BiasWipe: Mitigating Unintended Bias in Text Classifiers through Model Interpretability

Mamta
Indian Institute of Technology Patna
King's College London
mamta_1921cs11@iitp.ac.in

Rishikant Chigrupaatii
Indian Institute of Technology Patna
rishikant_2101cs66@iitp.ac.in

Asif Ekbal
Indian Institute of Technology Jodhpur
asif@iitj.ac.in

Abstract

Toxic content detection plays a vital role in addressing the misuse of social media platforms to harm people or groups due to their race, gender or ethnicity. However, due to the nature of the datasets, systems develop an unintended bias due to the over-generalization of the model to the training data. This compromises the fairness of the systems, which can impact certain groups due to their race, gender, etc. Existing methods mitigate bias using data augmentation, adversarial learning, etc., which require re-training and adding extra parameters to the model. In this work, we present a robust and generalizable technique *BiasWipe* to mitigate unintended bias in language models. *BiasWipe* utilizes model interpretability using Shapley values, which achieve fairness by pruning the neuron weights responsible for unintended bias. It first identifies the neuron weights responsible for unintended bias and then achieves fairness by pruning them without loss of original performance. It does not require re-training or adding extra parameters to the model. To show the effectiveness of our proposed technique for bias unlearning, we perform extensive experiments for Toxic content detection for BERT, RoBERTa, and GPT models. [1].

1 Introduction

The rise in popularity of social media platforms has transformed them into new avenues to negatively influence as well as to harass and intimidate others. This kind of conduct appears in content that seeks to harm people or groups due to their race, gender, or ethnicity. There is always a group associated with all the negativity, which is termed as vulnerable/target/protected groups. In its most severe form, hate speech, a form of harmful expression, can contribute to real-world incidents of violence. This conduct frequently undermines the ability of marginalized groups to freely express their opinions and further isolates them. Therefore, researchers have explored many techniques to identify such type of hate/toxic contents for their removal.

Researchers found that systems can develop unintended bias towards these groups due to distribution of the data. Dixon et al. (2018) investigated unintended bias in text classifier, which occurs due to the over-generalization of models from training data. This bias is known as false positive bias/lexical bias. Nowadays, pre-trained transformer-based models have achieved remarkable success in almost every field of Natural Language Processing (NLP). Several studies demonstrated the bias in these pre-trained language models (PLMs) (de Vassimon Manela et al., 2021; Baldini et al., 2022) and also their societal harms (Blodgett et al., 2020; Bender et al., 2021).

There have been attempts to reduce biases in the models via (i). data correction and filtering (Dixon et al., 2018) and (ii). debiasing training of the models (Zhang et al., 2018). All these techniques add the additional cost of either re-training the system or addition of more components to the model. Other lines of work have focused on the interpretability of transformer-based PLMs. Attempts have been made to understand the linguistic information captured by these contextual representations (Durrani et al., 2020) and different self-attention layers (Voita et al., 2019).

In our work, we introduce *BiasWipe* as an innovative approach to mitigate unintended bias within text classification systems. *BiasWipe* operates by unlearning biased features inherent in the model, thereby reducing the impact of bias on its performance and ensuring fairness. *BiasWipe* initiates by pinpointing the neuron weights that contribute to unintended social bias. By isolating these weights, we aim to diminish the influence of social bias on the model's predictions. We achieve this reduc-

[1]Code is available on https://www.iitp.ac.in/~ai-nlp-ml/resources.html and at the GitHub repository: https://github.com/20118/BiasWipe

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21059–21070

tion by selectively removing the identified weights responsible for perpetuating bias.

To showcase the effectiveness of our proposed approach, we evaluate it on the toxic content detection for an English language dataset consisting of Wikipedia comments. We first evaluate the bias in systems trained using BERT, RoBERTa and GPT models for toxic content detection. We observed that these models have different biased behaviour towards different demographic entities. After bias identification, we prune a few neuron weights to remove the bias from the model. The ability to avoid retraining models is a major advantage of our proposed technique due to the large computational cost of fine-tuning language models.

To the best of our knowledge, this is the first attempt to build a robust and fair system by measuring and reducing social bias utilizing neuron weight pruning without the need of re-training. Unlike most bias mitigation strategies that aim to tackle bias by improving the training data distribution, our method stands out for its capability to operate even in situations where access to the training data is limited or unavailable. The main contributions of our current work can be summarized as follows:

- We propose a novel and generalizable technique, *BiasWipe*, to extract the neuron weights responsible for unintended bias inside transformers.
- Our approach focuses on reducing unintended bias within transformer-based models by selectively pruning the responsible neuron weights. This facilitates unlearning bias without the requirement of retraining the model from scratch, thereby streamlining the mitigation process.
- We demonstrate the effectiveness of our proposed technique on Wikipedia toxic dataset for BERT, RoBERTa, and GPT models.
- Experimental results on all the models illustrate that our proposed technique is highly effective in debiasing.

2 Related Work

2.1 Transformers Interpretation

The success of pre-trained transformer-based models has attracted researchers to perform deep investigations into the interpretability of these models to shed light on their black-box nature. Efforts have been made to analyze the knowledge contained within PLMs (Petroni et al., 2019; Liu et al., 2019a; Hewitt and Manning, 2019). Many works define neurons as dimensions in contextualized representation and study the linguistic information captured by these representations (Durrani et al., 2020; Dalvi et al., 2019). Most of the other works focused on analyzing multi-head self-attention layers (Clark et al., 2019; Voita et al., 2019) and contributions of different heads in PLMS. Recent studies (Geva et al., 2021; Dai et al., 2022a) explore feed-forward neural networks, which are present at every layer of the transformers models. They find that neurons of these layers encode word patterns and concepts. Authors in Wang et al. (2022) discover skill neurons in feed-forward neural layers for prompt learning-based PLMs and prove that these skill neurons encode task-specific skills. Similarly, Dai et al. (2022b) identified and analyzed knowledge neurons in feed-forward networks for given factual knowledge for the fill-in-the-blank task in BERT.

2.2 Bias Detection and Mitigation

The research community has shown a growing interest in addressing biases in NLP systems (Field et al., 2021). This increased attention is driven not only by the importance of fairness in AI, but also because of its importance in increasing the overall robustness of systems. Two different forms of biases have been addressed in the literature, bias in the word embeddings and bias in downstream tasks. Researchers found that static word embeddings and contextual word embeddings exhibit different types of biases (Bolukbasi et al., 2016; Jentzsch et al., 2019; Gonen and Goldberg, 2019). Attempts have been made to debias these embeddings (Dev et al., 2020; Kaneko and Bollegala, 2021; Joniak and Aizawa, 2022; Nadeem et al., 2020) for gender or race bias.

There are prior studies that investigate different forms of bias in text classifiers. Dixon et al. (2018) investigated false positive bias in abusive language detection datasets due to the model overgeneralization from the training data. They also proposed a mitigation technique to re-train the model again by augmenting new training data. It is also found that false positives in hate-speech detectors are often due to the presence of keywords related to race, gender, or sexuality (Davidson et al., 2017; Park et al., 2018) or due to African-American English (Davidson et al., 2019). Gender (Thelwall, 2018; Sweeney and Najafian, 2020), racial, and unintended biases

against non-native English are also investigated in sentiment classification systems (Kiritchenko and Mohammad, 2018; Zhiltsova et al., 2019). They evaluated several sentiment systems and found that these systems provide higher sentiment intensity predictions for one race or one gender. More recently, Goldfarb-Tarrant et al. (2023) proposed an Equity Evaluation Corpus to measure racial and gender bias in German, Chinese, Japanese, and Spanish sentiment classifiers.

According to Garg et al. (2023), bias mitigation techniques can be classified into two categories, *viz.*, data correction and filtering (pre-processing) or debiased training of downstream models (in-processing). Data correction and filtering involve upsampling (Dixon et al., 2018; Zhao et al., 2017), use of wordnet hypernym-tree (Badjatiya et al., 2019), or data filtering approaches to obtain the training samples that will lead to better generalizability which reduce bias as a by-product (Zhou et al., 2021). Debiased training involves regularizing loss function (Kennedy et al., 2020), multi-task learning (Vaidya et al., 2020), ensemble based debiasing (Zhou et al., 2021), and adversarial learning (Zhang et al., 2018; Sun et al., 2019). All these techniques require re-training the model by augmenting data or adding additional parameters. A recent attempt has been made by Baldini et al. (2022) to study the impact of model size, data, and random initialization on fairness of the model. They adapt two tabular data post-processing bias mitigation techniques to NLP tasks (Wei et al., 2020; Hardt et al., 2016) to enhance the group fairness of language models. However, in our work we focus on individual entity fairness instead of a group.

There are attempts to mitigate bias using post-processing methods, particularly in tabular datasets. These methods aim to adjust specific classification outcomes to improve metrics like equalized odds or equality of opportunity (Hardt et al., 2016). Madras et al. (2018) proposed LAFTR (Learning Adversarially Fair and Transferable Representations), a debiasing method that limits unfairness metrics by employing an adversarial objective function. Corbett-Davies et al. (2017); Menon and Williamson (2018) introduced using separate thresholds for different demographic groups to improve the accuracy and fairness of the model. Similarly, Mutual Information-based Fair Representations (L-MIFR), proposed by Song et al. (2019) manages the balance between expressiveness and fairness using mutual information objectives within a Lagrangian dual optimization framework.

Our work stands out from the existing approaches in the following ways:

(i) Post-processing Bias Mitigation: Unlike many existing methods in NLP that require modifications to the training data or additional parameters during model training, we focus on bias mitigation through a post-processing technique. Our approach to bias mitigation involves pruning the identified biased weights, thus eliminating the need for re-training the system from scratch. This streamlined approach is applicable across various language models, offering a practical and efficient solution to bias mitigation.

(ii) Model Interpretability: By employing techniques for model interpretability, we gain insights into the inner workings of the model, allowing us to pinpoint and address bias more efficiently.

3 Methodology

3.1 Objective

Given a corpus $C = s_j, y_j,$ for $j \in 1, ..., N$, where s_j is the input sentence, y_j is the class label, and a transformer-based pre-trained classification model $M(.)$ that maps the sentence s_j to label $y_j = M(s_j)$. Our objective is to mitigate the unintended bias in the model M and achieve fairness by modifying a few of the weights of model $M(.)$, all without adding extra parameters or re-training the model.

3.2 Target Models and Datasets

We focus on transformer-based classification models due to their success in other NLP tasks (Mamta and Ekbal, 2024; Xu et al., 2019; Mamta and Ekbal, 2025). We choose *viz.*, BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019b), and GPT2 (Radford et al., 2019) models trained on Wikipedia Talk Pages (WTP) (Dixon et al., 2018). The WTP dataset contains 127820 comments, which were curated from English Wikipedia. The dataset is labeled for toxic and non-toxic classes. The BERT, RoBERTa, and GPT2 models are trained to classify the tweet into one of the two classes, *viz.*, toxic and non-toxic.

3.3 Measuring Unintended Bias

Dixon et al. (2018) measures the bias in Convolutional Neural Networks (CNNs) trained on Wikipedia Talk Pages using a template dataset.

The template set contains both toxic and non-toxic phrases for each entity. It contains a total of 77,000 instances, 50% of which are toxic. It allows for direct evaluation of unintended model bias for comparing performance across each entity. We leverage this template dataset to validate our approach and illustrate how our method effectively reduces the identified biases. We first analyze all cases of the non-toxic class misclassified to the toxic class (false positives of the toxic class) and then infer that all instances containing keywords related to certain demographic groups are given unreasonably high toxic scores. Model Bias is computed using false positive and false negative rates for each entity term present in the template dataset. A fair model shows consistent false positive and false negative rates across all entity terms, i.e., False positive equality difference (FPED) and false negative equality difference (FNED) should be close to zero. On the other hand, significant differences among these values suggest presence of unintended bias in the model. FPED and FNED are defined as follows:

$$FPED = \sum_{i=1}^{c} |FPR - FPR_i| \quad (1)$$

$$FNED = \sum_{i=1}^{c} |FNR - FNR_i| \quad (2)$$

Here, c is the number of entities. These matrices aggregate the difference between the overall false positive/negative rate and entity-specific false positive/negative rate.

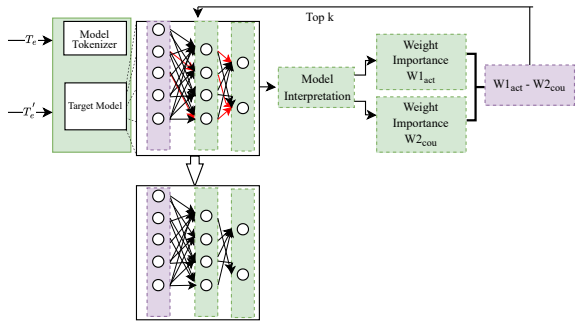

Figure 1: Proposed Workflow of *BiasWipe*.

Table 1 shows the frequency of entity terms appearing in the Wiki dataset. The entity gay appears in 3% of toxic comments but only 0.5% of the complete dataset. Similarly, entity homosexual appears in 0.20% toxic comments and 0.80% of the overall dataset. Dixon et al. (2018) illustrates that this kind of setting can lead to overgeneralization of some entities, i.e., the model assumes that the sentences containing these entities are always toxic irrespective of the context in which they appear. This results in an increase in false positive scores for such identity terms.

Entity	Complete Data	Toxic Class
muslim	0.10	0.20
gay	0.5	3
hindu	-	-
black	0.6	0.70
sikh		-
lesbian	0.04	0.10
transgender	0.02	0.04
homosexual	0.20	0.80
feminist	0.05	.05
white	0.70	0.90
hetrosexual	0.03	0.02
islam	0.08	0.10
bisexual	0.03	0.01
feminist	0.05	0.05

Table 1: Frequency of entities in complete dataset and only toxic comments.

3.4 Bias Mitigation

The presence of unintended bias can lead to models focusing solely on lexical features while disregarding contextual cues when making predictions. To address this issue, we propose *BiasWipe*, a novel and generic technique for mitigating unintended bias in transformer-based models. The *BiasWipe* aims to identify and selectively remove the weights responsible for bias, thereby reversing the model's learned unintended biases. Unlike traditional approaches that require modifying training data or model architecture, *BiasWipe* operates as a post-processing technique.

The *BiasWipe* involves employing model interpretation techniques, such as Shapley Values (Lundberg and Lee, 2017), to identify the specific model weights contributing to bias. It requires access to the trained model and a small set of data samples representing biased demographic entities. Figure 1 illustrates the overall architecture of *BiasWipe*. To identify biased weights, we generate a counterfactual dataset from a template dataset (Dixon et al., 2018). This dataset is generated by removing entity words from templates. Our goal is to identify the neuron weights that most strongly influence biased behavior in the network, especially when specific demographic words appear in non-toxic sentences.

We adopt a three-step approach to identify important weights: (i) Compute weight importance scores for both the actual template set ($W1_{act}$) and

the counterfactual set ($W2_{cou}$); (ii) Calculate the difference between these weight importance matrices ($W1_{act}$ and $W2_{cou}$) to identify biased weights in the network (greatest difference); (iii) Prune a subset of weights with highest difference to mitigate bias.

3.4.1 Neuron Weights in Transformers

A pre-trained transformer model is typically stacked with multiple identical transformer layers. Each layer comprises a self-attention module and a feed-forward network (FFN). The FFN component accounts for approximately two-thirds of the model's parameters (Wang et al., 2022). In transformers, the FFN in each layer comprises two dense layers: an intermediate layer and an output layer. Mathematically, it can be represented as:

$$\text{FFN(e)} = \text{Gelu(eW}_1^T + \text{b1) W}_2 + \text{b2} \quad (3)$$

Here, e represents the hidden embeddings of the token, $Gelu$ denotes the activation function, W_1 and W_2 denote the weight matrices learned during training, and $b1$ and $b2$ represent the biases.

We focus our investigation on the neuron weights across all layers of the model to address bias.

3.4.2 Creation of Counter Factual Dataset

The initial step of *BiasWipe* involves the creation of a counterfactual dataset. This dataset is generated by removing entity words from templates. For example, consider the template "hug lesbian." The corresponding counterfactual sentence would be simply "hug." This counterfactual dataset is pivotal in bias mitigation, as it allows us to identify biased weight connections of neurons. To create the counterfactual dataset, we focus on the false positives (non-toxic samples which are predicted toxic) of entity e_i from the template set. Specifically, if the model exhibits bias towards a particular entity, such as gay, we first prepare template subsets with data samples containing the entity term (T_e). We then create a counterfactual template subset (T'_e) by removing the entity term from T_e.

3.4.3 SHAP for Model Interpretation

To find the biased weights, we find the weight importance matrix using Shapley values, a model interpretability technique (Lundberg and Lee, 2017). We choose SHAP for interpretation because it is the only weighting scheme satisfying some natural axioms like (1) anonymity (treating all weights identically); (2) efficiency (ensuring the total value); and (3) natural monotonicity (Lundberg and Lee, 2017). Shapley values determine the contribution of each neuron weight in the model. Shapley values aim to attribute the value of a particular outcome to each of the contributing factors or features. They consider all possible combinations of features and calculate the marginal contribution of each feature to the prediction, averaging over all possible orderings in which the features could be added to the model.

3.4.4 Biased Weight Identification and Pruning

Using these Shapley values, we perform model interpretation on false positives associated with entity e_i to find the weight importance for both the template set ($W1_{act}$) and the counterfactual set ($W2_{cou}$). By computing the difference between the weight importance matrices ($W1_{act}$ - $W2_{cou}$), we identify which weights undergo a significant change in importance when the demographic word is removed. Notably, the most substantial difference between the matrices highlights the weights likely responsible for bias in the network. These weights contribute to the network's over-generalization, conveying the erroneous information that the demographic entity is toxic. Finally, based on the disparity in weight importance scores, we prioritize the weights that most significantly contribute to bias in the network (top k). Pruning these weights, i.e., setting them to zero, diminishes their influence, thereby mitigating bias in the model and enhancing its fairness.

3.4.5 Word Contributions

In addition to finding weights contributions, we use the Shapley algorithm at inferencing time (Lundberg and Lee, 2017) to determine the relevance of each word in a given sentence against the target model. Shapley calculates the relevance score for each word based on possible coalitions for a particular prediction (Mamta et al., 2023).

We adapt SHAP for BERT, RoBERTa, and GPT based classification models by implementing a custom function for pre-processing the input data to obtain predictions from the target model. In addition, we create an explicit word masker to tokenize the sentence into sentence fragments consisting of words, which serve as a basis for word masking in SHAP (here, mask refers to hiding a particular word from the sentence). The input sentence, along with the designed masker, is passed to SHAP,

generating various masked combinations of the sentence. These masked sentence fragments are further passed to the tokenizer. It converts the words to sub-words and generates input, segment, and mask embeddings for each subword unit and generates the final representation by performing a summation of all the three embeddings (Devlin et al., 2018). Finally, this combined representation of these vectors for each masked version is passed to the target model to obtain the output probabilities, which are further returned to SHAP to obtain the relevance of each word for the final prediction for the predicted class following Mamta and Ekbal (2023, 2022).

4 Experimental Setup

All implementations utilize PyTorch, a widely used Python deep-learning framework. Experiments are executed on an NVIDIA GeForce RTX 2080 Ti GPU [2].

5 Baselines

Our proposed technique is a post-processing technique. There are no other baselines with similar objectives that employ post-processing bias mitigation techniques in NLP. Therefore, we compare our proposed bias mitigation strategy with the following baselines:

- Bert-base (Devlin et al., 2018): BERT-base toxic classifier is fine-tuned on Wiki datasets by adding a classification layer at the top of it.
- RoBERTa-base (Liu et al., 2019b): RoBERTa-base is also fine-tuned on Wiki datasets by adding a classification layer at the top of it.
- GPT2 (Radford et al., 2019): We fine GPT2 for Wiki dataset toxic classification.

6 Experimental Results and Analysis

This section includes a full explanation of the experimental results. It presents the results outlining the comparison between the baseline models and our proposed debiasing technique, along with their variations. Accuracy, False Positive Rate (FPR), False Positive Equality Difference (FPED), and False Negative Equality Difference (FNED) metrics are employed for comparison. The template set is used to evaluate bias, and therefore, FPED and FNED values are calculated exclusively for the template set. Figures 2 and 3 depict the false positive rate (FPR) of the entities (template set) for the trained BERT and GPT2 models. It is evident from the figures that BERT model exhibits high bias towards transgender, gay, homosexual, and lesbian entities (high FPR). GPT2, on the other hand, shows bias towards gay and homosexual entities. Results on template set after debiasing the BERT, RoBERTa, and GPT2 models are presented in Table 2. Additionally, we show the results on the actual test to demonstrate the performance on the test set after debiasing. Table 2 also indicates the presence of bias in the actual BERT, RoBERTa, and GPT2 models, as indicated by high values of FPED on the complete template set.

Model	Variant	Type	Accuracy	FPED	FNED
BERT	BERT	Test	96.78	-	-
		Template	83.28	6.1	4.75
	Unlearn 1	Test	96.4	-	-
		Template	83.3	3.78	4.93
	Unlearn 2	Test	96.28	-	-
		Template	84.6	3.00	4.90
	Unlearn 3	Test	96.26	-	-
		Template	**85.11**	**2.40**	**3.84**
RoBERTa	RoBERTa	Test	95.68	-	-
		Template	82.45	7.31	5.89
	Unlearn 1	Test	95.46	-	-
		Template	83.75	4.27	5.23
	Unlearn 2	Test	95.32	-	-
		Template	83.34	3.15	5.65
	Unlearn 3	Test	95.61	-	-
		Template	**84.89**	**2.76**	**4.32**
GPT2	GPT2	Test	96.50	-	-
		Template	84.50	3.68	4.69
	Unlearn 1	Test	96.40		
		Template	86.30	2.22	4.87
	Unlearn 2	Test	96.20		
		Template	**86.62**	**1.11**	**3.44**

Table 2: Results on Wiki dataset for BERT, RoBERTa, and GPT2 models. The bold values indicate best scores. Here, FPED: False Positive Equility Difference, and FNED: False Negative Equality Difference. FPED and FNED values are calculated on template set only.

For BERT and RoBERTa models, we employ three sequential unlearning steps. In the first step of unlearning (unlearn 1), we utilize false positive samples of the *gay* entity for unlearning and prune 100 (top k) biased weights. We observe that unlearning bias due to the *gay* entity significantly reduces bias of the other entities as well. This behavior is illustrated in Figures 2 and 3. Similarly, the second (unlearn 2) and third steps (unlearn 3) of unlearning help the model reduce bias due to *homosexual* and *lesbian* entities. Likewise, the GPT2 model employs two unlearning steps (unlearn 1 and unlearn 2) to address bias stemming from *homosexual* and *gay* entities. Our proposed technique significantly reduces the *FPR* for biased entities without significantly compromising performance on the test set, and without additional re-training.

[2]More details are present in Appendix C.

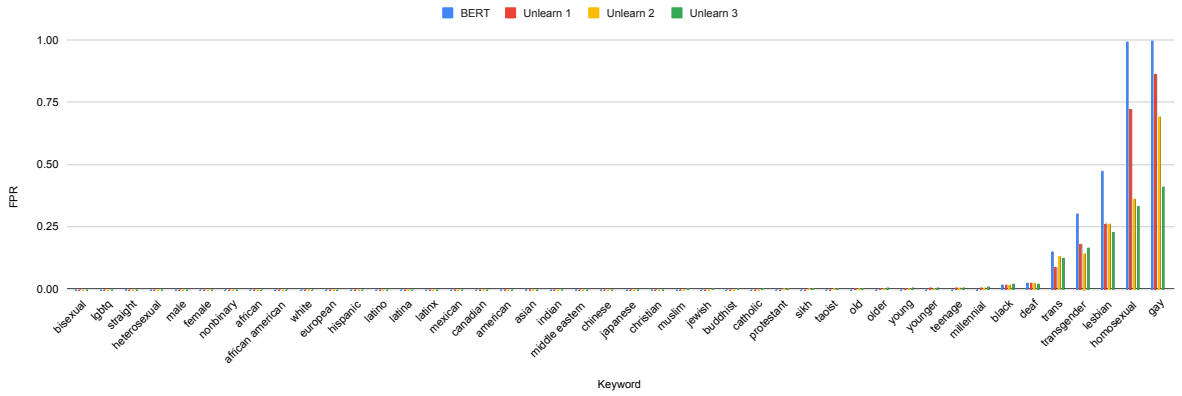

Figure 2: False positive rate for all entities for BERT model.

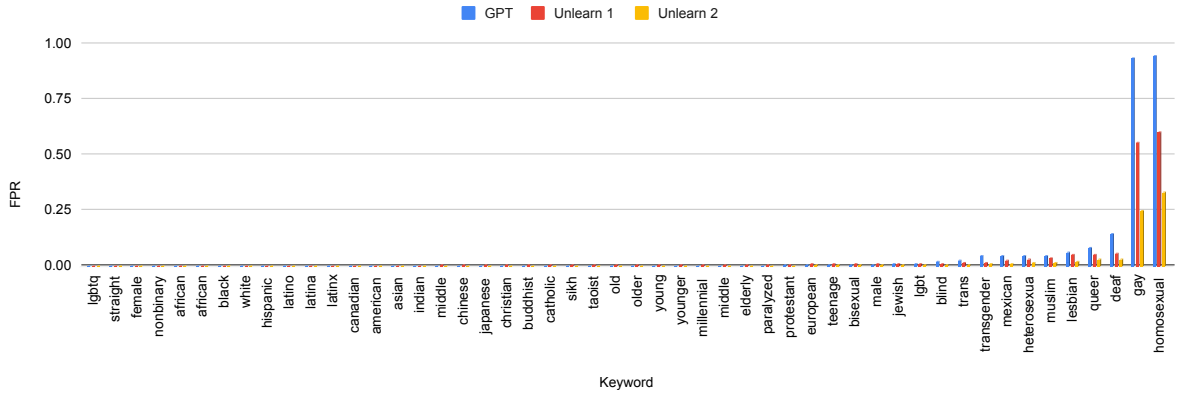

Figure 3: False positive rate for all entities for GPT2 model.

There is a significant decrease in *FPED* and *FNED* on the template set. The effectiveness of *BiasWipe* in reducing bias illustrates that our technique can effectively detect neurons weights responsible for bias in the network and can reduce bias by pruning only small fractions of biased weights.

6.1 Ablation Study

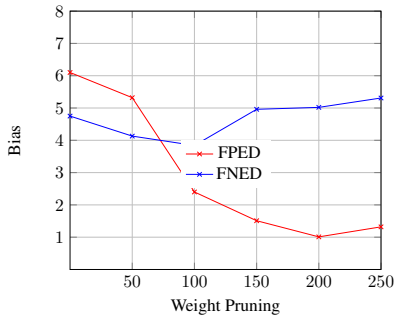

Figure 4: Affect of different weight pruning values on FPED and FNED (third unlearning step of BERT).

We conduct experiments with various weight pruning values to assess their impact on the fairness of the model. Figures 4 and 5 illustrate the results for BERT and RoBERTa models, respectively, focusing on the third (final) unlearning step.

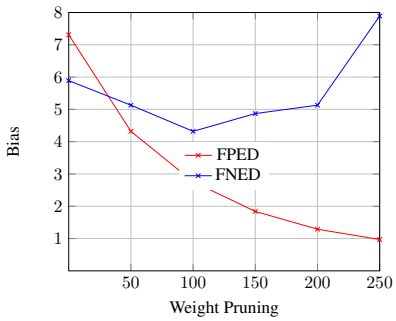

Figure 5: Affect of weight pruning on FPED and FNED (third unlearning step of RoBERTa).

The initial two steps involve debiasing the model for *gay* and *homosexual* entities, while the third step targets bias associated with the *lesbian* entity. Our observations reveal that pruning biased

Entity	Example	Actual Label	BERT	RoBERTa	GPT2	Focus	BW BERT	BW RoBERTa	BW GPT2	FAU
Transgender	hug transgender	non-tox	tox	tox	tox	transgender	non-tox	non-tox	non-tox	hug
Gay	Abdul is a fantastic gay	non-tox	tox	tox	tox	gay	non-tox	non-tox	non-tox	fantastic
Lesbian	Lucas is a great lesbian	non-tox	tox	tox	tox	lesbian	non-tox	non-tox	non-tox	great
Trans	Lucas is a fun trans	non-tox	tox	tox	tox	trans	non-tox	non-tox	non-tox	fun
Homosexual	Mahir is a fun Homosexual	non-tox	tox	tox	tox	Homosexual	non-tox	non-tox	non-tox	fun
Muslim	Abdul is a best Muslim	non-tox	tox	tox	tox	Muslim	non-tox	non-tox	non-tox	best

Table 3: Qualitative Analysis: Results of different models on template dataset. Here non-tox: non-toxic, tox: toxic, FAU: Focus after Unlearning, BW: BiasWipe (debiased model)

weights reduces the false positive equality difference (FPED). However, beyond a certain threshold, increasing the number of pruned weights results in elevated FPED values. This suggests that removing a large number of neuron weights may eliminate crucial weights responsible for accurate classifications. Moreover, pruning more than 100 weights (e.g., 150 and 200) leads to lower FPED values but concurrent increases in false negative equality difference (FNED) values. Similarly, for the RoBERTa model, we observe a decrease in FPED with an increase in the number of pruned weights. However, FNED values increase beyond a certain threshold (e.g., 100). Based on our experimentation, we determine 100 to be the optimal pruning value in our case, as it stabilizes both FPED and FNED. We hypothesize that pruning a large number of neuron weights may adversely affect the fairness of the model.

6.2 Qualitative Analysis

To demonstrate the effectiveness of the *BiasWipe* technique, we utilize model explainability to explain the predictions of toxic classification models. We extracted the tokens responsible for classification using the Shapley algorithm (Section 3.4.5).

Table 3 showcases examples of non-toxic cases from the template dataset that were originally misclassified by BERT, RoBERTa, and GPT2 models. This misclassification can be attributed to the presence of an entity keyword, where the model's focus on entities resulted in a toxic prediction. In Table 3, we show the first important word predicted by Shapley, illustrating the model's tendency to overgeneralize entities as toxic, thereby leading to toxic predictions. Upon pruning biased neuron weights using the *BiasWipe* approach, we observed a shift in the model's focus to another keyword (FAU: Focus after unlearning). This shift in focus helps the model to correctly classify the examples.

7 Conclusion

We proposed a novel technique *BiasWipe* to enhance the fairness of toxic classification systems. Our proposed post-processing technique utilizes model interpretability to mitigate within the model. It effectively detected the neuron weights responsible for unintended bias in the model and pruned them to ensure fairness. Our proposed technique mitigated the bias without additional re-training or components. We demonstrated the effectiveness and generalizability of the proposed method for BERT, RoBERTa, and GPT models. To illustrate the effectiveness of *BiasWipe* qualitatively, we used model explainability to explain how it helped the model in unlearning bias.

In the future, we plan to extend this work to other classification tasks involving other monolingual and code-mixed languages. We believe that our proposed approaches can also be used to reduce the bias of the systems in code-mixed settings.

Limitations

Like most studies, this study has some limitations that could be addressed in future research. The scope of our current work is confined to the English language. However, bias can be present in other language datasets also. In the future, our efforts will focus on enhancing our work in other monolingual and code-mixed languages.

Ethics Statement

We use freely available datasets for our experiments. The dataset has been used only for academic purposes, and in complete compliance with the license.

Acknowledgment

Mamta gratefully acknowledges the partial support from the Engineering and Physical Sciences Research Council (EPSRC, grant number EP/X04162X/1).

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. 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.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022a. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022b. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*, pages 67–73.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in nlp. 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 1905–1925.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. 2023. Bias beyond english: Counterfactual tests for bias in sentiment analysis in four languages. *arXiv preprint arXiv:2305.11673*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 37–44.

Przemyslaw Joniak and Akiko Aizawa. 2022. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 67–73, Seattle, Washington. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2005.02439*.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *NAACL HLT 2018*, page 43.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR.

Mamta and Asif Ekbal. 2023. Service is good, very good ornbsp;excellent? towards aspect based sentiment intensity analysis. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, page 685–700, Berlin, Heidelberg. Springer-Verlag.

Mamta and Asif Ekbal. 2024. Atmosphere kamaal ka tha (was wonderful): A multilingual joint learning framework for aspect category detection and sentiment classification. *IEEE Transactions on Computational Social Systems*.

Mamta and Asif Ekbal. 2025. Quality achhi hai (is good), satisfied! towards aspect based sentiment analysis in code-mixed language. *Computer Speech Language*, 89:101668.

Mamta Mamta, Zishan Ahmad, and Asif Ekbal. 2023. Elevating code-mixed text handling through auditory information of words. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15918–15932, Singapore. Association for Computational Linguistics.

Mamta Mamta and Asif Ekbal. 2022. Adversarial sample generation for aspect based sentiment classification. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 478–492, Online only. Association for Computational Linguistics.

Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173. PMLR.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *Association for Computational Linguistics (ACL 2019)*.

Chris Sweeney and Maryam Najafian. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 359–368.

Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*, 42(1):45–57.

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.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. 2020. Optimized score transformation for fair classification. *Proceedings of Machine Learning Research*, 108.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Alina Zhiltsova, Simon Caton, and Catherine Mulway. 2019. Mitigation of unintended biases against non-native english texts in sentiment analysis. In *AICS*, pages 317–328.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2021. Challenges in automated debiasing for toxic language detection. *arXiv preprint arXiv:2102.00086*.

A Ablation Study

We also conduct experiments with various weight pruning values to assess their impact on the accuracy of the model. Figure 6 illustrates the results for BERT and RoBERTa models, respectively, focusing on the third (final) unlearning step. The initial two steps involve debiasing the model for *gay* and *homosexual* entities, while the third step targets bias associated with the *lesbian* entity. We have analysed the tradeoff between accuracy and the number of pruned weights for BERT and RoBERTa models. We observed that pruning a large number of neuron weights adversely affects the model's performance. For example, in case of BERT model, pruning upto 100 weights has increased the accuracy of the model on template set. However, pruning 250 weights has reduced the accuracy by 4.14%.

B More Experiments

To demonstrate the effectivenss *BiasWipe*, we perform experiments on Hate-Speech-and-Offensive-Language (HSOL) dataset (Davidson et al., 2017).

Entity	Original BERT	Unlearn 1 (lesbian)	Unlearn 2 (queer)
gay	43.06	1.06	1.04
hetrosexual	45.71	6.47	2.06
white	46.76	0.40	0.26
bisexual	59.31	5.02	4.36
homosexual	93.53	10.17	6.61
lesbian	96.21	43.59	32.76
queer	95.12	85.75	2.77

Table 4: False positive rates of entities for BERT and debiased models (Hate-Speech-and-Offensive-Language dataset)

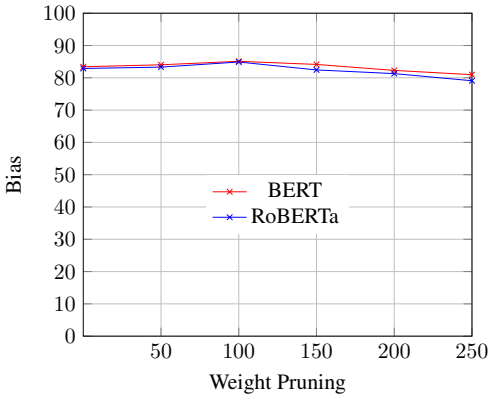

Figure 6: Affect of weight pruning on Accuracy (BERT and RoBERTa). **These results are for third unlearning step.**

Dataset is crawled from Twitter annotated using the Twitter API. with HSOL is a dataset for hate speech detection. Dataset contain 24,802 tweets annotated for three categories, namely, hate speech, offensive but not hate speech, or neither offensive nor hate speech. We treat the hate speech and offensive classes as one class. Therefore, we performed experiments on binary classes, *viz.*, hate-offensive and neutral.

Table 4 depicts the false positive rate (FPR) of the entities (template set) for trained BERT model. We show the FPR for entities before unlearning (original BERT) and after unlearning. It is evident from the Table that BERT model exhibit bias towards gay, homosexual, white, bisexual, homosexual, queer, and lesbian entities (high false positive rate). For BERT model, we employ two sequential unlearning steps. In the first step of unlearning (unlearn 1), we utilize false positive samples of the *lesbian* entity for unlearning and prune 100 (top k) biased weights. We observe that unlearning bias due to the *lesbian* entity significantly reduces bias of the other entities as well. Similarly, the second

Hyper-parameter	**Values**	**BERT**	**RoBERTa**
Learning rate	1e5, 2e-5,3e-5,5e-5	3e5	3e-5
Batch size	8,16,32	16	16

Table 5: Hyper-parameter values for BERT and RoBERTa models

(unlearn 2) step of unlearning helps the model reduce bias due to *queer* entity. Table 4 illustrates that unlearning lesbian and queer entities have significantly reduced the bias due to *homosexual*, *bisexual*, *white*, *hetrosexual*, and *gay* entities.

We also analyzed the accuracy on the test set to observe the effect of unlearning on actual performance of the model. We observe the original accuracy on test set (original BERT) came out to be 95.92% and after second unlearning step, accuracy is 95.34%. Our proposed technique significantly reduces the *FPR* for biased entities without compromising performance on the actual test set without additional re-training. The effectiveness of *BiasWipe* in reducing bias illustrates that our technique can effectively detect neurons weights responsible for bias in the network and can reduce bias by pruning only small fractions of biased weights.

C Experimental Setup

The target models, BERT-base and RoBERTa-base, employ 12 transformer blocks with a hidden size of 768 and 12 self-attention heads. BERT and RoBERTa consist of 110 million and 125 million parameters, respectively. Optimization is performed using the Adam optimizer, updating weights based on categorical cross-entropy loss. Hyperparameters are fine-tuned using grid search to determine the optimal parameter sets. We experiment with hyper-parameters shown in Table 5 to train target models.