT	110M
	nCF BERT	
Hate speech	CF LR	13763
	nCF LR	14800
	CF BERT	110M
	nCF BERT	

Table 8: Number of model parameters for the CF and nCF models.