Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists

Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists

Giuseppe Attanasio[1,2], **Debora Nozza**[1], **Dirk Hovy**[1], **Elena Baralis**[2]

[1]Bocconi University, Milan, Italy
[2]Politecnico di Torino, Turin, Italy
{giuseppe.attanasio3,debora.nozza,dirk.hovy}@unibocconi.it,
elena.baralis@polito.it

Abstract

Warning: This paper contains examples of language that some people may find offensive.

Natural Language Processing (NLP) models risk overfitting to specific terms in the training data, thereby reducing their performance, fairness, and generalizability. E.g., neural hate speech detection models are strongly influenced by identity terms like *gay*, or *women*, resulting in false positives, severe unintended bias, and lower performance. Most mitigation techniques use lists of identity terms or samples from the target domain during training. However, this approach requires a-priori knowledge and introduces further bias if important terms are neglected. Instead, we propose a knowledge-free Entropy-based Attention Regularization (EAR) to discourage overfitting to training-specific terms. An additional objective function penalizes tokens with low self-attention entropy. We fine-tune BERT via EAR: the resulting model matches or exceeds state-of-the-art performance for hate speech classification and bias metrics on three benchmark corpora in English and Italian. EAR also reveals overfitting terms, i.e., terms most likely to induce bias, to help identify their effect on the model, task, and predictions.

1 Introduction

Online hate speech is growing at a rapid pace, with effects that can result in dangerous criminal acts offline. Due to its verbal nature, various Natural Language Processing approaches have been proposed (Qian et al., 2018; Indurthi et al., 2019; Attanasio and Pastor, 2020; Kennedy et al., 2020; Vidgen et al., 2021, inter alia). Recently, detection performance has significantly improved with the use of large pre-trained language models based on Transformers (Vaswani et al., 2017), such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). However, several works have shown that by fine-tuning neural language models on hate speech detection, the classifiers obtained contain severe *unintended bias* (Dixon et al., 2018), i.e. they perform better or worse when texts mention specific *identity terms* (such as *gay*, *Muslim*, or *woman*). As a result, a sentence like "As a Muslim woman, I agree" would be wrongly classified as hate speech, purely due to the presence of two identity terms, i.e., terms referring to specific groups based on their socio-demographic features. One cause of false positives is selection bias in the keyword-driven collection of corpora (Ousidhoum et al., 2020). Figure 1 shows a false positive example for a fine-tuned BERT model on hate speech detection. Ideally, the model should rely on the words *adore* and *you*. Instead, BERT overfitted to the word *Girl* and associated it with a hateful context. This unwanted effect demonstrates the issues of lexical overfitting, and how they cause uninteded bias on identity terms.

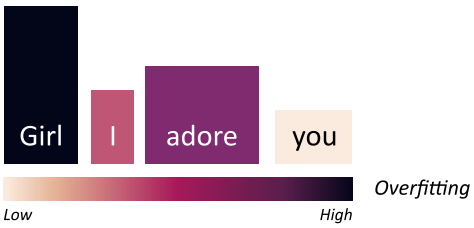

Figure 1: False positive from BERT as a hate speech detector. The darker and taller the bar, the higher the overfitting on the term.

Various methods have been proposed to mitigate and measure (unintended) bias (Elazar and Goldberg, 2018; Park et al., 2018; Dixon et al., 2018; Nozza et al., 2019; Kennedy et al., 2020; Vaidya et al., 2020). However, all those methods rely on the availability of a set of *identity terms*. This is a severe limitation, which hinders the generalizability and applicability of hate detection models

Findings of the Association for Computational Linguistics: ACL 2022, pages 1105 - 1119

to real-world contexts. For example, a model designed to reduce the unintended bias on gender-related terms (such as *woman*, *wife*) will not address unintended bias for religious affiliation. So practitioners must decide a-priori *"which vulnerable groups are present in our data?"*

We propose an Entropy-based Attention Regularization (EAR) that forces the model to build token representations by attending to a wider context, i.e., consider a larger number of tokens from the rest of the sentence. We measure the attended context as the entropy of the self-attention weight distribution over the input sequence. We use EAR as a regularization term in the loss computation to maximize each token's entropy. We apply EAR to BERT. The resulting model (BERT+EAR) significantly improves performance on unintended bias mitigation in English and Italian. In addition, it requires no a-priori knowledge (e.g., sets of identity terms), making it fairer and more general. The contextualized representations EAR induces avoid basing the classification on individual terms and, ultimately, mitigate lexical overfitting and intrinsic bias from pre-trained weights.

As a training by-product, EAR lets us extract the overfitting terms, i.e., terms accounting for narrower context that most likely induce unintended bias. These terms can highlight possible weaknesses in the model: from the over-sensitivity of pre-trained weights to specific words (Sheng et al., 2019; Nangia et al., 2020; Vig et al., 2020), to over-specialization of training corpora on the keywords used for collecting data (Ousidhoum et al., 2020).

Note that while we show results on BERT, EAR is applicable to any attention-based architecture.

Contributions. EAR is a novel entropy-based attention regularization method to mitigate unintended bias by reducing lexical overfitting. It is applied to all terms, so it *does not need a-priori domain knowledge* (e.g, predefined term lists). Independent of domain-specific information, EAR *generalizes better to different languages and contexts* compared to similar approaches. Attention entropy is used to extract a list of the most likely *biased terms*. EAR code is available at `https://github.com/g8a9/ear`.

2 Entropy-based Attention Regularization

Attention was originally designed for aligning target and source sequences in machine translation (Graves, 2013; Bahdanau et al., 2015). However, in the Transformer architecture (Vaswani et al., 2017), it has become a means to account for lexical influence and long-range dependencies. It also provides useful information about the importance of a term for the output (Wiegreffe and Pinter, 2019; Brunner et al., 2020; Sun and Marasović, 2021). Here, we use the notion of attention entropy, and EAR's use of it in BERT. Note, though, that EAR can be used with *any* attention-based architecture.

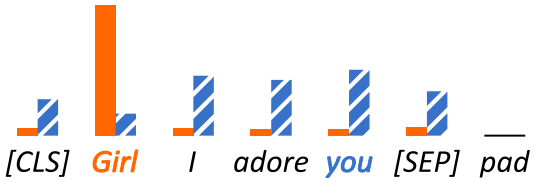

Figure 2: Self-attention distribution on tokens *Girl* (solid orange) and *you* (shaded blue). Attention for *Girl* is concentrated on its representation: its entropy is low. Attention for *you* is spread: its entropy is high.

Self-attention in Transformers. The Transformer model consists of two connected units, an encoder and a decoder, designed for sequence-to-sequence tasks.

A transformer encoder applies scaled-dot product self-attention over the input tokens to compute N independent attention heads.[1] Let $E = [\boldsymbol{e_0}, ..., \boldsymbol{e_{d_s}}]$ be the sequence of input embeddings, with $\boldsymbol{e_i} \in \mathbb{R}^{d_m}$. For the h-th attention head and i-th position, each embedding $\boldsymbol{e_i}$ is projected into a query $\boldsymbol{q_{h,i}}$, a key $\boldsymbol{k_{h,i}}$ and value $\boldsymbol{v_{h,i}}$. So each token expresses an attention distribution over all input embeddings as

$$a_{h,i} = \text{softmax}\left(\frac{\boldsymbol{q}_{\boldsymbol{h,i}}^T K_h}{\sqrt{d_k}}\right) \quad (1)$$

where K_h is the matrix of keys and d_k their dimension.

Attention weights $a_{h,i} = [a_{h,i,0}, ..., a_{h,i,d_s}]$, where $a_{h,i,j} \in [0, 1]$ and $\sum_j a_{h,i,j} = 1$, can be seen as a soft-indexing over the values. Since the values are projections of the tokens themselves, each weight in self-attention measures the contribution of its token to the attention head and, in turn, to the new token representation. We provide additional details to the self-attention mechanism in Appendix A.

[1] In the following, we use *token* and *embedding* interchangeably. We represent vectors with lowercase bold letters.

Attention entropy. Information *entropy* was first introduced in Shannon (1948), and measures the average information content of a random variable X with the set $[x_0, ..., x_n]$ of possible outcomes. It is defined as

$$H(X) = -\sum_i P(x_i) \log P(x_i) \tag{2}$$

Following Ghader and Monz (2017), we compute the entropy in the self-attention heads by interpreting each token's attention distribution as a probability mass function of a discrete random variable. The input embeddings are the possible outcomes, and the attention weights their probability.

For the sake of simplicity, we now discuss the computation of attention entropy of a single token in a standard transformer encoder. Attention weights are first averaged over heads by defining $a'_{i,j} = \frac{1}{h}\sum_h a_{h,i,j}$ as the mean attention that the token at position i pays to the token at position j. Then, we define a probability mass function by applying a softmax operator:

$$a_{i,j} = \frac{e^{a'_{i,j}}}{\sum_j e^{a'_{i,j}}} \tag{3}$$

We define the attention entropy as follows

$$H_i = -\sum_{j=0}^{d_s} a_{i,j} \log a_{i,j} \tag{4}$$

Intuitively, attention entropy measures the degree of contextualization while constructing the model's upper level's embedding. A large entropy suggests that a wider context contributes to the new embedding, while a small entropy tells the opposite: only a few tokens are deemed relevant. From a broader viewpoint, contextualized tokens improve the information passage between continuous layers by re-distributing the information content for every unit involved.

Figure 2 shows a toy example of self-attention distributions for two arbitrary tokens. Solid orange bars correspond to $a_{\text{Girl},j}$, while shaded blue bars correspond to $a_{\text{you},j}$. The toy example illustrates the correlation between attention distributions and entropy. The representation of *you* uses a wider context and, thus, it has a higher attention entropy. Note that, if present, we discard padding tokens from the attention entropy computation. Conversely, we include special tokens when required by the downstream task.

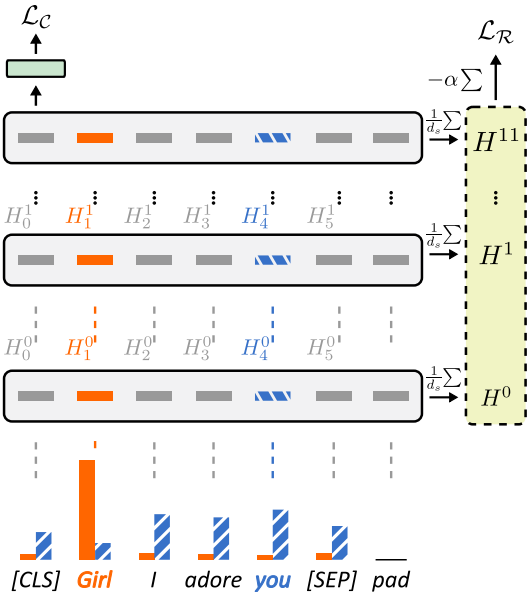

Figure 3: Overview of BERT+EAR. Grey boxes are Transformer layers. Each builds a token with attention entropy H_i^ℓ. Right green box pools layer-wise contextualization contributions and outputs regularization loss. First layer self-attention distribution (bottom) shown for *you* (shaded blue) and *Girl* (solid orange).

EAR in BERT. We introduced attention entropy as a proxy for the degree of contextualization of token representations above. Following this intuition, we propose BERT with EAR mitigation (BERT+EAR), a novel model trained to learn tokens with maximal self-attention entropy over the input sequence. We fine-tune BERT+EAR in the downstream task of hate speech detection. Note, though, that the approach is feasible for any classification task. In classification models, having more contextualized tokens avoids individual terms driving the classification outcome because they got over-attentioned.

Although EAR is applicable to any Transformer-based model, we base our approach here on the BERT (Devlin et al., 2019) base architecture. BERT provides an informative case study, given the number of architectures it has spawned and the recent interest in its attention patterns (Clark et al., 2019b; Kovaleva et al., 2019; Serrano and Smith, 2019). BERT consists of twelve stacked transformer encoders, each running self-attention on the output of the previous encoder. In BERT+EAR, we build new tokens with the maximal information content coming from the previous layer for every transformer layer in the architecture. Using Equation 4, we first compute the attention entropy of each token in the input sentence. We then take their

mean and define the *average contextualization* for the ℓ-*th* layer as

$$H^{\ell} = \frac{1}{d_s} \sum_{i=0}^{d_s} H_i^{\ell} \tag{5}$$

where H_i^{ℓ} is the attention entropy of the token at position i, and d_s is the length of the input sequence (excluding the padding tokens but including the [CLS] and [SEP] special tokens). Finally, we introduce a new regularization term to the model loss to maximize the entropy at each layer:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_R, \quad \mathcal{L}_R = -\alpha \sum_{l} H^{\ell} \tag{6}$$

$\mathcal{L}$ is the total loss, $\mathcal{L}_C$ and $\mathcal{L}_R$ are the classification and regularization loss, respectively, and $\alpha \in \mathbb{R}$ is the regularization strength. As in previous work, $\mathcal{L}_C$ is the Cross Entropy loss obtained with a linear layer on top of the last encoder as a classification head. It receives the [CLS] embedding and outputs the probability of the positive class (Hate).

The new regularization term $\mathcal{L}_R$ frames the task of maximal contextualization learning in the network. This framing has several advantages over existing approaches. First, it is a sum of differentiable terms and is hence differentiable. We can thus optimize BERT+EAR with classical back-propagation updates. Second, the regularization is agnostic to specific identity terms. It instead induces the network to learn contextualized tokens globally. This induction is crucial to regularize biased terms that might not be known in advance. Finally, note that the $\mathcal{L}_R$ pools each layer's entropy-based contributions H^{ℓ}. Each term H^{ℓ} is in turn dependent on the sole attention entropy defined in Equation 4. This makes the setup a general framework not limited to BERT. $\mathcal{L}_R$ can be used to evaluate and maximize the token contextualization in any attention-based architecture.

Figure 3 shows a graphical overview of BERT+EAR. Each layer provides a contextualization contributing to the loss independently, where layers with a low average contextualization increase the loss the most. Note also that, similarly to He et al. (2016), $\mathcal{L}_R$ introduces skip connections between layers and the classification head, so shorter paths for the contextualization information to flow.

Insights from attention entropy. On the one hand, we use attention entropy maximization to train BERT+EAR and test its classification and bias mitigation performance. On the other hand, we can leverage attention entropy to automatically extract the tokens with the lowest contextualization, which are the most likely to induce unintended bias. When a sentence is fed through a model like BERT, we can inspect the attention distribution of its terms[2].

We propose to exploit entropy, and hence contextualization, to gain insights into any attention-based model. Given a corpus and a model we want to inspect, we repeatedly query the model with sentences from the corpus and collect each token's attention entropy. Finally, we take each token's mean to measure the impact it has on bias, where lower is worse. Note that the same term can impact bias differently depending on the sentence.

While our approach works for any attention-based model and data set, we test it on fine-tuned classifiers to extract the biased terms learned on the training data set. We discuss this functionality in Section 5.

3 Experimental settings

In this work, we consider the problem of *unintended bias* (Dixon et al., 2018): "*a model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others*".

Datasets. Unintended bias is measured on synthetic test sets, artificially generated by filling manually defined contexts with identity terms (e.g., *I hate all* ___, *I love all* ___) . By construction, each identity term appears 50% of the time in hateful contexts and 50% in non-hateful ones. If a model then classifies the instances related to one identity term differently than the others, it means that the model contains unintended bias towards that term, e.g., if every instance containing the term *women* is labelled hateful, independently of the context. Synthetic test sets simulate new data, so a model that has low performance on this set demonstrates low generalization abilities and incapacity to be used in real-world contexts and applications.

We test BERT+EAR on hate speech datasets with associated synthetic test sets to measure unintended bias.

MISOGYNY (EN) (Fersini et al., 2018) is a state-of-the-art corpus for misogyny detection in English.

[2]For complex terms, we average the attention entropy of their sub-words.

	MISOGINY (EN)	MISOGINY (IT)	MLMA
# Train	4,000	5,000	5082
# Test	1,000	1,000	565
% Validation	10	10	10
% Hate (train, test)	45, 46	47, 53	88, 88
B_2	0.858	0.852	0.881
# Synthetic	1,464	1,908	77,000
# Identity terms	12	18	50
% Hate (Synthetic)	50	50	50

Table 1: Statistics of the data sets.

The related synthetic test set (Nozza et al., 2019) was created via several manually defined templates and synonyms for "woman" as identity terms.

MISOGYNY (ITA) (Fersini et al., 2020) is the benchmark corpus for misogyny detection in Italian. The synthetic test set has been generated similarly to the English one. This dataset allows us to study EAR's impact on cross-lingual adaptation.

MULTILINGUAL AND MULTI-ASPECT HATE SPEECH (MLMA) (Ousidhoum et al., 2019) consists of tweets with various hate speech targets. We choose to work on its English part. We use the synthetic test provided in Dixon et al. (2018), generated by slotting a wide range of identity terms into manually defined templates.

Table 1 reports statistics of the data sets. Alongside the size of train, test, and validation sets, we report also the percentage of hateful instances to show the class balance. Note that MLMA is highly unbalanced with 88% of instances associated with the hateful class. Note that the original MULTILINGUAL AND MULTI-ASPECT dataset comes in a multi-label, multiple class setting. Following Ousidhoum et al. (2021), we used the *Hostility* dimension of the dataset as target label and created a *Hate* binary from it as follows. We considered single-labeled "Normal" instances to be non-hate/non-toxic and all the other instances to be toxic.

To further characterize our data sets, we explore the aspect of selection bias, reporting the measure B_2 (Ousidhoum et al., 2020). The metric ranges from 0 to 1 and evaluates how likely topics of the data set are to contain keywords of the data collection. Values above 0.7 demonstrate high selection bias, implying the need for unbiasing procedures.

We report also the size and number of identity terms used in the synthetic test sets. The percentage of hateful content is perfectly balanced (50%) since each identity term should appear exactly in the same context as the others to measure the unintended bias. See Appendix B for the list of identity terms and further preprocessing details.

3.1 Metrics

We use the weighted and binary F1-score of the hateful class ($\mathbf{F1_w}$ and $\mathbf{F1_{hate}}$) as classification metrics. We consider both due to the class imbalance of test sets (see Table 1).

We compute the unintended bias metrics from Dixon et al. (2018) and Borkan et al. (2019). They are computed from differences in the score distributions between instances mentioning a specific identity-term (*subgroup distribution*) and the rest (*background distribution*). The three per-term AUC-based bias scores are:

1) $AUC_{subgroup}$ calculates AUC only on the data subset of a given identity term. A low value means the model performs poorly in distinguishing between hateful and non-hateful comments that mention the identity term.

2) *Background Positive Subgroup Negative* (AUC_{bpsn}) calculates AUC on the hateful background examples and the non-hateful subgroup examples. A low value means that the model confuses non-hateful examples that mention the identity term with hateful examples that do not.

3) *Background Negative Subgroup Positive* (AUC_{bnsp}) calculates AUC on the non-hateful background examples and the hateful subgroup examples. A low value means that the model confuses hateful examples that mention the identity with non-hateful examples that do not.

We report the averaged metrics across identity terms, i.e., $\mathbf{AUC_{subgroup}}$, $\mathbf{AUC_{bpsn}}$, and $\mathbf{AUC_{bnsp}}$.[3]

3.2 Baselines

We compare BERT+EAR against the following existing approaches: (1) *BERT* (Devlin et al., 2019), (2) *BERT+SOC mitigation* (Kennedy et al., 2020), where the authors modify BERT's loss to lower the importance weight of identity terms, computed with the Sampling-and-Occlusion (SOC) algorithm (Jin et al., 2019), (3) Nozza et al. (2019), a single-layer neural network architecture based on the Universal Sentence Encoder (USE) representation (Cer et al., 2018), (4) Lees et al. (2020), a multilingual BERT model fine-tuned on the training data, (5) Ousidhoum et al. (2021), a classifier based on TF-

[3]Statistical significance and results from Lees et al. (2020) on these metrics could not be computed due to data unavailability and label distribution assumptions.

	Unintended bias (synthetic)					test	
	$\mathbf{AUC_{subgroup}}$	$\mathbf{AUC_{bnsp}}$	$\mathbf{AUC_{bpsn}}$	$\mathbf{F1_w}$	$\mathbf{F1_{hate}}$	$\mathbf{F1_w}$	$\mathbf{F1_{hate}}$
Nozza et al. (2019), no mitigation	49.83	49.83	49.83	49.97	51.33	**72.29**	**71.62**
Nozza et al. (2019), debiased	50.27	50.21	50.21	45.40	29.31	71.43	69.37
Zhang et al. (2020)	69.99	62.19	62.19	43.01	66.70	31.35	63.21
BERT, no mitigation	70.97	66.62	66.62	58.19	64.61	69.60	70.21
BERT+SOC mitigation	78.11	**76.60**	**76.60**	51.88	58.89	57.39	60.47
BERT+SOC mitigation, missing ITs	68.58	67.38	67.38	38.49	41.38	51.14	43.65
BERT+EAR	**80.08**	75.18	75.18	**62.59** $^{\bullet\blacktriangle}$	**70.58** $^{\bullet\blacktriangle}$	70.90 $^{\blacktriangle}$	70.83 $^{\blacktriangle}$
Lees et al. (2020), debiased	-	-	-	**47.00**	58.58	79.87	82.45
Zhang et al. (2020)	48.10	48.29	48.29	33.33	**66.66**	33.54	66.69
BERT, no mitigation	47.30	47.54	47.54	39.72	61.17	81.57	83.56
BERT+SOC mitigation, translated ITs	45.54	45.88	45.88	46.34	51.62	80.28	81.73
BERT+EAR	**48.59**	**48.65**	**48.65**	40.64	62.71 $^{\bullet\blacktriangle}$	**83.29** $^{\bullet\blacktriangle}$	**84.68** $^{\circ\blacktriangle}$
Ousidhoum et al. (2021), no mitigation	63.87	60.80	61.10	33.33	66.66	82.84	**93.80**
Zhang et al. (2020)	74.14	64.74	65.76	33.33	66.66	82.84	93.79
BERT, no mitigation	69.38	67.12	67.12	**50.24**	39.65	64.70	70.14
BERT+SOC mitigation	56.15	55.83	55.58	33.79	59.89	76.49	86.24
BERT+EAR	**74.31**	**71.43**	**71.25**	40.09	**67.45** $^{\bullet\blacktriangle}$	**83.05** $^{\bullet\blacktriangle}$	91.88 $^{\bullet\blacktriangle}$

Table 2: Results (in %) on MISOGYNY (EN) (top), MISOGYNY (ITA) (middle), and MLMA. Significance of BERT+EAR over BERT without mitigation ($^{\bullet}$: $p \leq 0.01$) and BERT with SOC mitigation ($^{\blacktriangle}$: $p \leq 0.01$).

IDF and Logistic Regression, and (6) Zhang et al. (2020), a debiasing training framework based on instance weighting.

The *debiased* version proposed in Lees et al. (2020) is obtained by training the model on additional samples from Wikipedia articles (assumed to be non-hateful) to balance the distribution of specific identity terms. Nozza et al. (2019) extracted these additional non-hateful samples from an external Twitter corpus (Waseem and Hovy, 2016).

To address the impact of different term lists, we also consider two different versions of BERT+SOC mitigation, one where we test the effect of *missing identity terms* and the other where the identity terms are *translated* for adapting to a new language.

4 Experimental Results

Table 2 shows classification and bias metrics on both synthetic and test set for the three corpora, i.e., MISOGYNY (EN) (top), MISOGYNY (ITA) (middle), and MLMA (bottom). The top rows in each table section report the performance of hate speech detection models specifically proposed for the respective dataset. The lower rows show the results of baselines and BERT+EAR. BERT+SOC mitigation uses the identity terms from Kennedy et al. (2020) (see Appendix C), unless a different identity terms lists is specified (e.g., "BERT+SOC mitigation, translated ITs").

BERT+EAR obtains comparable and, in most cases, better performance on all three datasets than all state-of-the-art debiasing approaches, which are based on (i) the knowledge of identity terms and (ii) data augmentation techniques. However, identity terms are not always readily available, which severely limits the generalization of those approaches. Similarly, there are several drawbacks to data augmentation with (assumed) non-hateful samples containing the identity terms. 1) Data augmentation is expensive. It requires filtering a large dataset (usually Wikipedia) and retraining the model with a much larger set of instances. 2) Data augmentation with task-specific identity terms requires prior knowledge of those terms, and is therefore limited by the authors' knowledge. 3) The overlap between identity terms in the evaluation set and the augmented data inevitably (but somewhat unfairly) improves the performance on the synthetic dataset.

BERT+EAR is overall the best debiasing model considering the proposed bias metrics. The only exception is MISOGYNY (EN), for which BERT+EAR has lower $\mathbf{AUC_{bnsp}}$ and $\mathbf{AUC_{bpsn}}$ than BERT+SOC mitigation. The latter's advantage, however, comes with high variability in the results. BERT+SOC mitigation seems more sensitive to random initialization. The standard deviation over 10 runs is 37%, compared to 13% of BERT+EAR. Figure 4 shows the $\text{AUC}_{\text{subgroup}}$ metric separately by identity term on MISOGYNY (EN). We compare

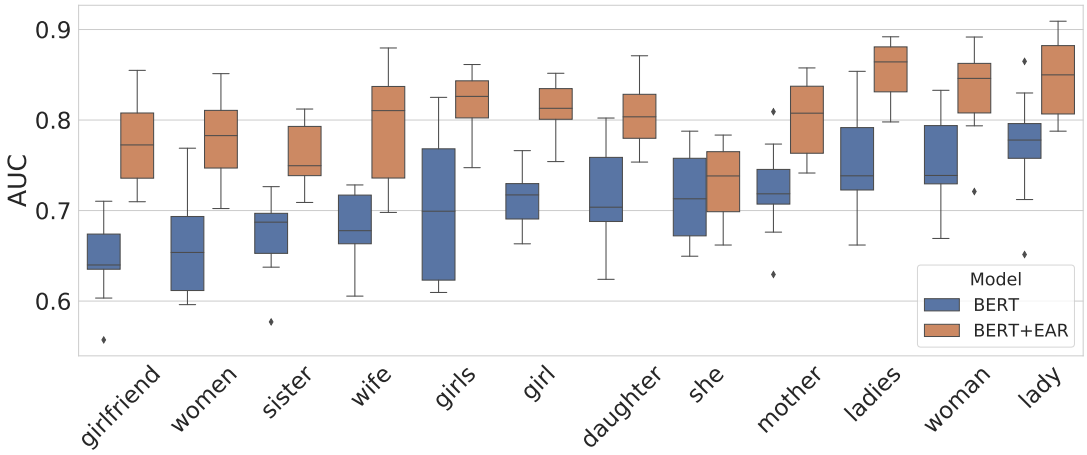

Figure 4: $AUC_{subgroup}$ results broken down by identity term on MISOGYNY (EN).

BERT and BERT+EAR over 10 different initialization runs. EAR improves BERT across all identity terms

Most existing models and AUC-based metrics for unintended bias focus only on the false positives (i.e., hateful instances wrongly recognized as non-hateful). While correctly recognizing hateful instances is important, we believe that the problem of false negatives is equally important. Since BERT+EAR does not rely on identity term lists, it regularizes terms that impact *both* the positive and negative class. BERT+EAR obtains an average decrease of 15.04% in false negative rate compared to BERT and BERT+SOC mitigation. Indeed, the performance difference between BERT+EAR vs. BERT and BERT+SOC is mainly due to non-hateful instances (∼95% of the time). Reducing the impact of overfitting terms like *f*ck* and *p*ssy* in MISOGYNY (EN) causes BERT+EAR to consider a larger context, and correctly labels them as non-hateful.

4.1 Error Analysis

Table 3 shows tweets from the MISOGYNY (EN) data set which have been correctly predicted by BERT+EAR but misclassified by BERT or BERT+SOC. These tweets serve as qualitative examples of the effectiveness of forcing the model to attend to a wider context and not overfit to training-specific terms, exploiting the richness of information (Nozza et al., 2017). The examples are an excerpt of the most common cases where BERT+EAR classifies the non-hateful examples correctly: (1) when slurs or negative words (such as *sk*nk*) are used in a non-hateful context, like slang or lyrics, (2) when many words associated with misogyny appear in the sentence (e.g., *rape*, *abuse*) and (3) when the hateful target is male and the instance should not be classified as misogynous. The use of a wider context by BERT+EAR allows the model identify such non-misogynous instances compared to BERT and BERT+SOC. In particular, BERT+SOC is even more biased in these cases because its debiasing techniques overly rely on specific terms (e.g. *woman*) and increase overfitting to training-specific examples.

4.2 Impact of predefined identity terms

We also analyze the impact of predefined identity term lists on performance by evaluating the effect of (i) missing identity terms, and (ii) adapting to a new language where the list is unavailable.

First, we remove every identity term of BERT+SOC from MISOGYNY (EN) that appears at least once in the evaluation set, here *women* and *woman* out of 24 terms. This reflects the real-world case where the identity term list does not contain a specific group present in the data. The significant performance drop resulting from this case (Table 2, top, "missing ITs") highlights a strong weakness of term-based mitigation strategies.

Second, we analyze the case where identity terms need to be adapted to a new language, e.g., Italian. We translated the English identity terms from BERT+SOC to Italian via Google Translate.[4] Table 2 (middle, "translated ITs") shows that the performance is lower than BERT+EAR. A simple translation of predefined identity terms is therefore not an option for cross-lingual settings. This aligns with the findings by Nozza (2021), that demon-

[4]For gendered Italian words, we kept both the masculine and the feminine (e.g., *muslim* → *musulmana*, *musulmano*).

text	Ground truth	BERT	BERT+SOC	BERT+EAR
I'm just a sk*nk for understanding the basics of life!	0	1	1	0
You're such a f*cking hoe, I love it - the new Kanye and Lil Pump I kings make women feel comfortable about their sexuality.	0	1	1	0
GIRL, YOU'RE HYSTERICAL. I AM DANCING SO HAPPY FOR TODAY	0	0	1	0
#metoo I'm a victim of rape, abuse and harrassment. Every woman who had any these experiences.	0	1	1	0
some people at school drive me insane. like cool b*tch! im depressed too!! doesnt mean im a f*cking c*nt	0	1	1	0
@male_user And you are a hysterical k*nt.	0	0	1	0
@male_user F*ck you p*ssy	0	1	1	0

Table 3: Examples of MISOGYNY (EN) tweets misclassified by BERT or BERT+SOC, and correctly classified by BERT+EAR. Next to the tweet text, we report the ground truth label and the prediction of each model. Exact phrasing changed to protect privacy.

strated that cross-lingual hate speech detection is limited by the use of non-hateful, language-specific taboo interjections that are not directly translatable.

In sum, we demonstrated that relying on a predefined list of identity terms is a strong limitation for performance and generalizability of the model. In contrast, BERT+EAR's independence from any predefined terms makes it the ideal model in real-world scenarios.

5 Extracting overfitting terms

While being the core of EAR, attention entropy serves another purpose. Once standard fine-tuning is concluded (i.e., with no regularization involved), models have overfitted specific terms. We identify these terms using attention entropy.

To extract the most indicative terms, we replicate training conditions. Specifically, we run inference using all the training data using a fine-tuned checkpoint and a standard BERT tokenizer. We collect attention entropy values for each term and average them over all training instances. Terms with lowest average entropy show the highest overfitting as the model learned them with a narrow context.[5]

Retrieving these terms after training allows us to gain insights into the domain and language-specific aspects driving the outcome.

Table 4 shows the top 10 terms with highest lexical overfitting on the studied datasets extracted from the corresponding fine-tuned model. We extract terms strongly correlated with the positive class, e.g., *womens*ck* (97%), *shut* (96%), *n*gger* (92%), *sb*rro* (97%), *c*lone* (95%). Note that these terms are *not* frequent in the corpus. Overfitting terms appear with an average document frequency of only 4.7%, while the most frequent terms have 32.5% average document frequency across datasets. These results suggest that the higher the class polarization of a token, the narrower the context BERT will use to learn its representation, and the higher the overfitting.

[5]To filter out noise, we report only words with a document frequency higher than 1%.

6 Related Work

The first works to study bias measurement and mitigation in neural representation aimed at removing implicit gender bias from word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Romanov et al., 2019; Ravfogel et al., 2020). More recently, researchers have started to focus on contextualized sentence representations and effective neural models for understanding the presence and resolution of bias (Nozza et al., 2021; Ousidhoum et al., 2021).

While the majority of proposed approaches focus on data augmentation (Dixon et al., 2018; Nozza et al., 2019; Sharma et al., 2020; Bartl et al., 2020; de Vassimon Manela et al., 2021), different approaches have been proposed for bias mitigation intervening directly in the objective function. Kennedy et al. (2020) proposed to apply regularization during training to the explanation-based importance of identity terms, obtained with Sampling-and-Occlusion (SOC) explanations (Jin et al., 2019). Kaneko and Bollegala (2021) pro-

Dataset	Overfitting terms
MISOGYNY (EN)	girls, womens*ck, hoes, c*ck, shut, stupid, hoe, p*ssy, trying, f*ck
MISOGYNY (ITA)	pezzo, bel, bellissima, scoperei, p*ttanona, zitta, sb*rro, t*ttona, bella, c*lone *(piece, nice, very nice, I'd f*ck, sl*t, shut up, c*m, b*sty, beautiful, fat*ss)*
MLMA	n*gger, n*gro, shut, chong, ching, d*ke, okay, sp*c, tw*t, f*ggot

Table 4: Terms with highest lexical overfitting identified using attention entropy.

posed a method for debiasing pre-trained contextual representation by retaining the learned semantic information for gender-related words (e.g., *she*, *woman*, *he*, *man*) and simultaneously removing any stereotypical biases in the pre-trained model. Zhou et al. (2021) exploited debiasing methods for natural language understanding (Clark et al., 2019a) to explicitly determine how much to trust the bias given the input. Vaidya et al. (2020) proposed a multi-task learning model for predicting the presence of identity terms alongside the toxicity of a sentence.

The main drawback of all aforementioned works is their strict reliance on a set of predefined identity terms. This list can be either defined manually by experts or extracted a-priori from the data set. In both cases, the subsequent debiasing models will be strongly affected by these biased terms, limiting the applicability of the trained model to new data. This is a severe limitation, since it is not always possible to retrain a model on new data to reduce bias, resulting in limited use in real-world cases.

7 Conclusion

We introduce EAR, a regularization approach applicable to any attention-based model. Our approach does not require any a-priori knowledge of identity terms, e.g., lists. This feature (i) allows us to generalize to different languages and contexts, and (ii) avoids neglecting important terms. Thus, it prevents the introduction of further bias. As part of the training procedure, EAR also discovers the impact of relevant domain-specific terms. This automatic term extraction provides researchers with an analysis tool to improve data collection and bias mitigation approaches.

EAR, applied to BERT, reliably classifies data with competitive performance and substantially improves various bias metrics. BERT+EAR generalizes better to new domains and languages than similar methods.

In future work, we will apply EAR-based models to different downstream tasks to both improve bias mitigation and automatically extract biased terms.

Acknowledgments

We would like to thank the anonymous reviewers and area chairs for their suggestion to strengthen the paper. This research is partially supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR), and by Fondazione Cariplo (grant No. 2020-4288, MONICA). DN, and DH are members of the MilaNLP group, and of the Data and Marketing Insights Unit at the Bocconi Institute for Data Science and Analysis. EB is member of the DataBase and Data Mining Group (DBDMG) at Politecnico di Torino. GA did part of the work as a member of the DBDMG and is currently a member of MilaNLP. Computing resources were partially provided by the SmartData@PoliTO center on Big Data and Data Science.

Ethical Considerations

In this paper, we propose term attention entropy as a proxy for unintended bias in attention-based architectures. Our approach allows us to extract, for a given classifier and data set, a list of terms that induce most of the bias in the model. While this list is intuitive and easy to obtain, we would like to point out some ethical dual-use considerations.

The process of collecting the list is a data-driven approach, i.e., it is strongly dependent on the task, collected corpus, term frequencies, and the chosen model. Therefore, the list might lack specific terms or include terms that do not strictly perpetrate harm, but are prevalent in the sample. Because of these twin issues, the resulting lists should *not* be read as complete or absolute. We discourage users from developing new models based solely on the extracted terms. We want, instead, the terms to stand as a starting point for debugging and searching for potential bias issues in the task at hand, be it in data collection or model development.

Further, while the probability is low, we can not exclude the possibility that future users run EAR on other tasks and data sets to derive private information or profile vulnerable groups.

References

Giuseppe Attanasio and Eliana Pastor. 2020. PoliTeam @ AMI: Improving sentence embedding similarity with misogyny lexicons for automatic misogyny identificationin italian tweets. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, pages 48–54. Accademia University Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *International Conference on Learning Representations*.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019a. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019b. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. 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.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR-WS.org.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: Automatic Misogyny Identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2019. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

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

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw@ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2017. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280, Valencia, Spain. Association for Computational Linguistics.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Nedjma Ousidhoum, Yangqiu Song, and Dit-Yan Yeung. 2020. Comparative evaluation of label-agnostic selection bias in multilingual hate speech datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2532–2542, Online. Association for Computational Linguistics.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language 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 4262–4274, Online. 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.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Hierarchical CVAE for fine-grained hate speech classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3550–3559, Brussels, Belgium. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. 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 4187–4195, Minneapolis, Minnesota. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 358–364, New York, NY, USA. Association for Computing Machinery.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. 2014. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan. Association for Computational Linguistics.

Kaiser Sun and Ana Marasović. 2021. Effective attention sheds light on interpretability. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4126–4135, 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. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

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.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

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

Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

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.

A Details on self-attention in Transformers

The Transformer (Vaswani et al., 2017) is the building block of many recent neural language models. A Transformer model consists of two connected encoder and a decoder units which align a source and a target sequence. Differentiating from the original formulation, large language models, such as BERT, drop the encoder and use the remaining encoder to process a single input sequence.

A transformer encoder consists of a multi-head self-attention block and a position-wise, fully connected feed forward neural network. Both the self-attention block and the feed forward network adopt a residual skip connection and batch normalization. We provide details for a standard forward pass in the encoder. In attention blocks, the multi-head output is computed with Scaled Dot-Product Attention between a set of queries and keys of dimension d_k, and a set of values of dimension d_v. Let Q, K and V be the respective matrix representations. The attention is then computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

To improve expressiveness, the operation is performed on N different, independent linear projections of the same queries, keys and values, so that N attention heads are produced. The heads are then concatenated, projected back to the original input space, and finally fed through the fully connected neural network to produce the next layer embeddings. Let $E = [e_0, ..., e_{d_s}]$ be the sequence of input embeddings[6], with $e_i \in \mathbb{R}^{d_m}$. In the specific case of a transformer encoder, queries, keys and values correspond to the input embeddings - i.e. $Q = K = V = E$. As such, the output of the multi-head self-attention block is computed applying the previously presented Equation to the N token projections, concatenating and projecting back to the original space:

$$\text{MultiHead}(Q, K, V) = (\text{o}_0 || \ldots || \text{o}_N)\, W^O$$

where

$$\text{o}_h = \text{Attention}\left(QW_h^Q, KW_h^K, VW_h^V\right)$$

and W^O and each W_h^Q, W_h^K, W_h^V are projection matrices.

[6]The input embeddings for the first layer are the static token embeddings plus their position encoding.

B Experimental setup

Hyper-parameters All our experiments use the Hugging Face transformers library (Wolf et al., 2020). We base our models and tokenizers on the `bert-base-uncased` checkpoint for English tasks and on the `dbmdz/bert-base-italian-uncased` checkpoint for Italian. We pre-process and tokenize our data using the standard pre-trained BERT tokenizer, with a maximum sequence length of 120 and right padding. We train all models with the following hyperparameters: batch size=64, learning rate=0.00002, weight decay=0.01, learning rate warmup steps=10%, full precision, maximum number of training epochs=30, and early stopping on non-improving validation loss after 5 epochs. Table 2 report results of BERT+EAR trained for 20 epochs with no early stopping, and regularization strength $\alpha = 0.01$. We chose the latter parameters with grid search on $\alpha \in [0.0001, 0.001, 0.01, 0.1, 1]$ and epochs $\in [10, 20, 30, 40, 50]$. When fine-tuning on MULTILINGUAL AND MULTI-ASPECT, we use a weighted cross-entropy classification loss ($\mathcal{L}_C$) to discount class unbalance. Specifically, we normalize the loss for data points belonging to class C by the prior probability of C, evaluated as its relative frequency in the training set.

For Kennedy et al. (2020), Nozza et al. (2019), Lees et al. (2020), and Ousidhoum et al. (2021), we kept all the parameters as specified by the respective authors. Please refer to our repository (`https://github.com/g8a9/ear`) for further details or the respective publications.

We trained all models with 10 different initialization seeds per parameter configuration and averaged over them to obtain stable results and meaningfully compute significance.

Statistical significance We compute the statistical significance of BERT+EAR over BERT and BERT with SOC mitigation via bootstrap sampling, following Søgaard et al. (2014), using $^\circ$ and $^\triangle$ (and their filled counterparts for a stronger significance) symbols, respectively. We use 1000 bootstrap samples and a sample size of %20. For Hate Speech, significance can only be computed on F1-scores, since bias metrics require an assumption about the label distribution across identity terms that is not given.

Selection bias We computed the $\mathbf{B_2}$ metric following Ousidhoum et al. (2020). Specifically, we run the authors' code on each of our training dataset, using the query keywords used to sample each dataset. In case of queries composed of multiple words, we split and considered them separate keywords.

Dataset preprocessing The original MULTILINGUAL AND MULTI-ASPECT dataset comes in a multi-label, multiple class setting. Following Ousidhoum et al. (2021), we used the *Hostility* dimension of the dataset as target label and created a *Hate* binary from it as follows. We considered single-labeled "Normal" instances to be non-hate/non-toxic and all the other instances to be toxic.

Computation time We report NVIDIA Tesla V100 PCIE-16GB -equivalent computation time for the tested models. Averaging across the three presented data sets, training and evaluating 10 seeds of BERT+EAR (without early stop) requires 22 hours, compared to 72 hours for BERT+SOC and 7 hours for BERT. The regularization of attention entropy does not affect the computation time by a significant amount.

CO_2 emission Experiments were conducted using a private infrastructure, which has an estimated carbon efficiency of 0.432 kgCO_2eq/kWh. A cumulative of 319 hours of computation was performed on the hardware of type Tesla V100-PCIE-16GB (TDP of 300W). Total emissions are estimated to be 41.34 kgCO_2eq.

Estimations were conducted using the Machine Learning Impact calculator presented in (Lacoste et al., 2019).

C List of identity terms

In the following, we report the list of identity terms used in the considered data sets and methods.

(Kennedy et al., 2020): *muslim, jew, jews, white, islam, blacks, muslims, women, whites, gay, black, democrat, islamic, allah, jewish, lesbian, transgender, race, brown, woman, mexican, religion, homosexual, homosexuality, africans*

(Nozza et al., 2019): *woman*, *women*, *daughter*, *girl*, *girls*, *mother*, *she*, *wife*, *lady*, *ladies*, *girlfriend*, *sister*

(Fersini et al., 2020): *nonne*, *matrone*, *mamme*, *casalinghe*, *compagne*, *morose*, *femmine*, *donne*, *fidanzate*, *nonna*, *matrona*, *casalinga*, *morosa*, *femmina*, *mamma*, *donna*, *fidanzata*, *compagna*

(Dixon et al., 2018): *lesbian*, *gay*, *bisexual*, *transgender*, *trans*, *queer*, *lgbt*, *lgbtq*, *homosexual*, *straight*, *heterosexual*, *male*, *female*, *nonbinary*, *african*, *african american*, *black*, *white*, *european*, *hispanic*, *latino*, *latina*, *latinx*, *mexican*, *canadian*, *american*, *asian*, *indian*, *middle eastern*, *chinese*, *japanese*, *christian*, *muslim*, *jewish*, *buddhist*, *catholic*, *protestant*, *sikh*, *taoist*, *old*, *older*, *young*, *younger*, *teenage*, *millenial*, *middle aged*, *elderly*, *blind*, *deaf*, *paralyzed*