Counterfactually Augmented Data and Unintended Bias
The Case of Sexism and Hate Speech Detection
Sen, Indira; Samory, Mattia; Wagner, Claudia; Augenstein, Isabelle

Published in:
Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

DOI:
10.18653/v1/2022.naacl-main.347

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection

Indira Sen1 Mattia Samory1 Claudia Wagner1,2 Isabelle Augenstein3
1GESIS – Leibniz Institute for the Social Sciences 2RWTH Aachen University 3University 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/Unintended_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). 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).
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).

Table 2: 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.

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.

Table 4: 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.

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. This can be especially harmful if the misclassified posts happen to be reports or disclosures of facing hate

4.1 Test Sets for Measuring Unintended Bias

To understand if CF models facilitate this type of unintended bias, we leverage two test 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 Subgroup.
### Table 3: Macro F1 and FPR on Hatecheck for models trained on CAD (CF) vs those trained on original data (nCF).

<table>
<thead>
<tr>
<th>Model</th>
<th>Mode</th>
<th>F1</th>
<th>FPR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>nCF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logreg</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logreg</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 shows the result of CF and nCF models on ISG while Table 3 shows the results for HateCheck.

#### 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 false 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.
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 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

We thank the members of the Computational Social Science department at GESIS, the CopeNLU group, and the anonymous reviewers for their constructive feedback. Isabelle Augenstein’s research is partially funded by a DFF Sapere Aude research leader grant with grant number 0171-00034B.
References


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.


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

This is the appendix for the paper, “Counterfactually 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), performance 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 the results for logistic 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. Furthermore, 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 different 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 models were trained or finetuned on a 40 core Intel(R) Xeon(R) CPU E5-2690 (without GPU).

9.2 Model Training Details: Hyperparameters and Time Taken

We preprocess all the data by removing social media features such as hashtags and mentions. The hyperparameter bounds for LR models are:
1. stopwords: English, none, English without negation 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 hyperparameters, where the metric for selection was macro F1. Run times and hyperparameter configurations for the best performance for all CF (with randomly sampled 50% data) and nCF models (RQ1) are included in Table 6. The hyperparameters and run times for the CF models trained on different types of CAD (RQ2) are in Table 7.
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.

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.
<table>
<thead>
<tr>
<th>construct</th>
<th>model</th>
<th>best model hyperparameters</th>
<th>time to train (one run)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sexism</td>
<td>CF LR</td>
<td>english, 12, 0.01, 12</td>
<td>5.42s</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>epochs: 5, learning rate: 2e-5</td>
<td>3h42m20s</td>
</tr>
<tr>
<td></td>
<td>nLR</td>
<td>none, 12, 0.01, 12</td>
<td>4.87s</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>epochs: 5, learning rate: 2e-5</td>
<td>3h38m57s</td>
</tr>
<tr>
<td>hate speech</td>
<td>CF LR</td>
<td>english without negation, 12, 0.01, 12</td>
<td>26.27s</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>epochs: 4, learning rate: 5e-5</td>
<td>17h54m03s</td>
</tr>
<tr>
<td></td>
<td>nLR</td>
<td>english without negation, 12, 0.01, 12</td>
<td>26.67s</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>epochs: 5, learning rate: 5e-5</td>
<td>17h39m29s</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>construct</th>
<th>model</th>
<th>best model hyperparams</th>
<th>time to train (one run)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sexism</td>
<td>CF c</td>
<td>english, 11, 1, 11</td>
<td>5.91s</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>english without negation, 11, 1, 11</td>
<td>6.15s</td>
</tr>
<tr>
<td></td>
<td>r</td>
<td>english, 12, 0.1, 12</td>
<td>5.27s</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>epochs: 5, learning rate: 5e-5</td>
<td>3h42m20s</td>
</tr>
<tr>
<td></td>
<td>CF a</td>
<td>epochs: 5, learning rate: 3e-5</td>
<td>3h34m36s</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>epochs: 5, learning rate: 2e-5</td>
<td>3h50m18s</td>
</tr>
<tr>
<td>hate speech</td>
<td>CF c</td>
<td>english without negation, 11, 1, 11</td>
<td>33.35s</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>english without negation, 11, 0.1, 11</td>
<td>30.08s</td>
</tr>
<tr>
<td></td>
<td>r</td>
<td>none, 11, 0.1, 11</td>
<td>32.67s</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>epochs: 5, learning rate: 3e-5</td>
<td>18h09m11s</td>
</tr>
<tr>
<td></td>
<td>CF a</td>
<td>epochs: 5, learning rate: 3e-5</td>
<td>17h58m33s</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>epochs: 5, learning rate: 2e-5</td>
<td>17h49m46s</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>construct</th>
<th>model</th>
<th>#params</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sexism</td>
<td>CF LR</td>
<td>4750</td>
</tr>
<tr>
<td></td>
<td>nCF LR</td>
<td>5505</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>110M</td>
</tr>
<tr>
<td></td>
<td>nCF BERT</td>
<td>110M</td>
</tr>
<tr>
<td>Hate speech</td>
<td>CF LR</td>
<td>13763</td>
</tr>
<tr>
<td></td>
<td>nCF LR</td>
<td>14800</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>110M</td>
</tr>
<tr>
<td></td>
<td>nCF BERT</td>
<td>110M</td>
</tr>
</tbody>
</table>

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