The Other Side of Compression: Measuring Bias in Pruned Transformers

Irina Proskurina, Guillaume Metzler, Julien Velcin

The Other Side of Compression: Measuring Bias in Pruned Transformers

Irina Proskurina(✉), Guillaume Metzler, and Julien Velcin

Univ Lyon, Univ Lyon 2, UR ERIC
Lyon, France
irina.proskurina@univ-lyon2.fr

Abstract. Social media platforms have become popular worldwide. Online discussion forums attract users because of their easy access, speech freedom, and ease of communication. Yet there are also possible negative aspects of such communication, including hostile and hate language. While fast and effective solutions for detecting inappropriate language online are constantly being developed, there is little research focusing on the bias of compressed language models that are commonly used nowadays. In this work, we evaluate bias in compressed models trained on Gab and Twitter speech data and estimate to which extent these pruned models capture the relevant context when classifying the input text as hateful, offensive or neutral. Results of our experiments show that transformer-based encoders with 70% or fewer preserved weights are prone to gender, racial, and religious identity-based bias, even if the performance loss is insignificant. We suggest a supervised attention mechanism to counter bias amplification using ground truth per-token hate speech annotation. The proposed method allows pruning BERT, RoBERTa and their distilled versions up to 50% while preserving 90% of their initial performance according to bias and plausibility scores.

Keywords: Hate speech recognition · Model fairness · Structured pruning · Compressing transformers

1 Introduction

The spread of offensive speech in social media is considered a precursor of numerous existing social issues, such as the distortion of victims' portrayal in society, social tension, dissemination of entrenched stereotypes, provoking hostility and hate crime, not to mention the mental toll. Rational content moderation and filtering in social networks is the primary tool for preventing these consequences of offensive speech. Given the number of everyday social media posts, the need for automated content monitoring looks inevitable. Automated solutions also help to prevent moral damage and the negative impact of disturbing texts on annotators [20]. Recently, algorithmic moderation has become a ubiquitous tool for the vast majority of social networks, including Facebook, YouTube and Twitter. Nevertheless, existing challenges of the hate speech detection task form a stumbling block to guaranteeing accurate and unbiased models' predictions. Context

sensitivity and an unclear author's intention are the main challenges at the data annotation stage. These factors are the primary sources of the annotators' disagreement during the dataset creation. And the annotation bias in data influences learning bias accumulated when training a classifier, so the risk of the annotators' bias inheritance increases. In the case of hate speech classification, there is a risk of unintended identity-based bias. For example, non-hateful texts containing mentions of gender, nationality or other protected attributes can be classified as a harmful utterances. The cases of biased decision-making are governed by law. For example, the social media platforms that signed the EU hate speech code [1] have to delete posts using offensive and inappropriate language within 24 hours. Given the number of everyday posts to check, automated moderation system feedback delay is highly restricted. For that reason, accelerated and compressed models receive more attention for the task.

Our paper presents one of the first attempts to analyze biased outcomes of compression in the context of hate and offensive language detection. In particular, we analyze the impact of encoder layer pruning in pre-trained Transformer Language Models (LMs, in short). Removing layers does not require additional fine-tuning and allows for explaining the contribution of the encoder blocks to model decision-making. We analyse the layers' contribution to rational model decision-making in terms of performance and fairness.[1]

The main contributions of this work are the following: *(i)* We measure identity-based bias in pruned Transformer LMs. *(ii)* We study which group of encoder layers (bottom, middle or upper) can be efficiently pruned without biased outcomes. *(iii)* We propose word-level supervision in pruned Transformer LMs as a debiasing method.

The paper is organized as follows. First, we report an analysis of related literature in Section 2. Section 3 provides the definition of pruning strategies, supervised token-wise attention learning methodology, and a list of evaluation criteria[2]. Section 4 provides the results and analysis of bias evaluation in compressed models.

2 Related Work

Inappropriate language with identity-targeted insults posted online provokes the dissemination of stereotypes about minority members [9]. To prevent hate and offensive language from being posted, automated hate speech detectors and filters are used [3]. The detectors vary depending on the task solved, such as profanity, individual cyberbullying, sexism, harassment, and othering language recognition.

Early research works approach the tasks using statistical and machine learning models trained on a suite of linguistic features extracted from text [23,7,21]. Recently, pre-trained Transformer LMs predominated over conventional machine learning methods [13]. Despite being efficient in a range of tasks associated with

[1] The implementation of the experiments can be found at https://github.com/upunaprosk/fair-pruning.

[2] In our work, we use token-wise and word-level supervision interchangeably.

hate speech classification, Transformer LMs can lack generalisation ability, increasing the risks of unintended bias [24,19]. There is little research studying whether compression could amplify bias, though novel model compression techniques in NLP are constantly being developed. Compression can be achieved, for instance, through pruning some parts of the Transformer LMs: neurons, heads, layers [4,14].

At the same time, in other fields, recent research shows that even when compressed models perform on par with the baselines, the predictions of pruned models can become considerably disproportionate and skewed. For example, the image features underrepresented in the training data could be misclassified by the compressed models [5].

To the best of our knowledge, our work is one of the first attempts to analyse bias amplification in compressed models in the context of a hate speech classification task. We transfer the hypothesis from the related work [5] to a compression impact study in Transformer LMs: if the impact of compression is uniform, then the shift in scores achieved on the texts mentioning a target community t should also be uniform compared to the overall scores shift β. That forms our null hypothesis H_0:

$$\begin{aligned} H_0 &: \beta_0^t - \beta_0 = \beta_c^t - \beta_c \\ H_1 &: \beta_0^t - \beta_0 \neq \beta_c^t - \beta_c, \end{aligned} \tag{1}$$

where β is an overall score, the superscript t is used to denote the score on texts mentioning community t, the subscript 0 is used for the scores of non-pruned full models, and the subscript c is used to denote the compressed models. We use fairness-related measures as β. We use the Wilcoxon Mann Whitney test to decide whether to accept the null hypothesis or an alternative one H_1, that the compression is not uniform and there is a relative difference in fairness for particular target subgroup t across 10 experiment runs.

3 Methodology

We approach the hate speech detection problem as a supervised multi-class classification with three classes: hate, offensive, and neutral. In this section, we first elaborate on Transformer LMs background and our pruning techniques and explain the motivation behind these compression strategies. Afterwards, we describe the experimental setup, including data, baselines, and evaluation criteria.

3.1 Transformer Background and Models

BERT is a Transformer LM known for achieving state-of-the-art results in various tasks, including hate language detection [12]. The BERT model configuration is defined by the number of encoder layers L and attention heads H. Each attention head receives a matrix $X_{n\times d}$ as an input with row-wise token representation, where n is the number of tokens in the input sequence, and d is the

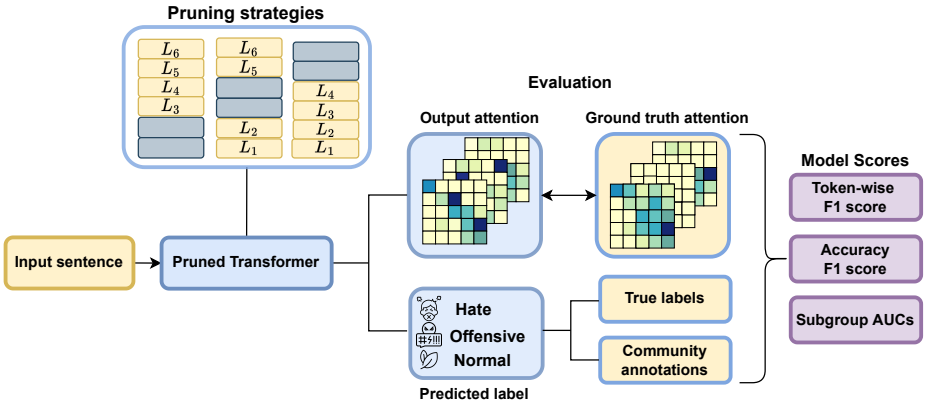


Fig. 1. End-to-end experimental pipeline. We prune the model by removing the layers, then use output attentions and predicted labels to evaluate Token F1 score, Accuracy/F1 scores, and Subgroup/BPSN/BNSP AUCs.

representation dimension. The output of the head is an updated matrix X_{out}:

$$X_{out} = W^A(XW^V),$$

where $W^A = \text{softmax}(\frac{(XW^Q)(XW^K)^T}{\sqrt{d}}) \in \mathbb{R}^{n \times n}$ is matrix with attention weights, and W^Q, W^V, W^K are projection matrices, the weights updated during the training. We consider a Transformer LM configuration defined by L encoder layers (blocks): $\{l_1, l_2, \ldots l_L\}$ and H attention heads.

3.2 Pruning Techniques

Following recently proposed pruning approaches, allowing for probing the importance of the layers [17], we explore six layer removal strategies: top, bottom, symmetric, alternate (odd and even), and contribution-based. Finally, we prune K of the layers selected via the pruning strategy, where $K = 2, 4, 6$ for architectures with $L = 12$ layers and $K = 1, 2, 3$ for $L = 6$ layers models. The end-to-end pipeline of experiments is illustrated in Figure 1.

Each pruning strategy is motivated by the redundancy of the layer that shows the relevancy of linguistic signals that the layer brings up. The syntactic and semantic information from the text is captured between the middle and upper layers [15]. However, the latter are more affected by the fine-tuning [11] and can be indifferent to decision-making. Therefore, the *top* pruning strategy for removing K upper layers (i.e., close to the model output) from pre-trained models could prevent overfitting issues. Surface features of the text being captured in bottom layers are necessary for various text classification tasks, making these layers more prominent for efficient distillation [22]. Bottom layer removal can, thus, cause considerable performance loss. We still consider that strategy since

the consequences of such pruning need to be clarified regarding bias. Middle layers store syntax information of the text [6], but can hold redundant knowledge from the bottom and upper layers accumulated during the training. To study the importance of middle layers, we consider a symmetric layers removal strategy by keeping the X top and bottom layers and removing the K layers in the middle such that $2X + K = L$. Alternate pruning consists of removal K layers starting from the upper ones; for example, $\{9, 11\}$ (odd alternate) and $\{10, 12\}$ (even alternate) pruning, when $K = 2$ and $L = 12$. Alternate pruning is motivated by the similar attention matrices of the close layers. One of the two consecutive layers can be dropped since the other holds almost the same information about the input text.

Lastly, we also estimate each layer's contribution to the decision-making. Given an input text sequence s_i, we measure the contribution of the layer l with cosine similarity between an input and output representations of the [CLS]-token, corresponding to the input sequence representation:

$$\phi_{s_i}(l) = \cos(Z_{l-1}, Z_l),$$

where Z_l is a vector of hidden states of the layer l, corresponding to the [CLS]-token[3]. We average the values over the validation texts. We prune K layers for each model with the highest contribution scores. Based on the obtained contribution scores, we consider the following layers removal lists for the models: BERT $\{5, 10, 9, 7, 2, 4\}$, RoBERTa $\{1, 2, 6, 8, 9, 4\}$, DistilBERT $\{2, 3, 4\}$, DistilRoBERTa $\{6, 2, 3\}$.

The efficiency of pruning models following the observed strategies depends on the number of pruned layers. The ratio of removed layers decreases the number of model parameters, resulting in fine-tuning speed-up [17].

3.3 Debiasing Approach

For these experiments, we use attention weights W^A to interpret model-decision. We suggest a debiasing approach that prompts the model to assign the larger weights to truly important tokens for the prediction, i.e. word-level supervision. For that, we change the loss computed during the training:

$$Loss_{\sum} = Loss_{pred} + \lambda Loss_{attn}, \tag{2}$$

where $Loss_{pred}$ is conventional cross-entropy classification loss, $Loss_{attn}$ is attention loss, computed based on the rationales provided along with data annotations, and λ is a hyper-parameter regulating the contribution of attention loss to the overall loss. Here, we use ground truth attention for calculating attention loss, which we introduced above (2) that is also depicted in Figure 1. We calculate the difference between the final hidden state corresponding to [CLS]-token and ground truth attentions (rationales). At the same time, using ground truth

[3] That token is used for classification in Transformer LMs.

attention makes the model focus on truly important tokens (ground truth rationales) for classification, reducing bias in models [10]. So, we treat fine-tuning with supervised attention on training set to compensate for the knowledge lost during compression and simultaneously prevent biased outcomes in compressed models.

3.4 Experimental Setup

We use state-of-the-art Transformer LMs for our experiments: base uncased configurations ($L = 12$,$H = 12$) of BERT [22] and RoBERTa [8], and their distilled [18] versions ($L = 6$,$H = 12$): DistilBERT, DistilRoBERTa. As the baselines, we use LMs fine-tuned for ten epochs with the batch size 16 and learning rate $2 \cdot 10^{-5}$ on training data. We use the benchmark dataset for explainable offensive and hate language detection HateXplain [10]. That dataset contains 20,148 posts collected from Twitter and Gab, each labelled as hateful, offensive, or normal. The dataset was annotated through crowdsourcing and contains extra annotations: hate and offence target communities and textual highlights, marked by annotators as reasoning for decision labelling, i.e. rationales. Rationales are represented as binary arrays, with one corresponding to the words marked by annotators as the ones influencing their labelling decision (offensive, hate or normal) and 0 for the rest of the words. To our knowledge, no other datasets have a similar range of annotations. For the experiments devoted to debiasing, we consider the following ranges of hyper-parameter, regulating the contribution of attention loss to the overall loss: $\lambda \in \{10^{\{-2,-1,0\}}\}$.

3.5 Evaluation

We use the train, development and test stratified split provided along with the dataset for the three following steps: models fine-tuning (train), hyper-parameters search (development), and evaluation (test).

We use a suite of evaluation metrics when establishing the baselines [10]. We report accuracy and macro F1-score reflecting the ability of the model to distinguish between hate, offensive and normal classes.

We measure identity-based bias in pruned models with the threshold-agnostic fairness metrics first introduced in [2]. These measures are AUC scores on the selected subset of the data. In particular, the data is divided into four domains: D^+, D_t^+, D^-, and D_t^-, where D^+ are posts labelled as hateful or offensive, D^- are normal posts, and D_t are the posts mentioning target community t. We use the following metrics: (1) Subgroup AUC = AUC($D_t^+ \cup D_t^-$), (2) Background Positive Subgroup Negative BPSN = AUC($D_{\setminus t}^+ \cup D_t^-$), and (3) Background Negative Subgroup Positive BNSP = AUC($D_{\setminus t}^- \cup D_t^+$). Here, Background refers to the texts not mentioning the community t. BPSN (BNSP) measures the false-positive (false-negative) rates for the texts mentioning target community t. We report aggregated scores for communities computed with Generalized Mean of

Bias (GMB):

$$GMB\left(m_t\right) = \left(\frac{1}{N}\sum_{s=1}^{N} m_t^p\right)^{\frac{1}{p}},$$

where m_t is a bias metric calculated for community t, N is a number of communities, and p is a constant exponent. We set $p = -5$ to emphasize the contribution of the lowest values m_t to the generalized score. The value p that is used is the same as the one used by the authors of the dataset [10].

Lastly, we estimate whether the models focus on relevant context when making the predictions. For that, we compare the context marked by annotators as influencing their class labelling decision, i.e. aforementioned (ground truth) rationales, and the model output rationales (Figure 1). As for the model output rationales, we select top-5 tokens with the largest attention weights. Given ground truth rationales, we compute the token F1-scores calculation based on precision and recall for model output rationales. The token F1 score refers to the plausibility suite of metrics [10].

4 Results

4.1 Pruning Impact

We find a typical pattern across the layer removal strategies: pruning leads to unintended identity-based bias, and the risks of unethical predictions increase with the ratio of pruned weights. Furthermore, layer removal provokes statistically significant differences in community-level fairness between a range of compressed and non-compressed models. Table 1 reports the results obtained for different models when pruning upper layers. The token F1 scores are low; the rationales annotation procedure can explain that. Most tokens can be labelled as 0, including articles, prepositions and other probably related tokens to the hate span. Low per-token alignment between predicted and ground truth rationales also provokes high variance in the Token F1 scores.

We find similar trends according to fairness loss between different pruning strategies and present results only for the upper layers of pruning. We observe that the disparate effect of pruning on a target-level basis is less common for BERT than for RoBERTa. For BERT, the maximum number of target communities with statistically significant difference scores shift is 4 (out of 10 most frequent communities in the data). In contrast, for RoBERTa, that number is maximum and equal to 6. DistilBERT is more robust to pruning in terms of both fairness and performance. DistilRoBERTa is also less sensitive to pruning, but only in terms of performance. We also find that the disproportionate effect of pruning takes place even when maintaining up to 90% of the original performance (for instance, that is the case of DistilBERT with 3/6 layers removed). That shows that there is also another side of pruning: performance loss does not necessarily go along with fairness loss.

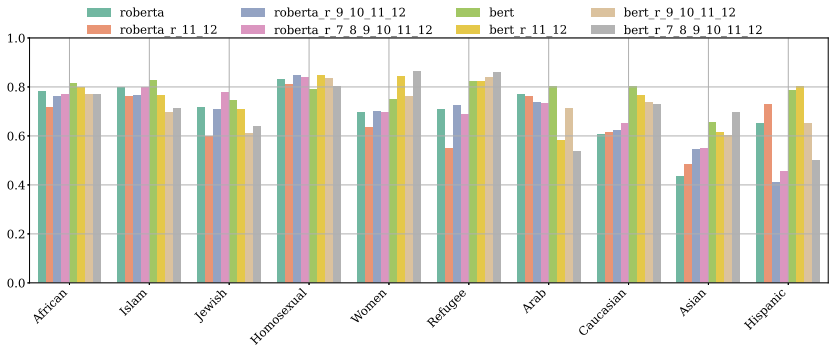
Table 1. Performance of original and pruned models on HATEXPLAIN test set. Layers correspond to the number of **upper** layers left. For the pruned models, we report the number of target communities for which the assumption H_0, formulated in (1), of compression uniform impact, is rejected, which means the compression has increased the biases.

Model	Layers	F1 score	Token F1 score	Count Signif Target Classes		
				Subgroup	BNSP	BPSN
BERT	12/12	$67.28_{\pm 0.13}$	$48.58_{\pm 3.28}$	-	-	-
	10/12	$65.31_{\pm 0.17}$	$38.35_{\pm 4.11}$	2	0	1
	8/12	$64.82_{\pm 0.15}$	$32.57_{\pm 4.06}$	2	0	2
	6/12	$63.46_{\pm 0.21}$	$34.4_{\pm 3.87}$	4	0	2
DistilBERT	6/6	$66.19_{\pm 0.44}$	$43.31_{\pm 3.42}$	-	-	-
	5/6	$66.08_{\pm 0.62}$	$42.77_{\pm 4.13}$	0	0	0
	4/6	$65.66_{\pm 0.51}$	$42.1_{\pm 3.98}$	3	0	1
	3/6	$64.31_{\pm 0.83}$	$39.81_{\pm 4.22}$	3	1	2
RoBERTa	12/12	$83.42_{\pm 0.4}$	$46.64_{\pm 3.51}$	-	-	-
	10/12	$81.46_{\pm 0.41}$	$39.37_{\pm 4.61}$	4	2	2
	8/12	$78.67_{\pm 0.58}$	$38.49_{\pm 4.23}$	6	3	4
	6/12	$77.08_{\pm 0.33}$	$24.47_{\pm 4.08}$	6	5	5
DistilRoBERTa	6/6	$82.02_{\pm 0.36}$	$42.08_{\pm 5.24}$	-	-	-
	5/6	$81.08_{\pm 0.4}$	$33.2_{\pm 4.75}$	3	0	2
	4/6	$77.06_{\pm 0.48}$	$32.76_{\pm 5.21}$	3	2	4
	3/6	$74.05_{\pm 0.43}$	$32.6_{\pm 4.61}$	6	5	6

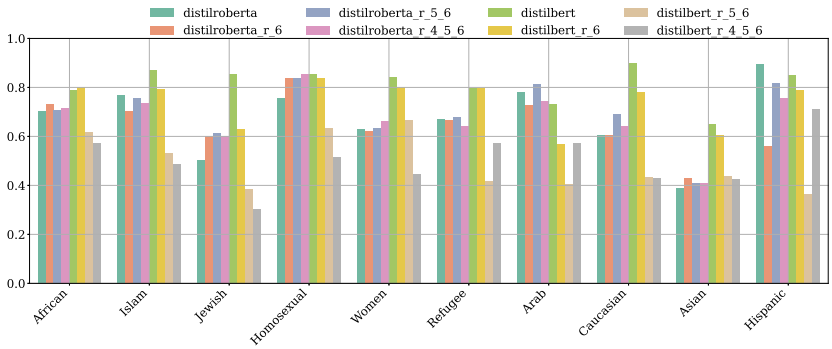
In Figure 2a, we plot BERT and RoBERTa Subgroup AUC scores for the ten most frequent communities in data. We find that pruning disproportionately affects some subgroups. For example, for RoBERTa with two last layers removed, there is a subgroup AUC score gain for some subgroups compared to the original model (Asian, Hispanic); for other cases, the score decreases considerably (Jewish, Refugee). At the same time, for BERT, the results are mostly stable, except for Women, Arab, and a few other subgroups. We also observe that there is sometimes an improvement between compressed and non-compressed BERT and RoBERTa models. We suggest that this is due to the dynamics of fine-tuning: some layers could learn wrong features from text and add bias. The results for distilled models are displayed in Figure 2b. The general trend is the same for distilled models: fairness steadily decreases with an increase in the number of removed layers. Figure 3 shows Subgroup AUC scores when removing bottom layers. We do not report results for other pruning techniques for the lack of space. The general pattern of fairness loss is the same for the bottom layer pruning strategy.

4.2 Debiasing with Word-level Supervision

The reported token F1 scores (Table 1, column 3) drop with an increasing number of pruned layers across all the models. That means that pruned models



(a) BERT/RoBERTa



(b) DistilBERT/DistilRoBERTa

Fig. 2. Community-wise Subgroup AUC scores on HateXplain test set. $r*$ = set of **upper** removed layers.

pay less attention to important contexts when making predictions. Recall that critical context is defined by ground truth rationales provided along with data annotations. We suppose that supervised attention learning can compensate for that loss during fine-tuning of the pruned model. We conduct the experiments on the models with the maximum of layers removed: pruned BERT with $L = 6$ and RoBERTa and distilled models with $L = 3$. Table 2 and Table 3 report fairness scores obtained for the models when pruning the upper and bottom layers. We present the scores for two strategies for the lack of space; the scores obtained when pruning other layers fall under the conclusion we draw.

We find that supervised attention reduces bias for all the models; the fairness improvement is substantial for non-distilled models: +0.172 for RoBERTa and +0.213 for pruned BERT when using $\lambda = 1$ (in comparison to models trained without attention learning). However, the performance loss is substantial for values greater than 1, so we do not report that result. For distilled models, the maximum improvements are +0.028 for DistilBERT and +0.03 for DistilRoBERTa.

Table 2. Performance and fairness scores (Subgroup AUC) of models trained with word-level supervision. The numbers in parentheses represent the ratio of the layers preserved when pruning **upper** layers. $\lambda = 0$ stands for non-supervised attention learning.

Model	λ	F1 score	Token F1 score	Subgroup AUC
BERT (6/12)	0	$63.46_{\pm 0.21}$	$34.4_{\pm 3.87}$	$0.59_{\pm 0.01}$
	0.01	$65.12_{\pm 0.38}$	$36.3_{\pm 4.01}$	$0.707_{\pm 0.11}$
	0.1	$65.92_{\pm 0.24}$	$39.26_{\pm 3.91}$	$0.784_{\pm 0.07}$
	1	$66.61_{\pm 0.17}$	$45.54_{\pm 3.29}$	$0.803_{\pm 0.12}$
DistilBERT (3/6)	0	$64.31_{\pm 0.83}$	$39.81_{\pm 4.22}$	$0.768_{\pm 0.24}$
	0.01	$64.35_{\pm 0.51}$	$40.4_{\pm 3.04}$	$0.748_{\pm 0.16}$
	0.1	$65.11_{\pm 0.7}$	$41.03_{\pm 3.28}$	$0.794_{\pm 0.31}$
	1	$66.71_{\pm 0.22}$	$42.67_{\pm 3.14}$	$0.796_{\pm 0.28}$
RoBERTa (6/12)	0	$77.08_{\pm 0.33}$	$24.47_{\pm 4.08}$	$0.519_{\pm 0.21}$
	0.01	$80.86_{\pm 0.22}$	$33.19_{\pm 3.28}$	$0.612_{\pm 0.29}$
	0.1	$78.58_{\pm 0.23}$	$36.49_{\pm 4.11}$	$0.681_{\pm 0.17}$
	1	$82.38_{\pm 0.26}$	$40.52_{\pm 3.81}$	$0.691_{\pm 0.14}$
DistilRoBERTa (3/6)	0	$71.05_{\pm 0.43}$	$32.6_{\pm 4.61}$	$0.62_{\pm 0.08}$
	0.01	$79.14_{\pm 0.47}$	$34.41_{\pm 4.11}$	$0.634_{\pm 0.04}$
	0.1	$81.25_{\pm 0.33}$	$36.51_{\pm 3.5}$	$0.635_{\pm 0.08}$
	1	$81.96_{\pm 0.51}$	$43.02_{\pm 4.14}$	$0.65_{\pm 0.09}$

We also report F1 scores showing how supervised attention learning improves performance, similar to fairness increase. The scores are on par with the baselines when using $\lambda = 1$. We show that the debiasing conducted via supervised attention learning improves all models' fairness scores.

5 Conclusion

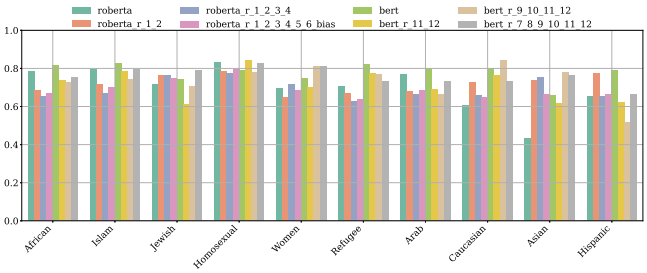
In this work, we conducted two chains of experiments to analyse the effect of Transformer LMs pruning in the context of hate speech classification tasks. We performed the experiments on a dataset containing Twitter and Gab data. First, we analysed the effect of pruning in terms of both fairness and performance loss for BERT, RoBERTa, and their distilled versions. We also estimated to which extent the pruned models rely on relevant context when making predictions. Our results show that removing any layer from Transformer LMs results in fairness loss even when the performance loss could be negligible. We statistically prove that there is a deviation in target community-level predictions when removing the layers from the models. Second, we conduct supervised attention-learning experiments to reduce bias in pruned models. Our results show that fairness score improvement depends on the hyper-parameter regulating the addition of attention loss to the overall loss. The pruned models achieve the best scores when $\lambda = 1$.

From the theoretical perspective, our work suggests a new research direction, focusing on fairness loss that should not be ignored when designing and

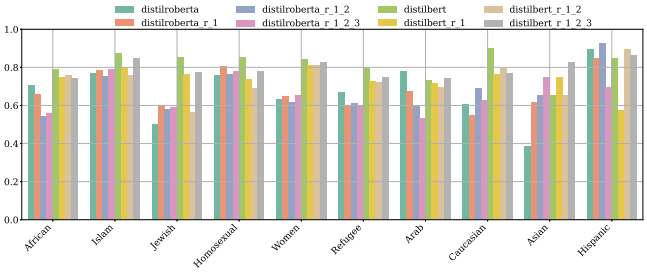
evaluating compressed models, including the classification task. We also suggest using supervised attention learning to compensate for the knowledge lost for pruned models. That correspondingly highlights the usefulness of relevant context annotations when designing the dataset.

The main limitations of our work are caused by the scope of data we use to study compression impact. Due to the demand for other datasets with similar fine-grained supervision, we are working on building new datasets in other languages. Future work may focus on compression impacts study (tensor decomposition, quantization, parameters sharing) and other debiasing techniques. The latter can be applied to the original model before the compression to estimate the consequences of initial bias in compressed versions. When compared to other debiasing approaches, the results of current research may serve as the baselines. The results of our work can also be used for further linguistic analysis, focusing on functional attributes of text [16].

Appendix



(a) BERT/RoBERTa



(b) DistilBERT/DistilRoBERTa

Fig. 3. Community-wise Subgroup AUC scores on HateXplain test set. $r*$ = set of **bottom** removed layers.

Table 3. Performance and fairness scores (Subgroup AUC) of models trained with word-level supervision. The numbers in parentheses represent the ratio of the layers preserved when pruning **bottom** layers. $\lambda = 0$ stands for non-supervised attention learning.

Model	λ	F1 score	Token F1 score	Subgroup AUC
BERT (6/12)	0	$62.97_{\pm 0.11}$	$30.5_{\pm 5.02}$	$0.52_{\pm 0.09}$
	0.01	$62.5_{\pm 0.18}$	$33.2_{\pm 4.67}$	$0.54_{\pm 0.07}$
	0.1	$63.25_{\pm 0.24}$	$34.05_{\pm 4.47}$	$0.591_{\pm 0.12}$
	1	$65.93_{\pm 0.26}$	$35.77_{\pm 3.88}$	$0.692_{\pm 0.54}$
DistilBERT (3/6)	0	$64.22_{\pm 0.36}$	$37.18_{\pm 4.04}$	$0.738_{\pm 0.17}$
	0.01	$63.08_{\pm 0.27}$	$38.07_{\pm 4.71}$	$0.736_{\pm 0.23}$
	0.1	$63.32_{\pm 0.4}$	$40.11_{\pm 3.96}$	$0.75_{\pm 0.09}$
	1	$64.1_{\pm 0.28}$	$40.05_{\pm 2.88}$	$0.791_{\pm 0.22}$
RoBERTa (6/12)	0	$78.18_{\pm 0.32}$	$25.32_{\pm 4.51}$	$0.683_{\pm 0.31}$
	0.01	$78.77_{\pm 0.29}$	$29.9_{\pm 4.42}$	$0.669_{\pm 0.34}$
	0.1	$78.92_{\pm 0.35}$	$31.54_{\pm 4.06}$	$0.684_{\pm 0.28}$
	1	$79.98_{\pm 0.32}$	$39.06_{2.88}$	$0.693_{\pm 0.31}$
DistilRoBERTa (3/6)	0	$78.13_{\pm 0.48}$	$34.18_{\pm 3.85}$	$0.625_{\pm 0.27}$
	0.01	$77.05_{\pm 0.53}$	$36.05_{\pm 4.06}$	$0.618_{\pm 0.14}$
	0.1	$78.21_{\pm 0.41}$	$43.61_{\pm 3.92}$	$0.626_{\pm 0.11}$
	1	$78.83_{\pm 0.36}$	$44.5_{\pm 2.92}$	$0.643_{\pm 0.15}$

References

1. Bisht, A., Singh, A., Bhadauria, H., Virmani, J., et al.: Detection of hate speech and offensive language in twitter data using lstm model. In: Recent trends in image and signal processing in computer vision, pp. 243–264. Springer (2020)
2. Borkan, D., Dixon, L., Sorensen, J.S., Thain, N., Vasserman, L.: Nuanced metrics for measuring unintended bias with real data for text classification. Companion Proceedings of The 2019 World Wide Web Conference (2019)
3. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR) **51**(4), 1–30 (2018)
4. Gupta, M., Varma, V., Damani, S., Narahari, K.N.: Compression of deep learning models for nlp. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. p. 3507–3508. CIKM '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3340531.3412171
5. Hooker, S., Courville, A., Clark, G., Dauphin, Y., Frome, A.: What do compressed deep neural networks forget? arXiv preprint arXiv:1911.05248 (2019)
6. Jawahar, G., Sagot, B., Seddah, D.: What does BERT learn about the structure of language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3651–3657. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1356
7. Lima, L., Reis, J.C., Melo, P., Murai, F., Benevenuto, F.: Characterizing (un) moderated textual data in social systems. In: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 430–434. IEEE (2020)

8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
9. Maass, A., Cadinu, M.: Stereotype threat: When minority members underperform. European Review of Social Psychology **14**(1), 243–275 (jan 2003). https://doi.org/10.1080/10463280340000072
10. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: A benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14867–14875 (2021)
11. Merchant, A., Rahimtoroghi, E., Pavlick, E., Tenney, I.: What happens to bert embeddings during fine-tuning? arXiv preprint arXiv:2004.14448 (2020)
12. Mozafari, M., Farahbakhsh, R., Crespi, N.: Hate speech detection and racial bias mitigation in social media based on bert model. PloS one **15**(8), e0237861 (2020)
13. Mutanga, R.T., Naicker, N., Olugbara, O.O.: Hate speech detection in twitter using transformer methods. International Journal of Advanced Computer Science and Applications **11**(9) (2020)
14. Neill, J.O.: An overview of neural network compression. arXiv preprint arXiv:2006.03669 (2020)
15. Niu, J., Lu, W., Penn, G.: Does bert rediscover a classical nlp pipeline? In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 3143–3153 (2022)
16. Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., Pierrehumbert, J.B.: Hatecheck: Functional tests for hate speech detection models. arXiv preprint arXiv:2012.15606 (2020)
17. Sajjad, H., Dalvi, F., Durrani, N., Nakov, P.: Poor man's BERT: smaller and faster transformer models. CoRR **abs/2004.03844** (2020)
18. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019)
19. Soares, I.B., Wei, D., Ramamurthy, K.N., Singh, M., Yurochkin, M.: Your fairness may vary: Pretrained language model fairness in toxic text classification. In: Annual Meeting of the Association for Computational Linguistics (2022)
20. Steiger, M., Bharucha, T.J., Venkatagiri, S., Riedl, M.J., Lease, M.: The psychological well-being of content moderators. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM (may 2021). https://doi.org/10.1145/3411764.3445092
21. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop. pp. 88–93 (2016)
22. Xu, C., Zhou, W., Ge, T., Wei, F., Zhou, M.: Bert-of-theseus: Compressing bert by progressive module replacing. arXiv preprint arXiv:2002.02925 (2020)
23. Xu, J.M., Jun, K.S., Zhu, X., Bellmore, A.: Learning from bullying traces in social media. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies. pp. 656–666 (2012)
24. Yin, W., Zubiaga, A.: Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science **7**, e598 (2021)