

Is My Model Biased? Exploring Unintended Bias in Misogyny Detection Tasks

Daniela Godoy¹, Antonela Tommasel¹

¹ISISTAN Research Institute (CONICET/UNCPBA), Tandil, Bs. As., Argentina

Abstract

Although hate speech detection has been extensively tackled in the literature as a classification task, recent works have raised concerns about the robustness of such systems. Understanding hate speech remains a significant challenge for creating reliable datasets and automatizing its detection. An essential goal for detection techniques is to ensure that they are not unduly biased towards or against particular norms of offense. For example, ensuring that models are not reproducing common biases in society associating certain terms with hateful content. This situation is known as unintended bias, in which models learn usual associations between words (commonly called identity terms) which causes them to classify content as hateful just because it contains one identity word. In this work, we tackle the issue of measuring and explaining the sensitivity of models to the presence of identity terms during model training. To this end, focusing on a misogyny detection task, we study how models behave in the presence of the identified terms, and whether they contribute to biasing the performance of trained models.

Keywords

Bias, misogyny detection, hate speech

1. Introduction

Social media is being increasingly used to spread hateful content, and, at the same time, its real-life consequences also grow. Hate speech mirrors not only existing opinions but also induces new negative feelings towards its targets [1]. To counteract the massive scale to which hate speech is occurring, there is an urgent need for effective counter-measures. From a natural language processing perspective, hate speech detection can be seen as a classification task, in which, given a text, it is determined whether it contains hate speech.

Although hate speech detection has been extensively tackled in the literature as a supervised learning task, recent works have raised concerns about the robustness of such systems [2]. For training a classifier, a large volume of data is required. This data is usually obtained by manually annotating a set of texts. Thereby, the reliability of human annotations is essential. Meanwhile, researchers have questioned the ability to let big data speak for itself as its representativeness, spatiotemporal extent, and uneven demographic information can make it subjective [3].

In this sense, hate speech detection (and related) tasks are especially challenging because the concept of hate or toxicity depends on the social context of the example, including the identity


AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Montreal, CA

✉ daniela.godoy@isistan.unicen.edu.ar (D. Godoy); antonela.tommasel@isistan.unicen.edu.ar (A. Tommasel)

ORCID 0000-0002-5185-4570 (D. Godoy); 0000-0001-6091-8305 (A. Tommasel)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

of the speaker [3]. Hate speech models could also capture and reproduce common biases in society. For example, when detecting misogynist content, “innocent” terms (also known as *identity terms* or *bias sensitive words*, such as “woman”) are frequently (mistakenly) associated with the misogynist class [4]. This situation often stems from the selection and skewed sampling process used for collecting the training data. In general, training data contains more hateful examples using such terms than non-hateful ones, which can induce models to solely associate such words with hateful content [5]. This situation is known as *unintended bias* by which a model performs differently across the identified groups (e.g., demographic groups, genders), thus leading to low-quality results and challenging the fairness of the technique [6].

While there have been efforts on identifying and characterizing bias in social data [7], and on identifying the identity terms on the (perhaps) biased training data, the removal or mitigation of bias has received comparatively less attention. Particularly, most works do not question the implications of other decisions made during model selection and training [8]. For this reason, in this work, we tackle the issue of measuring and explaining the sensitivity of models to the presence of identity terms during model training. To this end, focusing on a misogyny detection task and starting from existing techniques for identifying identity terms [9], we study how models behave in the presence of the identified terms and whether they contribute to biasing the performance of trained models.

The rest of this paper is organized as follows. Section 2 describes related works. Section 3 describes the methodology and the results of the performed study. Section 4 presents the conclusions and outlines future work.

2. Related works

Hate speech (and another abusive content) has received increasing attention over the last few years for its profound effects to society [10]. For example, the propagation of hate speech risks harming its targets, polluting the public discourse, escalating acts of violence, discrimination and even fostering social tensions. In this context, the real-time and accurate detection of online abuse is crucial [10]. Among the different particular types of hate speech lies misogyny. Misogyny can be defined as the hate of prejudice against women, that can linguistically manifest as social exclusion, discrimination, threats and sexual objectification [11].

During the last few years, and fostered by the organized shared tasks at multiple conferences [11, 12, 13], different approaches have been proposed for tackling misogyny detection. In general, research leverages diverse textual features, ranging from lexical and syntactic features to semantic features derived from word embeddings in combination with both traditional and deep learning classifiers [12, 4]. The shared tasks have not only addressed the problem in English, but also on Spanish and Italian, and even multilingual approaches have been proposed [14].

As previously mentioned, addressing bias is crucial, not only due to its potential impact in real-world applications, but also to improve the robustness of techniques in a cross-dataset (or even cross-domain) scenario [4]. In this regard, a few works have started to explore the fairness of techniques [4, 9, 15]. For example, the EVALITA 2020 shared task [12] was concerned with producing fairer classifications and not explicitly assessing the causes of bias and how it affected the trained model. Nozza *et al.* [4] proposed a set of model agnostic metrics to

Table 1
Summary of datasets

	Train		Test	
	Pos	Neg	Pos	Neg
EVALITA-2018	1785	2215	460	540
TRAC-2	309	3954	175	1025
Urban Dictionary	724	875	310	376

assess unintended bias based on the classifiers’ score distribution across the protected groups. Vidgen and Derczynski [9] proposed metrics for detecting the words inducing bias based on term distribution in the training set and on how the trained model classified such word. Then, multiple replacement strategies were proposed for transforming the training data and mitigating the effect of the biased words. Finally, Nozza *et al.* [4] manually identified patterns of identity terms which were used as an unbiased test set. In this case, the mitigation strategy was based on extending the training data with texts including the biased terms but associated with the minority class.

Our research aims to take a step further in the analysis of biased sensitive terms by analyzing the importance and impact on the models of identified bias sensitive words. First, the occurrence of potential identity terms in misogyny-oriented datasets is automatically calculated based on existent metrics. Second, the importance and impact of these terms in trained classification models is assessed through explanation techniques. Ultimately, the effects of removing these terms during model learning are quantified and analyzed.

3. Study description

3.1. Data description

In this study, we explored three available datasets for misogyny detection. The first dataset included in the analysis was made available for the Automatic Misogyny Identification (AMI) shared task at EVALITA 2018¹. AMI@EVALITA 2018 dataset [16] consist of 5000 annotated tweets aiming at discriminating misogynistic posts from non-misogynistic ones. The second dataset comes from the Misogynistic Aggression Identification shared task at TRAC-2², which includes posts from both social media and other popular streaming and sharing platforms [17]. Aggressive posts were annotated as Non-gendered or Non-Misogynous and Gendered or Misogynous. Finally, the Urban Dictionary dataset³ [18] comprises definitions gathered from the Urban Dictionary platform⁴ and annotated as misogynistic or non-misogynistic. Table 1 summarizes the statistics of the selected datasets according to the training and test partitions. In the case of the Urban Dictionary dataset, partitions were randomly generated with a stratified strategy.

¹<https://amievalita2018.wordpress.com/>

²<https://sites.google.com/view/trac2/shared-task>

³<https://data.mendeley.com/datasets/3jfwskryy/3>

⁴<https://www.urbandictionary.com/>

Table 2

Bias sensitive words according to SOAC values

EVALITA-2018			TRAC-2			Urban Dictionary		
word	%pos	%neg	word	%pos	%neg	word	%pos	%neg
bitch	76.32	17.58	homosexu	61.03	6.49	femal	76.11	9.52
hoe	74.91	19.54	armi	52.94	13.23	vagina	92.41	0.00
suck	80.48	17.07	bitch	60.41	20.83	pussi	91.01	2.24
shut	88.13	11.86	bastard	81.25	12.50	dick	73.91	14.49
lil	77.27	20.45	randi	80.00	13.33	fat	67.92	15.09

Table 3

Bias sensitive words according to SPCPD values

EVALITA-2018		TRAC-2		Urban Dictionary	
word	$p(c w)$	word	$p(c w)$	word	$p(c w)$
bitch	1.00	bitch	1.00	femal	1.00
whore	1.00	randi	0.98	pussi	1.00
femal	0.92	madarchod	0.87	vagina	1.00
vagina	0.91	homosexu	0.87	woman	1.00
ladi	0.91	kutiya	0.77	lesbian	0.99

3.2. Bias sensitive words

According to Dixon *et al.* [6], a model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others. In this situation, a word w is defined as a bias sensitive word for a classifier if the classifier is unreasonably biased with respect to w to a very high degree.

For identifying identity terms, Dixon *et al.* [6] used a dictionary of hand-curated biased words, whereas Nozza *et al.* [4] defined templates denoting common patterns in which these words tend to appear. In this work, we searched for stereotypical words using the two metrics defined by Badjatiya *et al.* [9].

The first metric, Skewed Occurrence Across Classes (SOAC), indicates how frequent a word appears in the training set for a particular class (e.g., misogynistic). The classifier is assumed to learn to classify any text containing the high frequency terms into its predominant class, solely based on the presence of such words. Table 2 shows the top-5 words according to SOAC for the three analyzed datasets, sorted by word document frequency ($df(w)$), and then by their frequency in the positive class. In this context, words having a stronger presence in the positive class, while being underrepresented in the negative class could be a source of bias.

The second metric, also proposed by Badjatiya *et al.* [9], is the Skewed Predicted Class Probability Distribution (SPCPD). This metric allows determining whether the classifier has stereotyped a word as belonging to a certain class. Let $p(c|w)$ denote the classifier prediction probability of assigning a sentence containing only word w to class c and c_ϕ to the remaining classes. Then, the SPCPD score is defined as the maximum probability of w to belong to any class but c_ϕ . A word w is deemed as a bias sensitive word if $SPCPD(w) \geq \tau$, where τ is a pre-defined threshold. For binary classification τ is usually set close to 0.5.

Table 3 shows the top-5 sensitive words for the three datasets according to SPCPD scores ranked by the probability of being assigned to the relevant class (i.e., misogyny). Scores were computed based on a Random Forest classifier. When considering a 0.5 threshold, a total of 32 (2.84%), 27 (1.55%) and 24 (1.38%) bias words were identified for EVALITA-2018, TRAC-2 and Urban Dictionary, respectively.

As it can be observed, the words selected by both metrics differ. In the case of EVALITA-2018 only one word was identified by both metrics, while for TRAC-2 and Urban Dictionary 3 words were identified by both. Despite the TREC-2 dataset being in English, several of the identified words are in Hindi. These differences imply that having a high presence in the misogyny class (as SOAC evaluates) during training does not necessarily translates to a biased classification. Similarly, a high probability of being identified as misogynist does not strictly correlates with a high presence in the misogyny class during training.

The two analyzed metrics are concerned with either the inputs or outputs of the trained model, but disregard the analysis of how the presence of such terms affects the internal state of the classifier. In consequence, it cannot be directly concluded that the sole presence of these words induces bias. It is necessary to study the impact of the identified words to effectively determine the presence of bias and whether bias mitigation techniques are needed.

3.3. Model learning

To study whether the identified terms induce bias in the trained models, we first trained several classical classification algorithms for each of the three selected datasets. Table 4 summarises the performance of the trained models. Implementation was based on the Natural Language Toolkit (NLTK)⁵ and scikit-learn⁶ libraries. As the focus of this work is exploring the existence of feature bias, and not misogyny detection, all models were trained using the default algorithm configurations. Performance was evaluated based on macro-averaged F_1 measure, AUC and False Positive Rates (FPR). As the classifier is assumed to incorrectly classify texts containing the identified terms into the misogyny class (i.e., the positive class), unintended bias can manifest through high FPR. In this sense, beyond the global perspective given by F_1 , AUC scores and false positive rates allow to have a better understanding of classification errors.

Results showed that the best scores were obtained by the Multi-layer Perceptron, Linear SVM Classification and Random Forest classifiers, while Naïve Bayes achieved the lowest results in all cases. The lowest FPR were observed for TRAC-2 and Urban Dictionary. The highest performance was observed for Urban Dictionary. These differences could be related to the different nature of the collected datasets. While EVALITA-2018 and TRAC-2 were collected from naturally occurring social media posts, Urban Dictionary presents definitions specifically provided by users, in some cases with a clear misogynist intention. Based on these results, it seems that identifying the potential causes of unintended bias is crucial for the adequate selection (and implementation) of bias mitigation strategies.

Table 4
Summary of classification results

	EVALITA-2018			TRAC-2			Urban Dictionary		
	F1	AUC	FPR	F1	AUC	FPR	F1	AUC	FPR
Linear SVM	60.26	60.25	35.37	<u>65.80</u>	<u>63.23</u>	4.39	87.69	87.46	7.98
Random Forest	61.37	61.34	32.96	63.55	61.23	4.39	<u>89.82</u>	<u>89.72</u>	7.98
Multi-layer Perceptron	<u>62.40</u>	<u>61.51</u>	<u>27.40</u>	46.63	50.24	<u>0.10</u>	87.85	87.62	7.98
Naive Bayes	53.60	55.51	68.33	34.78	32.54	46.34	65.00	65.45	39.10
Logistic Regression	61.38	61.43	28.88	55.37	54.79	1.85	87.43	87.02	<u>5.32</u>

3.4. Model explanation

Explainability techniques allow to associate feature values of an instance to the model prediction in a way it can be understood in human terms. In this regard, once models are trained, such techniques can be used to assess the role of the potential bias sensitive words in the performed misogyny predictions.

Model analysis was based on SHAP (SHapley Additive exPlanations) [19]. This method explains individual predictions by measuring the impact of variables in relation to their interactions with others. The SHAP explanation method, originated in coalitional game theory, and computes Shapley values [20] as the average marginal contribution of a feature value across all possible coalitions. The idea behind SHAP is that features with large absolute Shapley values significantly contribute to predictions. For each feature, its individual importance per prediction is computed, and then combined to obtain the global explanations, which allows to explain the entire model.

Figure 1 shows the top-5 features ranked according to their importance assessed by SHAP values for the three misogyny datasets. Models were trained using a Random Forest classifier so that explanation could be performed with TreeSHAP, an efficient estimation approach for tree-based models [21]. When comparing the top bias sensitive words detected based on the SOAC metric and the ones identified with SHAP, it can be observed that for the three datasets not every SOAC term greatly contributed to predictions. Hence, this could imply that the assumptions on which SOAC is based are not enough for inferring the effect that features will have on a model when interacting with others, and thus they capabilities for inducing bias.

As observed when comparing SOAC and SPCPD, there are differences between the terms identified by SHAP and the ones identified by the other metrics in the top-5 rankings. In the case of EVALITA-2018 only the word “bitch” was identified for the three analyzed metrics, while for TRAC-2 both “bitch” and “homosexu” were identified by the three metrics. On the other hand, for Urban Dictionary four words were identified by SPCPD and SHAP, and three by SOAC and SHAP. These observations highlight that only observing the input and output of a trained model is not enough for inferring whether such words would affect (or induce bias) in the trained model, as the internal learning mechanisms of algorithms may not be affected by features previously considered as bias sensitive.

Besides feature importance, we can summarize the effect of the identified features on the

⁵<https://www.nltk.org/>

⁶<https://scikit-learn.org/>

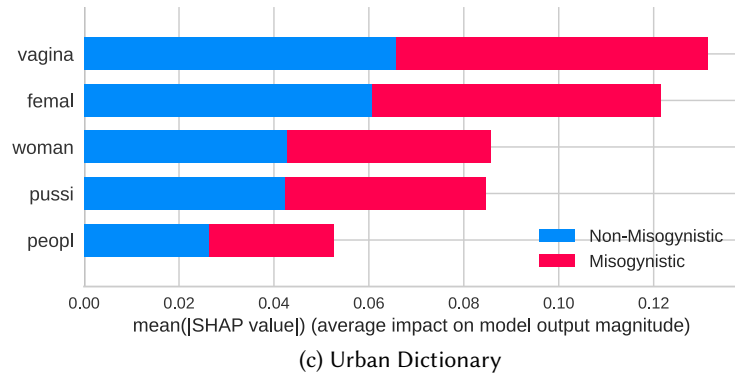
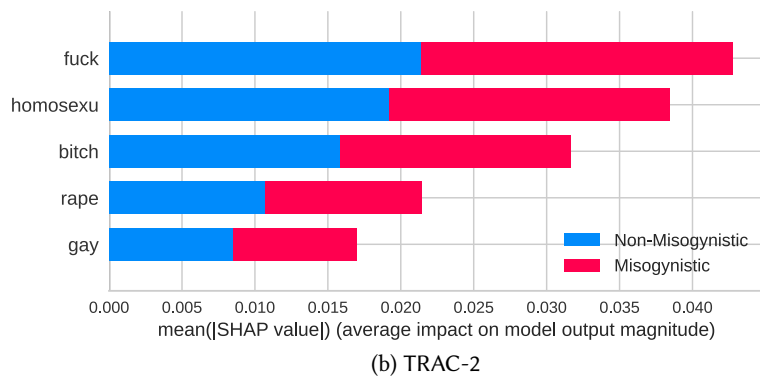
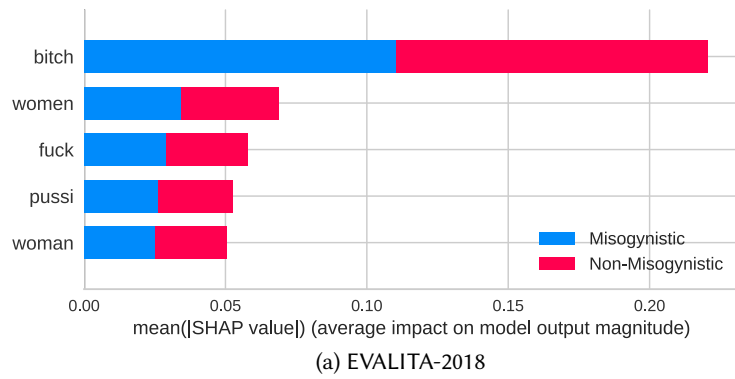


Figure 1: Features importance based on SHAP values

model by plotting the SHAP value for each feature and instance (Figure 2). Starting from the baseline (i.e., the mean feature importance), each feature (and its corresponding value) can be seen as a force that shifts the prediction towards the positive (misogynistic) class or the negative (non-misogynistic) class. The color represents the value of the feature within an instance from low to high. In this case, as features are weighted by their frequency, the minimum value

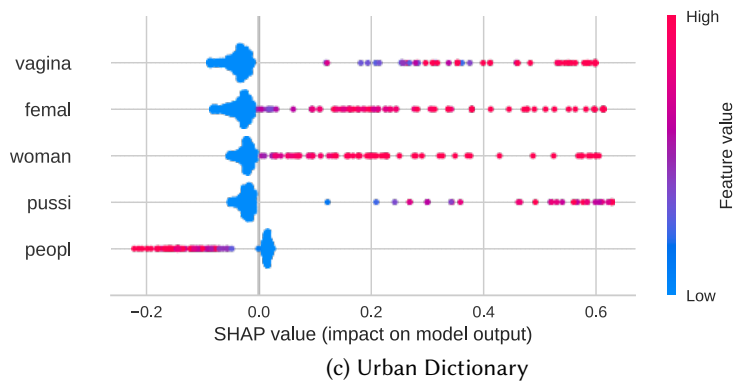
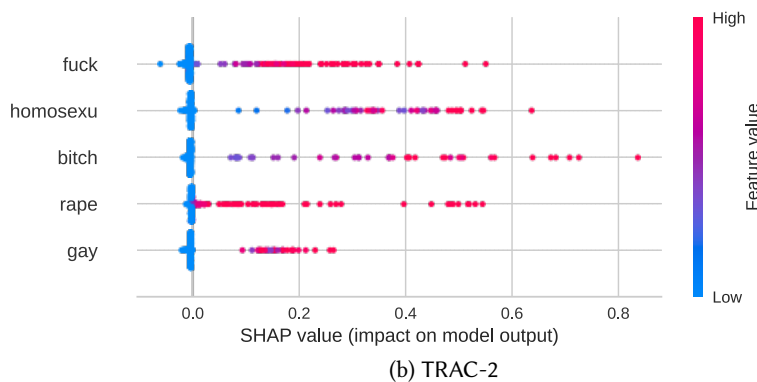
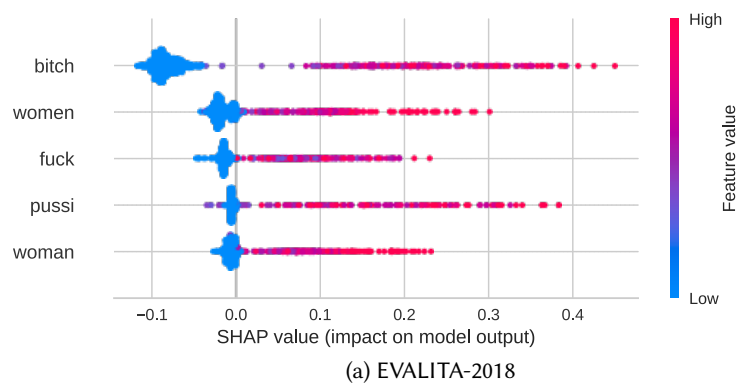


Figure 2: Features impact on prediction according to SHAP values

represents zero (blue), i.e., the absence of the feature.

As the figures show, different feature effects were observed for the three datasets. In the case of EVALITA-2018, all words showed a clear distinction between the effect of high feature values and their absence. Particularly, the highest positive effects were observed for “bitch”. Although high values of “women” and “fuck” showed a positive effect in the prediction, the

plot also reveals that low values on the top-5 features reduced the likelihood of predicting misogyny. In fact, in the case of the word “bitch”, its absence tended to incline predictions to the negative class. For the term “fuck”, and “women”, on the other hand, the magnitude of SHAP values is different, i.e. features with high values contribute less to predictions, and the push of low values towards the negative class is less visible (closer to zero Shapley value). For these two words, values are more concentrated in lower values, having a reduced impact on both classes. Similarly to the behaviour of the top-ranked word, the term “pussi” widens the span of importance, with high values driving the classifier to the positive class.

TRAC-2 showed the highest non-linear effects. The concentration of low values around the zero indicates that despite the presence of the word impacts on the misogynist classification, the model is not sensitive to its absence. In other words, while the absence of the word does not provide any information to the classifier, its presence induces a positive classification. Then, the analysis in this case should be oriented to determine the influence of high values. In this regard, the top-3 terms have a wider range of importance for the classifier, although the concentration of points is more scarce in the case of the word “bitch”. On the other hand, “rape” and “gay” are closer to the zero axis than the other words, implying that their absence or presence with low values have little influence on the classifier output.

Finally, for Urban Dictionary, the behaviour of the four top-ranked words seems to be alike. The absence of such terms showed to induce negative classifications, tending to move away from the zero SHAP values. Oppositely, high values are important for predicting the positive class, with all the instances distributed in a similar importance range. The word “peopl”, in turn, is also important but for predicting the negative class. High values of this word, induce the classifier to identify the content as non-misogynistic.

From looking at the model internals through the explanation results of the three datasets, we can identify distinctive feature behaviours: (1) words whose absence predicts the negative class, and its presence with high values the positive one (e.g., “bitch” in EVALITA-2018); (2) words whose absence is not tied to any class, but with high values indicating the positive class (e.g., “fuck” and “homosexu” in TRAC-2); (3) words whose absence is not tied to any class, and high values are not important either (e.g., “rape” and “gay” in TRAC-2). The consequences in classification of taking words with such distinctive behaviours into account during learning are analyzed in the next section.

3.5. Effect of feature removal

One of the simplest strategies to mitigate the impact of unintended bias is to add new instances to the training set in which the bias sensitive terms are used in the negative class. Nonetheless, this could not be a simple task due to not only the limited availability of labeled data, but also the possibility of introducing additional biases related to particularities of the newly introduced data and how they were annotated. As a result, before modifying the training data it could be useful to explore how the model behaves when removing the bias sensitive words.

Figure 3 shows the ROC curves when the complete set of features are considered and removing each of the important, bias sensitive words. The resulting AUC and F_1 scores are also included in the Figures. An AUC improvement would imply that removing the biased terms helped to avoid some false positives. As for model explanation a Random Forest classifier was used for

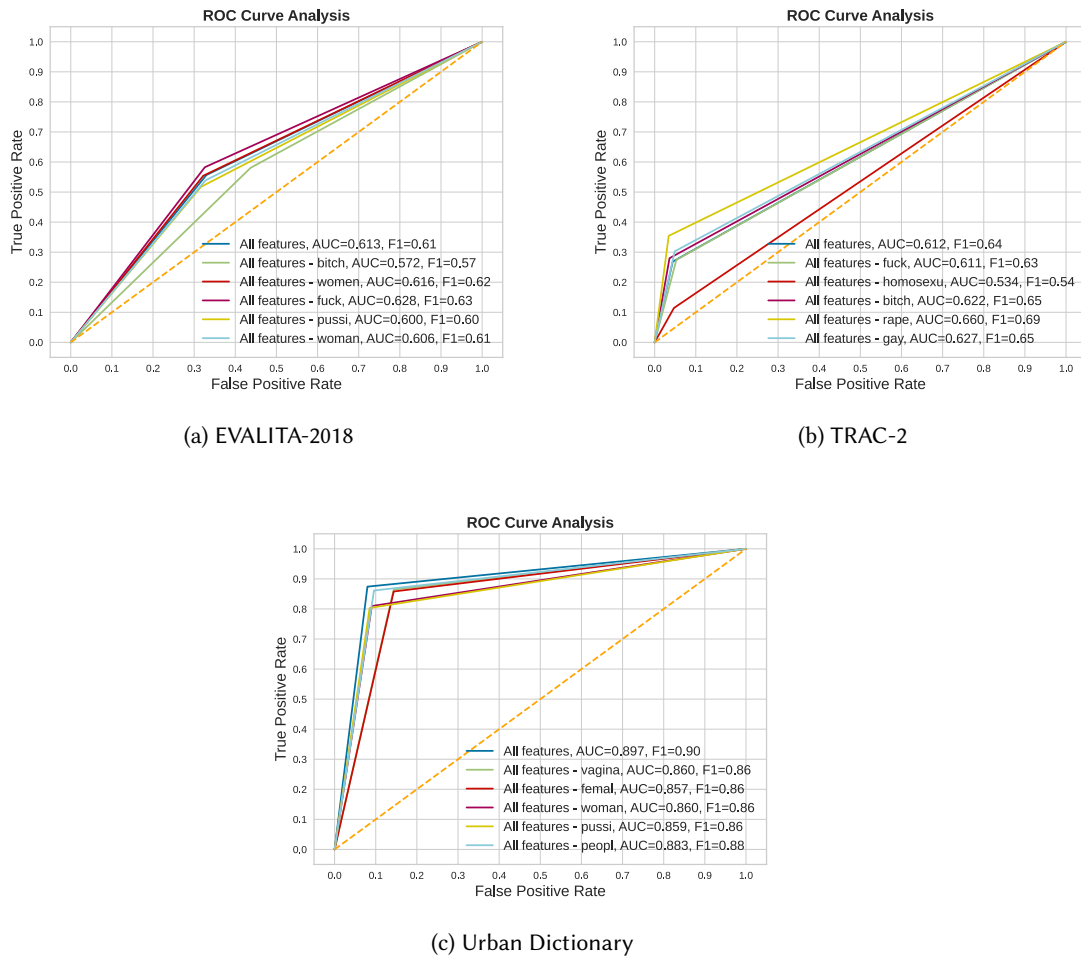


Figure 3: ROC curves after removing features

obtaining the results reported in this section.

For EVALITA-2018, the removal of the word “bitch” deteriorated model performance. Although identified by both SOAC and SPCPD metrics as a potential source of bias, its behaviour indicated that both its absence and presence had a high impact on prediction and, therefore, negative consequences during learning. The same happened with the word “pussi”. Instead, “women” and “fuck” exhibited a different behaviour to the previous words. In those cases, the opposite forces were more constrained and their removal led to an improvement in the model. As observed, the more reduced the feature values range, the higher the AUC improvement. This is, both words were deemed relevant when explaining the model, but they did not provide worthy information to distinguish misogynistic content from non-misogynistic one.

In the TRAC-2 dataset, removing “fuck” and “homosexu”, which were the most relevant features according to SHAP, did not improve prediction. Hence, they could be assumed to

be important to distinguish between classes, which is consistent with high values helping in the prediction of the positive class. Instead, removing “bitch” slightly improved the AUC score, also consistent with showing a higher dispersion of high feature values. The SHAP value distributions of “rape” and “gay” were closed to the zero axis, and thus high values did not relate to high SHAP values. As a result, removing such features allowed to improve the classification performance. On the other hand, “homosexu” and “bitch” were considered bias sensitive words by both SOAC and SPCPD metrics, and also for the model according to SHAP values. Nonetheless, their impact on model performance depended on how they contributed to identify the misogynistic content.

For Urban Dictionary, excluding the SHAP identified words (in some cases also identified by the SOAC and SPCPD metrics) did not lead to any significant change. This is consistent with the SHAP value behaviour, which showed that their absence was not necessarily helpful for distinguishing instances belonging to the negative class. The only exception was “peopl”, which showed a behaviour in the opposite direction of that of the other words.

To further understand the gains in AUC scores once features are removed, we computed the AUC-related metrics to measure unintended bias proposed by Borkan *et al.* [15]. These metrics rely on dividing the test data into identity or demographic subgroups, and then computing AUC for each group. Then, Subgroup AUC ($AUC_{Subgroup}$) measures the separability of instances within the subgroup containing the bias sensitive word.

In addition to computing AUC for each group, Borkan *et al.* [15] also defined a metric for comparing the subgroup results to the rest of the data (background data). In this regard, Background Positive Subgroup Negative (AUC_{BPSN}) calculates AUC on the positive examples from the background and the negative examples from the subgroup. A high AUC_{BPSN} means that few negative examples from the subgroup are classified as false positives at many thresholds. On the other hand, Background Negative Subgroup Positive (AUC_{BNSP}) calculates AUC on the negative examples from the background and the positive examples from the subgroup. A high AUC_{BNSP} implies that few positive examples from the subgroup are classified as false negatives at many thresholds.

Table 5 shows the AUC-related metric scores for the SHAP identified terms in the three datasets considering the complete set of features and after removing each of the identified words.

In the EVALITA-2018 dataset, the $AUC_{Subgroup}$ improved after removing the word “bitch”, denoting that the presence of such word helped to separate between misogynistic and non-misogynistic instances. However, after removing it, the model failed to distinguish non-misogynistic in the subgroup from misogynistic outside it (increasing false positives). Likewise, the model became worse at distinguishing misogynistic in the subgroup from outside non-misogynistic texts. The gain in $AUC_{Subgroup}$ was not enough to overcome the loses in AUC_{BPSN} and AUC_{BNSP} , so the overall AUC decreased. No improvement was observed when removing “women”, while removing “fuck” considerably improved the $AUC_{Subgroup}$ and AUC_{BNSP} , contributing to a better overall AUC score. Instead, the initial model was better at avoiding false positives having the word “fuck” as denoted by the slightly better AUC_{BPSN} .

For the TRAC-2 dataset, removing “bitch”, “rape” and “gay” improved the overall model AUC score and almost all AUC-related metrics. Particularly, removing “bitch” reduced the number of false positives, thus improving AUC_{BPSN} . This situation could indicate that removing the

word helped to reduce bias.

Finally, the removal of the SHAP identified words for Urban Dictionary did not contribute to better separate examples, as their absence did not improve the AUC metric. The three words in the table were considered bias sensitive by the SPCPD metric, which could be explained by the highly unbalanced term distribution in the two classes (e.g., the $AUC_{Subgroup}$ of “vagina” can not be calculated as it appears in all examples). The three terms were also important for the model according to SHAP values, but their presence seemed not to be the cause nor contribute to the false positive rate.

It is worth noting that we only removed important words as a means to illustrate their impact on model learning. However, the performance of models can not be solely explained by the presence or absence of a single term as terms are expected to interact in texts in multiple forms (e.g. correlation). More importantly, instead of simply removing individual features, feature interaction should be evaluated to discard any correlation effects. This would allow selecting an appropriate mitigation strategy to correct the discovered undesired feature effects.

4. Conclusions

The aim of this study was to measure and explain the effect of bias sensitive words on the models trained for misogyny detection. To this end, we studied how the presence of potential bias sensitive words could affect model prediction. The experimental study was based on three misogyny datasets.

After detecting and analyzing bias sensitive terms, we quantified their effect on model outcome by different AUC-related metrics. We compared the performance of the models including all words and when removing one by one the detected bias sensitive words. Explanation techniques allowed to discard some of the initially identified words as they did not show to greatly impact model predictions. Moreover, terms showing a strong effect on model outcome showed to induce avoidable false positives.

The exploration of the three datasets allowed us to draw some insights. First, the methods to identify bias sensitive words should focus not only in training data distribution (as SOAC metric) or the model output (as SPCPD metric), but also consider the role words have during model training. Second, even if individual terms could induce bias by themselves, their interaction with other terms should be considered when quantifying and mitigating the potential biases.

Although our work is preliminary, we hope that it can contribute to further developing the discussion of not only assessing bias in the training data and focusing on performance metrics, but also on how sensitive are models to biased data. Even though this work considered sensitive bias terms related to misogyny, the same analysis can be extended to other unintended or stereotypical related terms, or even other hate speech classification tasks. As future work, we envision an extended version of the performed study including the design and implementation of bias mitigation strategies.

Table 5
AUC-related metrics for subgroups

AUC	word	All features	After removal
EVALITA-2018			
$AUC_{Subgroup}$	bitch	52.28	<u>54.97</u>
AUC_{BPSN}		<u>61.15</u>	56.73
AUC_{BNSP}		<u>77.39</u>	65.68
$AUC_{Subgroup}$	women	<u>61.02</u>	56.50
AUC_{BPSN}		<u>61.26</u>	59.89
AUC_{BNSP}		<u>60.74</u>	51.12
$AUC_{Subgroup}$	fuck	58.44	<u>66.78</u>
AUC_{BPSN}		<u>63.76</u>	63.25
AUC_{BNSP}		74.95	<u>75.52</u>
TRAC-2			
$AUC_{Subgroup}$	bitch	47.86	<u>57.26</u>
AUC_{BPSN}		61.10	<u>62.40</u>
AUC_{BNSP}		<u>77.61</u>	59.26
$AUC_{Subgroup}$	rape	47.30	<u>52.87</u>
AUC_{BPSN}		62.79	<u>67.27</u>
AUC_{BNSP}		50.67	<u>53.24</u>
$AUC_{Subgroup}$	gay	55.94	<u>60.63</u>
AUC_{BPSN}		61.51	<u>62.68</u>
AUC_{BNSP}		58.72	<u>63.16</u>
Urban Dictionary			
$AUC_{Subgroup}$	vagina	-	-
AUC_{BPSN}		<u>89.12</u>	85.93
AUC_{BNSP}		<u>95.19</u>	87.08
$AUC_{Subgroup}$	femal	82.41	<u>82.75</u>
AUC_{BPSN}		<u>89.63</u>	86.04
AUC_{BNSP}		<u>93.42</u>	83.24
$AUC_{Subgroup}$	woman	<u>80.79</u>	76.65
AUC_{BPSN}		<u>89.64</u>	86.86
AUC_{BNSP}		<u>93.17</u>	77.66

References

- [1] K. Saha, E. Chandrasekharan, M. De Choudhury, Prevalence and psychological effects of hateful speech in online college communities, in: Proceedings of the 10th ACM Conference on Web Science, 2019, pp. 255–264.
- [2] J. H. Park, J. Shin, P. Fung, Reducing gender bias in abusive language detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018, pp. 2799–2804.
- [3] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1668–1678.
- [4] D. Nozza, C. Volpetti, E. Fersini, Unintended bias in misogyny detection, in: 2019

- IEEE/WIC/ACM International Conference on Web Intelligence (WI), Thessaloniki, Greece, 2019, pp. 149–155.
- [5] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, 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, Minneapolis, MN, USA, 2019, pp. 602–608.
 - [6] L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and mitigating unintended bias in text classification, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), New Orleans, LA, USA, 2018, pp. 67–73.
 - [7] A. Olteanu, E. Kiciman, C. Castillo, A critical review of online social data: Biases, methodological pitfalls, and ethical boundaries, in: Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM '18), Marina Del Rey, CA, USA, 2018, pp. 785–786.
 - [8] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of “bias” in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020, pp. 5454–5476.
 - [9] P. Badjatiya, M. Gupta, V. Varma, Stereotypical bias removal for hate speech detection task using knowledge-based generalizations, in: The World Wide Web Conference (WWW '19), San Francisco, USA, 2019, pp. 49–59.
 - [10] B. Vidgen, L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, PLOS ONE 15 (2021) 1–32.
 - [11] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at IberEval 2018, in: Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), Sevilla, Spain, 2018, pp. 214–228.
 - [12] E. Fersini, P. Rosso, M. Anzovino, AMI@EVALITA2020: Automatic misogyny identification, in: Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), Online, 2020.
 - [13] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., 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, Minneapolis, MN, USA, 2019, pp. 54–63.
 - [14] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in Twitter: A multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360.
 - [15] D. Borkan, L. Dixon, J. Sorensen, N. Thain, L. Vasserman, Nuanced metrics for measuring unintended bias with real data for text classification, in: Companion Proceedings of The 2019 World Wide Web Conference (WWW '19), San Francisco, USA, 2019, pp. 491–500.
 - [16] E. Fersini, D. Nozza, P. Rosso, Measuring and mitigating unintended bias in text classification, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), New Orleans, LA, USA, 2018, pp. 67–73.
 - [17] S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, A. K. Ojha, Developing a multilingual annotated corpus of misogyny and aggression, in: Proceedings of the 2nd Workshop on Trolling, Aggression and Cyberbullying (TRAC-2), Marseille, France, 2020, pp. 158–168.

- [18] T. Lynn, P. T. Endo, P. Rosati, I. Silva, G. L. Santos, D. Ging, Data set for automatic detection of online misogynistic speech, *Data in Brief* 26 (2019) 104223.
- [19] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Long Beach, USA, 2017, pp. 4768–4777.
- [20] L. S. Shapley, *A Value for N-Person Games*, RAND Corporation, Santa Monica, CA, 1952. doi:10.7249/P0295.
- [21] S. M. Lundberg, G. G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles, *CoRR abs/1802.03888* (2018).