Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection

Alan Ramponi
Fondazione Bruno Kessler
Trento, Italy
alramponi@fbk.eu

Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
satonelli@fbk.eu

Abstract

⚠ **Warning**: *this paper contains content that may be offensive or upsetting.*

Avoiding to rely on dataset artifacts to predict hate speech is at the cornerstone of robust and fair hate speech detection. In this paper we critically analyze lexical biases in hate speech detection via a cross-platform study, disentangling various types of spurious and authentic artifacts and analyzing their impact on out-of-distribution fairness and robustness. We experiment with existing approaches and propose simple yet surprisingly effective data-centric baselines. Our results on English data across four platforms show that distinct spurious artifacts require different treatments to ultimately attain both robustness and fairness in hate speech detection. To encourage research in this direction, we release all baseline models and the code to compute artifacts, pointing it out as a complementary and necessary addition to the data statements practice.[1]

1 Introduction

Hate speech in online social communities is a serious and pervasive concern, which requires fair and robust automated approaches to be tackled at scale. However, despite the great progress in natural language processing for detecting hate speech, current models have shown to be brittle when applied to real-world data, exhibiting limited out-of-distribution (OOD) robustness (Vidgen et al., 2019) and perpetuating and amplifying harmful social biases (Röttger et al., 2021). Noticeably, hate speech detection systems are typically trained on data from limited language varieties such as individual platforms, which inevitably exhibit differences in writing norms, language use, and hate targets, hampering generalization (Vidgen and Derczynski, 2020).

One of the main reasons for limited robustness and fairness of mainstream hate speech detection systems is largely ascribable to spurious statistical correlations between surface lexical items and labels in training data, which models exploit to derive predictions. These biases are commonly referred to as lexical **dataset artifacts**, and have recently attracted attention in the NLP community, particularly in natural language inference (NLI) studies (Belinkov et al., 2019; Gururangan et al., 2018; Poliak et al., 2018, *inter alia*). Efforts to tackle the issue in hate speech detection are instead rather scattered, and mainly focus on fairness using datasets from few platforms (Zhou et al., 2021; Kennedy et al., 2020b, *inter alia*), leaving the study on OOD robustness largely unexplored. We instead argue that fairness and robustness are strongly intertwined aspects (Table 1), and thus should be studied jointly, with the goal to understand to what extent these two dimensions are related.

	fair	robust
what have jews done to you?	✗	
RT [user]: I'm mad at this		✗
All black people literally go there	✗	✗

Table 1: Posts wrongly labeled as hateful by a fine-tuned BERT classifier due to the presence of spurious lexical artifacts (identity and non identity-related) and their negative impact (✗) on fairness and robustness.

Previous work has shown that state-of-the-art models overly rely on identity words (e.g., "jews", "gay") to predict hateful content (Zhou et al., 2021; Kennedy et al., 2020b, *inter alia*), further demoting voices of people from already marginalized groups (Bender et al., 2021). However, non identity-related lexical items – such as "sport", "announcer", and "football" in Waseem and Hovy (2016) – are also often spuriously associated with hate speech due to a biased data collection process (Wiegand et al., 2019), undermining OOD robustness. Despite the recent trend in minimizing

[1] Code and resources are available at https://github.com/dhfbk/hate-speech-artifacts.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3027 - 3040

topic bias in data sampling, we show that some spurious lexical artifacts still remain highly-predictive on certain distributions even if data has been sampled in a more attentive fashion (e.g., artifacts that are potentially data- or platform-specific – Figure 1, highlighted in gray).

We argue that disentangling artifacts into fine-grained categories by means of a cross-platform analysis may be beneficial to drive a broader debiasing of current hate speech models, ultimately improving both fairness and robustness to out-of-distribution data. To this purpose, we critically analyze artifacts in hate speech detection across multiple platforms and propose simple yet effective data-centric baselines exploiting spurious lexical items. We show that although we achieve substantial improvements in OOD fairness by exploiting spurious identity-related artifacts, this comes at the cost of robustness. This confirms that fairness and robustness are strictly interrelated dimensions that should be studied together in future research.

Contributions To the best of our knowledge, we are the first to (i) conduct a thorough investigation of lexical artifacts across online platforms; (ii) disentangle artifacts into fine-grained categories; and (iii) propose a viable data-centric approach based on masking that consistently improves fairness over *all* baselines across *all* platforms. To foster future research on the topic, we also release (iv) code to reproduce all experiments, and (v) disaggregated lexical artifact annotations, more broadly (vi) suggesting the inclusion of dataset artifacts in data statements (Bender and Friedman, 2018), which can be easily revealed using our codebase.

2 Lexical Artifacts are *not* all the Same

We conceptualize dataset artifacts at the lexical level as *emergent correlations between tokens and labels in input data*, consistently to lexical annotation artifacts in NLI (Gururangan et al., 2018). As such, given a target class c, we formally define lexical artifacts $\mathcal{L}_c$ as the set of highly-discriminating[2] tokens for c, which comprise **authentic artifacts** $\mathcal{A}_c$ – items that potentially carry useful information for the class at hand – and **spurious artifacts** $\mathcal{S}_c$ – items that are spuriously (or undesirably) associated to the target class – such that $\mathcal{L}_c = \mathcal{A}_c \cup \mathcal{S}_c$. In the context of hate speech detection, we consider the hateful class as c unless otherwise specified and simplify the notation (i.e., from "$\cdot_c$" to "$\cdot$").

[2]Highly-discriminating tokens can be computed and filtered using information theory and statistics measures.

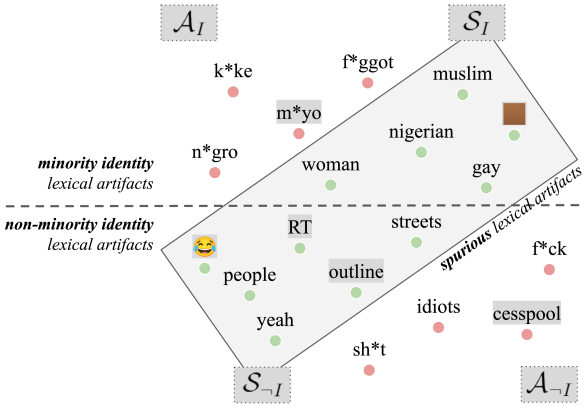

Figure 1: Illustration of lexical artifacts in hate speech detection, including relevant examples. $\mathcal{A}_I$: authentic artifacts, identity-related; $\mathcal{A}_{\neg I}$: authentic artifacts, non identity-related; $\mathcal{S}_I$: spurious artifacts, identity-related; $\mathcal{S}_{\neg I}$: spurious artifacts, non identity-related. Items highlighted in gray are potentially platform-specific.

We build our definitions on top of the categories of lexical biases by Zhou et al. (2021), which originally identify three bias groups: *i)* minority identity mentions which are not offensive, *ii)* minority identity mentions which are potentially offensive, and *iii)* non-identity mentions which are possibly offensive. We enrich this categorization by introducing a high-level separation into spurious (i.e., group *i)* in Zhou et al. (2021)) and authentic artifacts (i.e., group *ii)* and *iii)*), and including an additional spurious, non identity-related category (Section 2.2).

Indeed, given the broad nature of authentic and spurious artifacts, we further categorize them in Section 2.1 and 2.2 (see Figure 1 for an overview).

2.1 Authentic artifacts

We define authentic lexical artifacts $\mathcal{A}$ as the subset of highly-discriminating tokens which potentially convey hatefulness, profanity, or are otherwise frequently associated with hateful contexts. Intuitively, $\mathcal{A}$ is the set of artifacts which is likely to be informative to detect hate speech across distributions. Authentic artifacts enclose minority identity-related artifacts $\mathcal{A}_I$ and non-identity artifacts $\mathcal{A}_{\neg I}$.

Identity-related ($\mathcal{A}_I$) Potentially offensive or stereotyping terms towards minority identities (e.g., "n*gro", "f*ggot", "k*ke", "wh*re"), as well as reclaimed slurs (e.g., "n*gga") (Figure 1, top left).

Non-identity related ($\mathcal{A}_{\neg I}$) Swear words and profanities (e.g., "f*ck", "sh*t") as well as broad

terms typically associated with hateful contexts (e.g., "kill", "idiots") (Figure 1, bottom right).

2.2 Spurious artifacts

Spurious lexical artifacts $\mathcal{S}$ broadly enclose all tokens which we do not expect to be predictive for the target class at hand. As such, we postulate that those artifacts are a main reason for insufficient robustness and fairness of current hate speech detectors, and thus may play a positive role in lexical debiasing. We specifically focus on these artifacts in our experiments. As for authentic artifacts, spurious items can be grouped into minority identity-related artifacts $\mathcal{S}_I$ and non-identity artifacts $\mathcal{S}_{\neg I}$, the latter being currently disregarded in research investigating fairness only (Zhou et al., 2021).

Identity-related ($\mathcal{S}_I$) Terms describing a social minority, which are typically associated to hate speech due to their frequency on offensive statements on online fora (e.g., "muslim", "woman", "Islam", "nigerian", "LGBT") (Figure 1, top right).

Non-identity related ($\mathcal{S}_{\neg I}$) All non-identity tokens which are unexpectedly associated to hate speech, e.g., due to platform-specificity, bias in collection timeframe, etc. (e.g., "people", "RT", "streets", "Trump", "yeah") (Figure 1, bottom left).

3 Data

In this work we focus on hate speech, i.e., messages whose content spreads hatred or incites violence, or threatens people's freedom, dignity and safety, and whose target is a protected group, or an individual targeted for belonging to such a group and not for his/her individual characteristics (Poletto et al., 2021). Hate speech typically encompasses serious cases of offense with severe moral and legal implications, i.e., those cases that are of primary importance for content moderation.

We collect hate speech corpora that meet the following criteria: (i) they minimize topic and author biases in data collection (Wiegand et al., 2019), using alternatives to keyword and user searches such as pure or boosted random sampling, (ii) they pertain to different social media platforms, and (iii) they follow similar annotation guidelines, where hate speech is clearly defined and separated from other types of offensive language. For each corpus we create *hateful* and *non-hateful* examples. All datasets follow consistent preprocessing, deduplication, and anonymization (Appendices A.1 and A.2).

REDDIT ( ) We use the recently introduced Reddit dataset (v1.1) by Vidgen et al. (2021) which preserves a variety of grammar, topic, and style features due to a community-based sampling approach. The corpus contains 27,494 entries annotated following a hate speech taxonomy comprising abusive (*identity-directed*, *affiliation-directed*, *person-directed*) and non-abusive labels (*non-hateful slurs*, *counter speech*, and *neutral*). We follow the widely accepted definition of hate speech as "abuse targeting a protected group or its members for being a part of that group"[3] (Röttger et al., 2021; Banko et al., 2020; Vidgen et al., 2019, *inter alia*) to create the *hateful* label from *identity-directed* examples, and the *non-hateful* label from the remaining examples. For the purpose of this study, we discard instances marked as requiring previous content to be interpreted.[4] The final dataset after preprocessing consists of 1,688 *hateful* and 19,888 *non-hateful* examples, for a total of 21,576 unique instances.

TWITTER ( ) We select a widely used hate speech dataset which has been collected following a bootstrap random sampling approach (Founta et al., 2018). The dataset consists of 99,996 tweets annotated as *hateful*, *abusive*, *spam*, and *normal*. Similarly to previous work, we discard the *spam* category (Zhou et al., 2021), forming the *hateful* class following the original classification provided by the authors. This led to 3,937 *hateful* and 70,554 *non-hateful* examples, for a total of 74,491 tweets.

GAB ( ) We use the GAB hate corpus by Kennedy et al. (2020a), whose data has been sampled purely randomly due to the frequency of hate speech of the "free speech-preserving" (Zannettou et al., 2018) GAB social network. The corpus (v.2021-03-03) consists of 27,546 posts annotated with *(assault on) human dignity*, *call for violence*, and *vulgarity/offensive* labels. Similarly to previous work, we take the union of *human dignity* and *call for violence* labels for the *hateful* class (Kennedy et al., 2020b), whereas we create the *non-hateful* class from the remaining examples. We also leverage target annotations and consider messages towards ideology/political groups as *non-hateful*, to ensure consistency among datasets. As

[3] Groups based on age, disability, familial status, gender identity, national/ethnic origins, pregnancy, race, religion, sex or sexual orientation, as defined in Röttger et al. (2021), which in turn reflects the US 1964 Civil Rights Act, the EU's Charter of Fundamental Rights, and the UK's 2010 Equality Act.

[4] We leave the investigation of lexical artifacts in context-aware hate speech detection for future work.

a result, the final dataset is made up of 27,014 messages: 1,785 *hateful* and 24,829 *non-hateful*.

STORMFRONT (💬) We use the dataset pertaining to a white supremacist web forum collected by de Gibert et al. (2018) following a random sampling procedure. It consists of a total of 10,944 messages with annotations for *hate*, *no-hate*, *relation*, and *skip* labels. We remove *relation* and *skip* examples, since they require previous context, or they represent spam / content written in other languages, respectively. We then use the *hate* examples for the *hateful* class, and the *no-hate* instances for the *non-hateful* class. This led to a total of 10,448 examples, 1,192 *hateful* and 9,256 *normal*.

4 Disentangling Lexical Artifacts

In order to disentangle lexical artifacts, we first compute the correlation between each token and the hateful class for each dataset (Section 4.1), then assessing the cross-distribution indicativeness of each token (Section 4.2). For segmenting texts into tokens, we rely on the training portion of each dataset only (Section 5) and employ the WordPiece (Schuster and Nakajima, 2012) subword tokenizer as used in BERT (Devlin et al., 2019).[5] Finally, we perform lexical artifacts annotation (Section 4.3) following the categories defined in Section 2.

4.1 In-distribution artifacts

We follow Gururangan et al. (2018) and employ pointwise mutual information (PMI; Fano, 1961) to compute the discriminativeness of each token to the target class.[6] Since lexical artifacts are meant to be used for downstream debiasing, we argue that tokens should be consistent with inputs to the end model. As a result, we use tokens as given by the WordPiece subword tokenizer, the same tokenizer used by models employed in our experiments (Section 5). Formally, given a token t and a class c, the PMI is defined as follows:

$$\text{PMI}(t, c) = log \frac{p(t, c)}{p(t|\cdot)p(\cdot|c)} \quad (1)$$

We further apply reweighting to emphasize highly-discriminative token-class correlations, and normalize ≤ 0 values to zero since negative PMI scores are known to be unreliable on relatively small corpora (Jurafsky and Martin, 2021, Ch. 6).

[5] In preliminary experiments we found similar results when using the byte-level BPE tokenizer (Sennrich et al., 2016) as used in RoBERTa (Liu et al., 2019).

[6] A comparative assessment of different metrics for computing token-class correlations is out-of-scope in this study and will be investigated in future work.

Rank	🤖	🐦	🐸	💬	Avg.
1	##tar	##gga	white	n*gro	##s
2	##ded	hate	jews	white	white
3	##s	rt	##gger	black	black
4	fa	##s	##s	##s	jews
5	b*tch	[user]	jew	jews	hate
6	##gg	idiot	islam	whites	##es
7	gay	trump	muslim	blacks	women
8	women	ass	whites	jew	people
9	##ds	idiots	##gg	race	##tar
10	f*cking	people	women	##es	jew

Table 2: Top 10 most informative tokens for the hateful class according to PMI, divided per platform dataset (left), and after cross-distribution computation (right).

Discussion The top 10 tokens on each platform that are more associated with the hateful class are presented in Table 2 (left). All platforms exhibit a variety of lexical artifact types (cf. Section 2); however, we observe clear divergences across distributions. While artifacts in Stormfront data are mainly related to race, on Gab the focus is more on religion. Reddit and Twitter conversations are instead more varied, with higher occurrence of spurious, non-identity artifacts (e.g., "RT", "people").

4.2 Cross-distribution artifacts

When datasets from multiple platforms are available, we hypothesize that leveraging individual scores makes possible to better identify artifacts. Given $\text{PMI}(t, c)^d$ the score of a token t for the hateful class c on a given distribution $d \in D$ (e.g., platform), we normalize it in $[0, 1]$ by applying a min-max normalization function to enable cross-platform score comparability – obtaining $\text{PMI}(t, c)^d_{[0,1]}$ – further applying a log_2 transformation to mitigate the skewness of the original PMI distribution. As a result, the final cross-distribution score $S(t, c)$ for each token is given by the average of the corresponding individual scores:

$$S(t, c) = \frac{1}{|D|} \sum_{d=1}^{D} log_2(\text{PMI}(t, c)^d_{[0,1]}) \quad (2)$$

We then sort tokens by descending score, highlighting lexical artifacts that are highly discriminating for the hateful class across distributions. Table 2 (right) shows the top 10 tokens after the cross-platform computation is carried out.

Discussion As shown in Table 2 (right), cross-platform importance of tokens for the hateful class demotes scores (and thus, ranks) of lexical artifacts which are likely to be more indicative on some platforms only (e.g., "RT"), while consolidating the informativeness of cross-platform items (e.g., "jews", "hate", "##s", "##es",[7] "people"). This confirms our hypothesis that encompassing multiple platforms is beneficial for capturing lexical items that are likely to be predictive *across* distributions.

4.3 Artifacts annotation

In order to disentangle lexical artifacts for further debiasing, we select the k most predictive tokens given by the cross-distribution rank of discriminativeness (Section 4.2) to be manually annotated.[8] In our experiments, we set $k = 200$ as it matches the subset of tokens which are highly informative (≥ 0.33).[9] All k tokens have been labeled as potentially hateful and/or related to minority identities by two annotators – male and female, fluent in English – with background in linguistics and NLP, and past experience in hate speech activities with NGOs. Each annotator was provided with five examples of tokens in context for enabling more informed annotation decisions, represented by randomly sampled posts from the four platforms included in this study.

After annotation, the two annotators were involved in an adjudication session in order to discuss the cases of disagreement, followed by correction wherever possible. We calculate the inter-annotator agreement (IAA) score before and after adjudication using Cohen's kappa (Cohen, 1960). We obtain $\kappa = 0.6887$ before and $\kappa = 0.8311$ after the adjudication session, which is high agreement.

Discussion Although some cases of disagreement were easily resolvable (e.g., annotation errors), we found tokens which are difficult to discern due to ambiguity – mostly subwords – or due to real disagreement in the interpretation of the terms. This is in line with existing works showing that disagreement in toxicity annotation is inherent to the task and cannot always be solved through majority voting or adjudication (Aroyo et al., 2019; Leonardelli et al., 2021). Interestingly, this disagreement follows the trend of cross-distribution rank of artifacts (Figure 2). We decided to leave the analysis of annotators' disagreement for future work, and we release these cases as disaggregated labels. In Table 3 we show the most informative artifacts by type, whereas the full list of spurious artifacts used in the experiments is in Appendix B.

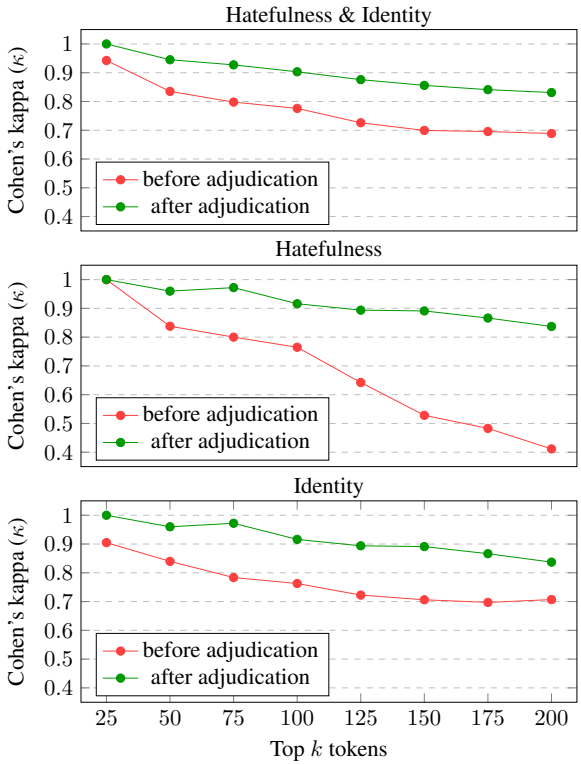

Figure 2: Cumulative Cohen's kappa (κ) scores for the full annotation of lexical artifacts (top), and for decisions on potentially hateful or identity-related artifacts only (middle and bottom, respectively), ordered by informativeness according to cross-distribution scores.

5 Experiments

We investigate the impact of spurious lexical artifacts on fairness and robustness in hate speech detection. Similarly to previous studies (Kennedy et al., 2020a; Röttger et al., 2021), we cast hate speech detection as a binary classification problem, where the two classes to be predicted are *hateful* and *non-hateful*, as defined in Section 3. We carry out in-distribution and OOD experiments, namely training and testing all models on the same or different platform data, respectively. We evaluate performance of models using macro F_1 score, whereas for fairness we use false positive rate (FPR) on test instances containing $\mathcal{S}_I$ mentions, consistently to

[7] We found "##s" and "##es" tokens typically correspond to plural suffixes of out-of-vocabulary words.

[8] The main advantage of token-level annotation compared to word-level annotation is that it allows to discern generic subwords from hateful or identity-related ones – e.g., "homophobia" $\mapsto$ {"homo", "##phobia"} – without losing important information when doing removal or masking (Section 5).

[9] We leave the investigation of larger thresholds for future work due to space and annotation constraints.

Authentic artifacts		*Spurious artifacts*	
$\mathcal{A}_I$	$\mathcal{A}_{\neg I}$	$\mathcal{S}_I$	$\mathcal{S}_{\neg I}$
retar*s	hate	white	_s
b*tch	dumb	black	_es
n*gro	stupid	jews	people
f*ggot	disgusting	women	country
n*gger	kill	jew	_ing
n*gga	racist	whites	anti
r*tarded	filthy	blacks	illegal
fa*s	evil	muslim	bunch
f*gs	ass	gay	_t
f*ggot	rape	muslims	kids

Table 3: Top 10 tokens per artifact type after annotation. Gray letters indicate the most common token prefixes/suffixes which occur with the lexical items. If a variety of them is available, we indicate it with "_".

previous work (Zhou et al., 2021).

We outline the experimental setup in Section 5.1, whereas data-centric baselines are presented in Section 5.2. Lastly, we present results and a thorough discussion in Section 5.3.

5.1 Experimental setup

For all our experiments, we employ the uncased BERT-base model (Devlin et al., 2019) as implemented in the MaChAmp v0.2 toolkit (van der Goot et al., 2021), since it has been shown to achieve state-of-the-art performance in hate speech detection (Tran et al., 2020). We use default hyperparameters, and perform a grid search to determine the number of epochs, the learning rate, and the batch size, using the search space suggested by Devlin et al. (2019) – i.e., [2, 3, 4] for epochs, [2e-05, 3e-05, 5e-05] for learning rate, and [16, 32] for batch size. We use stratified 80% train, 10% development, 10% test splits for each dataset, selecting the best model based on the average macro F_1 score on the development test across all platforms. During fine-tuning, we emphasize the minority hateful class using a cross-entropy loss with balanced class weights. The final hyperparameters are: 4 for epochs, 2e-05 for learning rate, and 16 for batch size. All experiments have been run on a NVIDIA Tesla V100-SXM2 GPU, with a training time ranging from 10 to 40 minutes each. The number of trainable parameter for all models are ≈110M.

5.2 Baselines

We investigate the impact of spurious identity-related and non identity-related lexical artifacts on the robustness and fairness of hate speech detection by employing the following data-centric baselines.

VANILLA We fine-tune the BERT-base model on each corpus, so that the proposed baselines can be directly compared to a commonly employed setup.

FILTERING Swayamdipta et al. (2020) have shown that the most ambiguous training data instances promote OOD generalization while preserving in-distribution performance. We thus leverage the VANILLA model's training dynamics to filter training data to contain the 33% most ambiguous instances only, in line with the subset size in Swayamdipta et al. (2020).[10] Intuitively, those are instances whose class probabilities fluctuate frequently across training epochs. We then fine-tune the BERT-base model on the resulting subset. This setup is similar in spirit to the one employed in Zhou et al. (2021); however, we assess it on data from multiple platforms, also removing duplicate instances (Appendix A.2) which may potentially confound the debiasing results.

REMOVAL Prior to fine-tuning, we naively remove any occurrence of spurious lexical artifacts from training and development data. This matches previously employed baselines for assessing fairness in hate speech detection (Kennedy et al., 2020b). However, since we are also interested in OOD robustness, we experiment with two removal variants: one for $\mathcal{S}_I$ and one for $\mathcal{S}_{\neg I}$ artifacts. We hypothesize that removing $\mathcal{S}_I$ tokens potentially improves fairness, whereas removing $\mathcal{S}_{\neg I}$ tokens mostly contributes to OOD robustness.

MASKING We propose a novel data-centric debiasing alternative based on token masking. Instead of removing spurious artifacts altogether, we reserve a special token in the vocabulary of the model that we use as replacement for spurious artifacts. We then fine-tune the model on the masked data. Intuitively, this way we encourage the model to blend all artifacts to a single contextualized representation that will never appear during testing, also avoiding to redistribute the informativeness of spurious lexical items to surrounding tokens. As for REMOVAL, we experiment with $\mathcal{S}_{\neg I}$ and $\mathcal{S}_I$ masking variants.

5.3 Results and discussion

In Table 4 we report the results for all baselines along the in-distribution and out-of-distribution dimensions from the lens of fairness and robustness.

[10] For fair comparison, we also provide results with less (50%) and more (25%) aggressive thresholds in Appendix C.1.

	In-distribution			*Out-of-distribution*							
				→		→		→		→	
		$F_{1\uparrow}$	$FPR_{\downarrow}$	$F_{1\uparrow}$	$FPR_{\downarrow}$	$F_{1\uparrow}$	$FPR_{\downarrow}$	$F_{1\uparrow}$	$FPR_{\downarrow}$	$F_{1\uparrow}$	$FPR_{\downarrow}$
VANILLA		$75.83_{0.3}$	$11.26_{0.9}$			$59.26_{0.7}$	$9.60_{1.3}$	$68.24_{0.4}$	$19.80_{1.4}$	$\mathbf{69.58}_{0.3}$	$16.36_{2.0}$
FILTERING		$72.79_{1.0}$	$14.57_{4.4}$			$58.95_{0.3}$	$12.05_{4.4}$	$65.57_{1.3}$	$19.19_{6.6}$	$67.68_{1.9}$	$19.67_{2.5}$
REMOVAL ($\mathcal{S}_{\neg I}$)		$74.96_{0.8}$	$10.39_{1.1}$			$\mathbf{59.28}_{0.3}$	$11.09_{0.5}$	$\mathbf{68.56}_{0.3}$	$19.29_{1.9}$	$67.65_{1.2}$	$15.32_{1.8}$
REMOVAL ($\mathcal{S}_I$)		$74.96_{0.8}$	$9.52_{0.9}$			$58.99_{0.4}$	$8.93_{1.7}$	$66.31_{0.6}$	$13.13_{1.1}$	$63.00_{1.5}$	$14.49_{3.6}$
MASKING ($\mathcal{S}_{\neg I}$)		$74.76_{1.1}$	$10.82_{0.7}$			$59.14_{0.2}$	$13.47_{2.5}$	$67.62_{0.6}$	$19.29_{2.3}$	$66.37_{0.9}$	$17.81_{3.6}$
MASKING ($\mathcal{S}_I$)		$\mathbf{76.41}_{0.6}$	$\mathbf{7.50}_{2.0}$			$58.83_{0.4}$	$\mathbf{8.26}_{2.3}$	$65.66_{0.8}$	$\mathbf{10.10}_{1.5}$	$63.48_{1.6}$	$\mathbf{6.63}_{1.9}$
VANILLA		$\mathbf{68.83}_{0.4}$	$11.01_{1.5}$	$60.61_{1.2}$	$29.87_{4.1}$			$65.95_{0.6}$	$41.72_{4.8}$	$\mathbf{67.68}_{0.6}$	$40.17_{5.3}$
FILTERING		$68.46_{0.3}$	$14.66_{1.1}$	$61.16_{0.2}$	$38.96_{1.5}$			$63.66_{0.7}$	$52.53_{2.1}$	$65.97_{0.7}$	$53.00_{1.4}$
REMOVAL ($\mathcal{S}_{\neg I}$)		$67.89_{1.0}$	$13.76_{1.4}$	$60.30_{1.9}$	$37.52_{4.3}$			$65.35_{0.6}$	$48.79_{5.6}$	$65.83_{1.4}$	$54.24_{6.2}$
REMOVAL ($\mathcal{S}_I$)		$67.65_{0.4}$	$6.99_{1.1}$	$58.77_{0.4}$	$17.46_{2.5}$			$65.71_{0.3}$	$27.98_{5.2}$	$66.95_{1.7}$	$21.74_{3.9}$
MASKING ($\mathcal{S}_{\neg I}$)		$68.50_{0.5}$	$9.60_{1.4}$	$\mathbf{61.17}_{1.2}$	$29.44_{2.3}$			$\mathbf{66.71}_{0.9}$	$40.61_{2.8}$	$67.11_{1.0}$	$36.85_{7.4}$
MASKING ($\mathcal{S}_I$)		$66.72_{0.7}$	$\mathbf{5.36}_{0.4}$	$57.72_{1.2}$	$\mathbf{12.55}_{2.0}$			$65.07_{0.6}$	$\mathbf{24.14}_{3.2}$	$63.13_{3.1}$	$\mathbf{11.59}_{3.1}$
VANILLA		$\mathbf{71.29}_{0.6}$	$31.41_{4.6}$	$64.51_{0.7}$	$29.15_{5.8}$	$61.12_{1.4}$	$10.04_{2.0}$			$67.66_{0.9}$	$37.68_{6.5}$
FILTERING		$71.13_{0.1}$	$27.47_{7.1}$	$64.31_{0.7}$	$23.67_{5.6}$	$61.09_{0.6}$	$9.90_{4.4}$			$68.15_{1.2}$	$31.88_{6.2}$
REMOVAL ($\mathcal{S}_{\neg I}$)		$71.04_{0.3}$	$30.00_{3.4}$	$64.19_{1.1}$	$27.71_{4.3}$	$61.58_{1.1}$	$10.49_{2.3}$			$68.44_{0.8}$	$34.99_{2.2}$
REMOVAL ($\mathcal{S}_I$)		$69.78_{0.5}$	$21.52_{5.4}$	$65.08_{0.4}$	$21.93_{6.3}$	$60.54_{1.3}$	$6.99_{1.4}$			$66.52_{1.4}$	$24.22_{7.8}$
MASKING ($\mathcal{S}_{\neg I}$)		$71.06_{0.4}$	$27.88_{4.7}$	$64.46_{0.7}$	$25.97_{2.8}$	$\mathbf{61.91}_{0.6}$	$8.63_{1.6}$			$\mathbf{68.86}_{0.6}$	$33.33_{4.7}$
MASKING ($\mathcal{S}_I$)		$69.72_{0.8}$	$\mathbf{13.64}_{0.9}$	$\mathbf{65.55}_{0.8}$	$\mathbf{15.01}_{2.2}$	$60.17_{1.5}$	$\mathbf{3.20}_{0.6}$			$66.64_{2.5}$	$\mathbf{13.04}_{1.9}$
VANILLA		$\mathbf{78.33}_{0.9}$	$15.73_{2.2}$	$60.20_{0.5}$	$17.89_{3.2}$	$58.22_{0.8}$	$5.58_{0.4}$	$64.76_{0.6}$	$25.56_{2.0}$		
FILTERING		$73.42_{3.1}$	$17.39_{1.6}$	$58.38_{1.1}$	$18.33_{1.3}$	$57.25_{1.6}$	$6.85_{1.8}$	$62.01_{0.8}$	$25.45_{3.2}$		
REMOVAL ($\mathcal{S}_{\neg I}$)		$76.77_{1.0}$	$17.81_{1.3}$	$\mathbf{61.32}_{1.6}$	$17.17_{1.5}$	$59.07_{2.0}$	$6.18_{0.8}$	$\mathbf{65.26}_{0.4}$	$24.75_{1.4}$		
REMOVAL ($\mathcal{S}_I$)		$75.62_{1.1}$	$15.32_{3.9}$	$58.58_{0.7}$	$20.06_{3.5}$	$58.95_{0.4}$	$7.29_{1.8}$	$61.96_{0.6}$	$22.12_{2.6}$		
MASKING ($\mathcal{S}_{\neg I}$)		$77.01_{1.0}$	$17.81_{0.7}$	$60.00_{0.7}$	$19.05_{4.2}$	$\mathbf{58.99}_{0.8}$	$6.18_{0.3}$	$64.44_{0.2}$	$27.27_{2.0}$		
MASKING ($\mathcal{S}_I$)		$76.39_{0.6}$	$\mathbf{9.94}_{1.6}$	$57.81_{1.2}$	$\mathbf{14.43}_{1.1}$	$57.33_{1.5}$	$\mathbf{4.32}_{0.6}$	$62.97_{0.4}$	$\mathbf{18.28}_{1.7}$		

Table 4: In-distribution and out-of-distribution results (F_1 for accuracy and FPR for fairness). Out-of-distribution results are on →: REDDIT, →: TWITTER, →: GAB, and →: STORMFRONT. Scores are averages of 3 runs with different seeds, whereas subscripts indicate standard deviation. ↑: greater the better; ↓: lower the better.

Since we argue that in-distribution performance is *not* a reliable measure for the performance of a hate speech detection system *in the wild*, due to space constraints we here focus on the more realistic yet more challenging out-of-distribution setup.

Filtering is not a one-size-fits-all solution Despite the improvements in OOD generalization on commonsense reasoning, question answering, and NLI tasks (Swayamdipta et al., 2020), training on ambiguous instances collected from training dynamics is not as effective in hate speech detection.[11] Instead, our results show that FILTERING leads to mixed results for OOD fairness compared to the VANILLA baseline. This is consistent with results on Twitter data (Zhou et al., 2021), and we further confirm it is the case also across platforms. Importantly, we also notice that FILTERING has a detrimental effect on OOD robustness, except for two cases only (i.e., → and →). This indicates that hate speech detection is a nuanced task requiring more targeted approaches than automated data filtering.

[11] We notice this holds true also when employing less/more aggressive filtering thresholds, as shown in Appendix C.1.

Removing $\mathcal{S}_I$ is not as strong as it has been previously thought Removing identity terms from data altogether is a commonly used baseline for testing downstream fairness (e.g., Kennedy et al., 2020b). Indeed, our results confirm that REMOVAL($\mathcal{S}_I$) consistently reduces the FPR on test instances containing $\mathcal{S}_I$ mentions compared to the VANILLA baseline – with the only exception of → . However, it only improves OOD robustness on → and → . Moreover, it consistently scores lower than MASKING($\mathcal{S}_I$) on fairness, as discussed below. This raises the question of whether REMOVAL($\mathcal{S}_I$) should continue to be used as fairness baseline in future studies.

Masking $\mathcal{S}_I$ improves fairness When masking $\mathcal{S}_I$, we notice a consistent improvement in fairness over *all* approaches, both in-distribution and out-of-distribution, on *all* platforms. Reduction in FPR over the VANILLA baseline is as large as 3×, as results for {; } → and → show. Most of the remaining train-test pairs show a 2× improvement in FPR, also compared to the common REMOVAL($\mathcal{S}_I$) baseline. We hypothesize the improved fairness performance with respect to RE-

MOVAL($\mathcal{S}_I$) is due to the way contextualized representations are formed during training, as discussed in Section 5.2. Despite being surprisingly simple, we envision MASKING($\mathcal{S}_I$) as a strong baseline for future work on fairness in hate speech detection.

$\mathcal{S}_{\neg I}$ artifacts are not as useful as $\mathcal{S}_I$ We observe that methods exploiting $\mathcal{S}_{\neg I}$ artifacts lead to mixed results. This suggests that while a substantial FPR reduction can be achieved exploiting $\mathcal{S}_I$ artifacts, robustness calls for more complex debiasing strategies to transfer well across distributions.

Fairness comes at the cost of robustness Overall, we observe an important trade-off between fairness and robustness. Data-centric approaches that achieve a consistently high level of fairness – namely, MASKING($\mathcal{S}_I$) and REMOVAL($\mathcal{S}_I$) – typically show a decrease in in-distribution and out-of-distribution performance – with the exception of → and → for MASKING($\mathcal{S}_I$), and → and → for REMOVAL($\mathcal{S}_I$). On one hand, this suggests that spurious, identity-related lexical artifacts do play an important role in performance *across* distributions. On the other hand, we believe this reflects the real performance of a prototypical model that is substantially fairer, and thus to which future work in hate speech detection should be compared to. We argue that MASKING($\mathcal{S}_I$) represents a starting point to achieve both fairness and OOD robustness, the latter requiring more complex, model-centric debiasing approaches. A summary of the results over all corpus pairs for each method is presented in Appendix C.2.

6 Towards Artifacts Documentation

The practice of *data statements* (Bender and Friedman, 2018) has been recently adopted by the NLP community as a way to include relevant information about the creators, the methodology and possible biases when a dataset is released. This should in turn have a positive impact on systems trained on such data, contributing to a better evaluation of models' generalization and fairness. We propose that an **artifacts statement** should be added to this documentation as a way to contribute to diagnosis (and thus mitigation) of pre-existing bias, which is also one of the goals of data statements.

In particular, we propose a template for lexical artifacts documentation and publicly release code to easily compute ranked correlations between tokens and target classes of interest for a given annotated corpus. To ensure the process of documenting lexical artifacts will be as smooth as possible – and thus allows widespread adoption of artifacts statement in the future – our code automatically generates outputs in different formats, from raw text to LaTeX code for seamless inclusion in publications.

We present the artifacts statement template below, and provide a full example in Appendix D.

I) TOP LEXICAL ARTIFACTS. Which are the k most informative tokens in the corpus for the class(es) of interest? This can be a ranked list of ($k \geq 10$) tokens in plain text or in a tabular format, optionally along with associated scores. If there are multiple classes of interest, top k lexical artifacts for each class should be included.

II) CLASS DEFINITIONS. Different definitions for the same class may exist across datasets. This impacts the annotation, which in turn has an effect on resulting lexical artifacts. An explicit definition of the target class(es) for which the top lexical artifacts are computed should be provided here.

III) METHODS AND RESOURCES. The method used to compute the correlation between tokens and class(es) (e.g., PMI, interpretability approaches) in the annotated corpus should be reported here, possibly with a link to code. If preprocessing and deduplication have been performed, they should be clearly reported. Resources such as full lists of lexical artifacts can be additionally included.

7 Related Work

The problem of models' generalizability related to hate speech detection has been extensively discussed in recent works (Vidgen and Derczynski, 2020; Yin and Zubiaga, 2021; Wich et al., 2021). Indeed, it has been shown that state-of-the-art performance on this task overestimates the capability of models to yield the same results over time (Florio et al., 2020) or across different domains (Wiegand et al., 2019). Possible mitigation strategies include domain adaptation techniques (Ramponi and Plank, 2020), augmenting smaller datasets with a larger dataset from a different domain (Karan and Šnajder, 2018), the use of a domain lexicon to transfer knowledge across domains (Pamungkas and Patti, 2019) and the fine-tuning of HateBERT (Caselli et al., 2021) on the target corpus (Bose et al., 2021), among others.

Concerning bias and fairness, several works have pointed out the presence of bias in hate and abusive language datasets (Wiegand et al., 2019; Sap

et al., 2019, 2020). This issue has been addressed in different ways, including functional tests for hate speech detection models (Röttger et al., 2021; Manerba and Tonelli, 2021) and post-hoc explanations to measure models' bias towards identity terms (Kennedy et al., 2020b). As regards bias mitigation, the task has been addressed through a number of approaches, e.g., via adversarial feature learning (Vaidya et al., 2020), by using debiased word embeddings and gender swap data augmentation (Park et al., 2018) or by adding non-toxic examples to better balance the data (Dixon et al., 2018). The work probably most related to ours is Zhou et al. (2021), which presents an analysis of lexical and dialectal biases in the dataset by Founta et al. (2018). The authors propose lexical bias categories which we extend in this work (see Section 2). However, they focus only on one dataset and on in-domain bias reduction. Moreover, they start from a list of "bad words", whereas we compute it from data. To our knowledge, this is the first work advocating for a joint view on fairness and robustness, both identified as critical aspects related to the classification of hate speech (Wich et al., 2021).

8 Limitations

Our work is a step forward towards a better understanding of the bias that can be encoded in hate speech detection corpora (Blodgett et al., 2020). However, we are aware of some limitations. First, all findings in this work are related to hate speech datasets in English. With the increasing availability of hate speech data in languages other than English, we aim to investigate our methods on other languages too. Second, annotated data from multiple platforms may not be available for some languages, and this can limit the cross-distribution computation of artifacts. Lastly, we acknowledge spurious statistical correlations may go beyond the token level. We believe our study is a first step towards contextual debiasing from spurious lexical artifacts, and thus can be of inspiration for future studies.

9 Conclusion and Future Directions

This paper investigates the impact of lexical artifacts on out-of-distribution fairness and robustness in hate speech detection, raising awareness on the interplay between the two dimensions that should be studied together in future work. We propose a fine-grained categorization of lexical artifacts and simple yet effective data-centric baselines, showing that while robustness calls for model-centric approaches, masking spurious identity artifacts is a viable approach that we argue should be used as strong baseline for fairness assessment in future research. In future work we aim to investigate the role of dialectal biases and non-lexical artifacts, extending the study on languages other than English. We release all baseline models, resources, and the code to compute lexical artifacts, broadly suggesting the inclusion of "artifacts statement" as a way to document potential lexical biases when a dataset is released, to provide a complementary view to data statements (Bender and Friedman, 2018).

Ethical Considerations

The annotation task described in Section 4.3 was carried out by two researchers regularly employed at Fondazione Bruno Kessler as part of their work.

Overall, we do not foresee any specific ethical concern related to this work. On the contrary, our goal is to propose artifacts statement as a desirable practice for documenting potential biases in newly released datasets, and improve current debiasing methods by distinguishing among different types of lexical artifacts. However, the (finite set of) identity-related and offensive tokens considered in this work are all in English and centered around Western cultural context. We leave the evaluation of our methodology to assess whether there are language- or more broadly culture-dependent changes for future work, following recent work on biases in geo-cultural contexts (Ghosh et al., 2021).

Acknowledgements

Part of this work was funded by the PROTECTOR European project (ISFP-2020-AG-PROTECT-101034216-PROTECTOR). This research was also supported by the KID ACTIONS REC-AG project (n. 101005518) on "Kick-off preventIng and responDing to children and AdolesCenT cyberbullyIng through innovative mOnitoring and educatioNal technologieS".

References

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

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.

Tulika Bose, Irina Illina, and Dominique Fohr. 2021. Unsupervised domain adaptation in cross-corpora abusive language detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 113–122, Online. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

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.

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*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Robert M Fano. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29:793–794.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12).

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2021. *Speech and Language Processing*, 3rd edition. Prentice Hall.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Brendan Kennedy, Mohammad Atari, Aida M. Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Gabriel Cardenas, Alyzeh Hussain, Austin Lara, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and

Morteza Dehghani. 2020a. The Gab hate corpus: A collection of 27k posts annotated for hate speech.

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

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-grained fairness analysis of abusive language detection systems with CheckList. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55:477–523.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Se Rim Park. 2020. HABERTOR: An efficient and effective deep hatespeech detector. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7486–7502, Online. Association for Computational Linguistics.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing

identity bias in toxic comment detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):683–693.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one*, 15(12):e0243300.

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, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

M. Wich, T. Eder, H. Al Kuwatly, and G. Groh. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *AI Ethics*, 19:1–23.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Appendix

A Data: Additional Details

A.1 Preprocessing and anonymization

We preprocess texts across platforms in a consistent way by anonymizing user mentions, URLs, and email addresses with `[USER]`, `[URL]`, and `[EMAIL]` placeholders, respectively. We segment hashtags into constituent words using the `wordsegment` package,[12] remove newlines, unescape HTML tags, and lowercase the texts.

A.2 Deduplication

We found many duplicates in the data for all platforms. We argue that retaining duplicates as done in most previous work could severely affect the reliability of any bias analysis (and debiasing method) and its subsequent conclusions. Specifically, duplicates can (i) skew the distribution of actual artifacts in the data, overamplifying some lexical items and demoting others, and (ii) result in unfair evaluations due to identical examples falling in multiple instances of the training, development and test sets, also potentially leading to overfitting.[13]

Following this intuition, we thus remove duplicate instances after preprocessing.[14] Specifically, we removed 485 duplicate instances from Vidgen et al. (2021), 10,911 duplicate instances from Founta et al. (2018), 521 from Kennedy et al. (2020a), and 255 from de Gibert et al. (2018). Moreover, for the purpose of this work we remove duplicates whose single instances exhibit opposing labels, leaving the exploration and exploitation of annotator disagreement for future work.

B List of Spurious Artifacts

In the following, we provide the list of all spurious lexical artifacts annotated as $\mathcal{S}_I$ and $\mathcal{S}_{\neg I}$ as described in Section 4.3. All these are the ones that exhibit full agreement. In our shared repository we also release all artifacts that exhibit disagreement even after adjudication, in order to encourage future work on this direction.

Identity-related ($\mathcal{S}_I$) "white", "black", "jews", "women", "jew", "whites", "blacks", "muslim", "gay", "muslims", "islam", "woman", "jewish", "islamic", "immigrants", "mexican", "asian", "homosexual", "americans", "lesbian", "homo", "females", "america", "brown", "israel", "arabs", "zionist", "trans", "lgbt", "girl", "hispanic", "refugees", "male", "african", "africa", "girls", "indians", "queer", "##grate", "guy".

Non identity-related ($\mathcal{S}_{\neg I}$) "##s", "##es", "people", "country", "##ing", "anti", "illegal", "bunch", "##t", "kids", "culture", "brain", "##ly", "##bt", "##d", "sex", "ho", "##nt", "countries", "##ic", "##ers", "liberal", "reason", "##y", "human", "genocide", "##ed", "##ists", "wrong", "lives", "bad", "god", "##oc", "lying", "##ard", "racism", "##e", "##oid", "##w", "yeah", "millions", "society", "##g", "leftist", "crime", "sp", "des", "##ist", "##ry", "mouth", "##ards", "##rs", "##ize", "burn", "murdered", "worship", "##ening", "##ism", "living", "##fa", "coming", "calling", "streets", "##ting", "force", "mis", "##ss", "blame", "typical", "##pe", "baby", "death", "talking", "##gen", "belong", "respect", "di", "##yp", "sexual", "##less", "mad", "war".

C Experiments: Additional Results

C.1 Filtering with different thresholds

In Table 5 we present results for the FILTERING baseline using different sampling thresholds. Specifically, in addition to using the 33% (1/3) most ambiguous training data instances as in Swayamdipta et al. (2020), we provide full results using more aggressive (i.e., 25%, 1/4) and less aggressive (i.e., 50%, 1/2) filtering thresholds. We notice mixed results that make hard to determine which is the best threshold across platforms. FILTERING (25%) improves OOD robustness on 🐸 → 👽, and FILTERING (50%) provides best overall in-domain performance on 🐸. However, MASKING($\mathcal{S}_I$) outperforms all FILTERING approaches according to the FPR metric.

C.2 Average results over all corpus pairs

We provide a summary of the results for all methods in Table 6, where we report average scores over all corpus pairs (refer to Table 4 for full results). On average, MASKING($\mathcal{S}_I$) improvement in FPR over the VANILLA baseline is as large as 2×, both in-distribution and out-of-distribution. This comes at

[12] https://github.com/grantjenks/python-wordsegment

[13] Among duplicates, we found some tweets with more than 100 duplicate instances in Founta et al. (2018).

[14] Most work do not explicitly mention if deduplication is carried *before* or *after* preprocessing texts. We believe this is an important detail to foster reproducibility – we found many examples with the same text but different URLs, unveiling possibly bot-generated messages we removed this way.

	In-distribution		*Out-of-distribution*							
			→		→		→		→	
		$F_{1\uparrow}$	$FPR_{\downarrow}$	$F_{1\uparrow}$	$FPR_{\downarrow}$	$F_{1\uparrow}$	$FPR_{\downarrow}$	$F_{1\uparrow}$	$FPR_{\downarrow}$	$F_{1\uparrow}$	$FPR_{\downarrow}$
FILTERING (25%)		$72.35_{0.7}$	$13.42_{1.9}$			$58.25_{0.6}$	$11.98_{2.8}$	$65.14_{0.9}$	$18.18_{3.2}$	$67.84_{2.3}$	$20.08_{9.9}$
FILTERING (33%)		$72.79_{1.0}$	$14.57_{4.4}$			$58.95_{0.3}$	$12.05_{4.4}$	$65.57_{1.3}$	$19.19_{6.6}$	$67.68_{1.9}$	$19.67_{2.5}$
FILTERING (50%)		$74.87_{0.9}$	$11.26_{2.6}$			$59.20_{0.4}$	$9.82_{2.5}$	$66.34_{0.3}$	$19.39_{2.8}$	$69.54_{2.4}$	$20.50_{4.5}$
FILTERING (25%)		$68.23_{0.6}$	$15.77_{2.3}$	$60.73_{0.3}$	$40.98_{1.0}$			$62.89_{0.7}$	$53.64_{2.6}$	$65.66_{0.6}$	$59.83_{6.6}$
FILTERING (33%)		$68.46_{0.3}$	$14.66_{1.1}$	$61.16_{0.2}$	$38.96_{1.5}$			$63.66_{0.7}$	$52.53_{2.1}$	$65.97_{0.7}$	$53.00_{1.4}$
FILTERING (50%)		$68.77_{0.5}$	$12.05_{1.6}$	$61.11_{0.4}$	$35.64_{2.5}$			$65.31_{1.2}$	$46.87_{3.8}$	$67.12_{0.5}$	$48.03_{5.6}$
FILTERING (25%)		$71.16_{0.6}$	$30.51_{2.6}$	$65.92_{0.1}$	$26.70_{1.6}$	$61.60_{0.8}$	$11.31_{1.6}$			$68.70_{0.8}$	$34.58_{2.9}$
FILTERING (33%)		$71.13_{0.1}$	$27.47_{7.1}$	$64.31_{0.7}$	$23.67_{5.6}$	$61.09_{0.6}$	$9.90_{4.4}$			$68.15_{1.2}$	$31.88_{6.2}$
FILTERING (50%)		$71.48_{1.0}$	$27.47_{5.4}$	$64.11_{1.1}$	$24.39_{6.1}$	$61.81_{0.5}$	$8.78_{2.5}$			$67.91_{0.9}$	$29.19_{8.8}$
FILTERING (25%)		$72.94_{2.2}$	$14.70_{2.2}$	$59.04_{0.7}$	$14.57_{1.6}$	$57.12_{2.1}$	$6.77_{1.4}$	$62.43_{0.2}$	$22.73_{4.2}$		
FILTERING (33%)		$73.42_{3.1}$	$17.39_{1.6}$	$58.38_{1.1}$	$18.33_{1.3}$	$57.25_{1.6}$	$6.85_{1.8}$	$62.01_{0.8}$	$25.45_{3.2}$		
FILTERING (50%)		$76.42_{1.1}$	$13.87_{1.4}$	$59.50_{1.0}$	$14.86_{2.9}$	$58.01_{1.3}$	$5.21_{0.3}$	$63.42_{0.4}$	$22.53_{2.0}$		

Table 5: Additional results for the FILTERING baseline using different sampling thresholds (25%, 33%, 50%).

	In-distr.		*Out-of-distr.*	
	$F_{1\uparrow}$	$FPR_{\downarrow}$	$F_{1\uparrow}$	$FPR_{\downarrow}$
VANILLA	**73.57**	17.35	**63.98**	23.62
FILTERING	71.45	18.52	62.85	25.96
REMOVAL ($\mathcal{S}_{\neg I}$)	72.67	17.99	63.90	25.63
REMOVAL ($\mathcal{S}_I$)	72.00	13.34	62.61	17.20
MASKING ($\mathcal{S}_{\neg I}$)	72.83	16.53	63.90	23.16
MASKING ($\mathcal{S}_I$)	72.31	**9.11**	62.03	**11.80**

Table 6: Average in-distribution and OOD results over all corpus pairs for each method.

the cost of a minimal in-distribution and OOD drop in macro F_1 (i.e., -1.26 and -1.95, respectively).

D Lexical Artifacts Statement Example

An example of lexical artifacts statement for the Reddit dataset (Vidgen et al., 2021) used in this study is presented in the following.

I) TOP LEXICAL ARTIFACTS. We present the top $k = 10$ most informative tokens for the *hateful* class along with their scores in Table 7.

Rank	Token	Score	*Rank*	Token	Score
1	##tar	1.00	6	##gg	0.80
2	##ded	0.91	7	gay	0.79
3	##s	0.86	8	women	0.76
4	fa	0.85	9	##ds	0.74
5	b*tch	0.83	10	f*cking	0.74

Table 7: Top 10 most informative tokens for the hateful class on the Reddit dataset according to PMI.

II) CLASS DEFINITIONS. The *hateful* class is represented by originally *identity-directed* labeled examples in CAD (Vidgen et al., 2021), and is defined as "*Content which contains a negative statement made against an identity. An 'identity' is a social category that relates to a fundamental aspect of individuals' community, socio-demographics, position or self-representation [...]. It includes but is not limited to Religion, Race, Ethnicity, Gender, Sexuality, Nationality, Disability/Ableness and Class.*" (Vidgen et al., 2021).

III) METHODS AND RESOURCES. In order to compute the correlation between tokens to the *hateful* class we employ PMI as implemented in *[this work]* (code: `https://github.com/dhfbk/hate-speech-artifacts`). Input texts have been preprocessed by anonymizing user mentions, URLs, and email addresses with `[USER]`, `[URL]`, and `[EMAIL]` placeholders. Hashtags have been segmented using `wordsegment`,[15] and we remove newlines, unescape HTML tags, and lowercase texts. Duplicate instances have been removed after preprocessing.

The full list of lexical artifacts along with associated scores is available at `https://github.com/dhfbk/hate-speech-artifacts`.

[15] `https://github.com/grantjenks/python-wordsegment`