Introducing Gender Stereotype-Aware Loss Regularizers to Reduce Gender-Based Bias in Language Models

Divyajeet Singh
divyajeet21529@iiitd.ac.in

Mehak Gopal
mehak21475@iiitd.ac.in

Siddhant Rai Viksit
siddhant21565@iiitd.ac.in

Sohum Sikdar
sohum20339@iiitd.ac.in

Abstract

Deep learning-based natural language models are becoming increasingly capable of learning patterns in various languages. Studies show that blind application of such models presents the risk of memorizing and amplifying the bias in the data on which they are trained. Motivated by recent works in the field of ethics in artificial intelligence, we analyze the presence of gender bias in Masked Language Models (MLMs), specifically the BERT Masked Language Model. This paper presents two novel loss regularizers to control and mitigate gender-related stereotypical bias in language models. These regularizers can be used with regular MLM objectives, using which the MLM learns to eliminate bias in datasets by penalizing stereotypical predictions made by the model. Using these approaches, we fine-tune the BERT MLM and quantitatively and qualitatively analyze the difference in predictions of the pre-trained and fine-tuned models.

1. Introduction

In recent years, the proliferation of artificial intelligence, particularly in natural language processing, has catalyzed profound advancements. However, these advancements have also given into the inherent biases encoded within these models, often reflecting and perpetuating societal biases and prejudices, which become more prominent as larger models are introduced and prepared using massive datasets. Among the most concerning biases is gender bias, which manifests in language models through skewed representations and predictions, potentially reinforcing harmful stereotypes and inequities.

The prevalence of gender bias in deep learning-based natural language systems has gained attention from modern researchers. Works such as Bolukbasi et al. [2], Jentzsch et al. [4], and Bhardwaj et al. [1] have highlighted instances where models based on BERT (Bidirectional Encoder Representations from Transformers) exhibit gender-based biases, which potentially influence decisions in sensitive areas such as hiring, loan approval, and content recommendation. Recognizing the ethical imperative to eliminate such biases, scholars strive to develop techniques that enforce fairness and inclusivity in these systems.

This paper presents two novel approaches to mitigate gender bias in BERT's Masked Language Model (MLM), a state-of-the-art pre-trained NLP model which has demonstrated remarkable performance across various NLP tasks. Like many deep learning models, BERT is susceptible to encoding and perpetuating gender biases in its training data [5]. Our proposed methodology addresses this concern by introducing two Gender Stereotype-Aware Loss Regularizers (Gals) designed to penalize predictions that exhibit gender bias during the fine-tuning process of BERT's MLM. We show that incorporating the GAL Regularizers into the MLM training objective encourages the model to generate more gender-neutral predictions, thereby mitigating the amplification of gender bias in its outputs. This results in a more gender-neutral model, obtained by targeting stereotypical model responses and reducing their prediction probabilities.

2. Related Work

2.1. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Bolukbasi et al. [2] study the nature of BERT embeddings and discover that BERT word embeddings, even when fine-tuned on real-world datasets like Google News, exhibit a concerning amount of male/female stereotypes. They showed that while the embeddings effectively capture the notion of gender using specific dimensions, they also pinpoint sexism implicit in the text. As an interesting example,

they observed that

$$\textbf{man} - \textbf{woman} \approx \textbf{king} - \textbf{queen}$$
$$\approx \textbf{computer programmer} - \textbf{homemaker}$$

The embeddings seem to capture predominantly male and female occupations as stereotypical *he-she* analogies such as shopkeeper-housewife and surgeon-nurse, along with gender-appropriate analogies like brother-sister and ovarian cancer-prostate cancer along the same dimensions. The paper focuses on the geometry and physical distributions of the embeddings and finds that along with the bias, the embedding space also contains information to help reduce the bias.

The authors study and identify the subspace of the embeddings that capture gender-based relationships and neutralize and equalize the space from gender-neutral words, ensuring that neutral words are equidistant from all words in each gendered set. For example, the **babysit** would be equidistant to **grandfather** and **grandmother**. This way, they present debiasing techniques to de-bias the embeddings geometrically while still preserving their most semantically meaningful properties.

2.2. Gender Bias in BERT – Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task

Jentzsch & Turan [4] analyze the gender-based biases in multiple publicly available BERT models. Their paper uses seven different publicly available BERT models in nine different training conditions (63 models in total) to analyze gender bias using the IMDB Large Movie Review Data. They introduce a novel method to measure sentiment bias, where bias is defined as the difference in sentiment valuation of female and male sample versions. Their bias measure is as follows:

$$\begin{aligned} Bias_{XY}(i) &= \Delta\text{sent} \\ &= sent(i_Y) - sent(i_X) \end{aligned}$$

where, $X = Female$ and $Y = Male$. $Bias_{XY}(i)$ represents the bias for a sample i with X version as i_X and Y version as i_Y. The experimental pipeline of this paper follows four major steps

1. The IMDB data collected was preprocessed to remove all punctuation and lowercase. The dataset was labelled using a thresholding method. Reviews with ratings ≤ 4 were labelled as negative, while those with ≥ 7 were marked as positive.

2. The sentiment classifiers were trained by fine-tuning seven common BERT models. Different sizes in the same architecture were also tested.

3. These models were then tested on the manipulated test data.

4. The results obtained were used to calculate the gender bias.

They concluded and showed that all trained classifiers show significant gender bias and highlighted that the biases are propagated from underlying pre-trained BERT models rather than learnt in task-specific training.

2.3. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation

Levy et al. [6] presented BUG, a large-scale gender-bias evaluation corpus consisting of diverse, real-world scenarios. Using this dataset, they evaluate gender bias on a large scale using machine translation and coreference resolution. Upon fine-tuning a model on the BUG dataset, they were able to show that that model was less prone to make gender-biased predictions.

To make BUG, they used 14 lexical-syntactic patterns, with two main anchors, a pronoun and a profession, which the pattern indicates are co-referring. To match these 14 patterns against real-world texts, they used SPIKE, which indexes large-scale corpora and retrieves matching instances given a lexical-syntactic pattern. Each instance in BUG has also been marked as stereotypical, neutral or anti-stereotypical. The results were all filtered with a predefined list of professions taken from the U.S. Census.

To evaluate their models, they used three main metrics

1. Accuracy: The F1 score of the gender prediction.

2. Population Bias (ΔG): This denotes the difference in the accuracy between sentences with entities that co-refer with a masculine pronoun versus those with entities that co-refer with feminine pronouns.

3. Historical Bias (ΔS): This denotes the accuracy difference between stereotypical and anti-stereotypical sentences.

Their results concluded that all tested models for machine translation and coreference resolution are prone to gender bias in real-world texts. A spanBERT model, fine-tuned on the anti-stereotypical portion of BUG, yielded an error reduction of 50% at the cost of an absolute 1% drop in overall performance accuracy.

3. Methodology

Our primary objective is to de-bias the BERT Masked Language Model. While working on an unbiased sentiment analyser, we note that Jentzsch & Turan [4] perform gender masking on gendered tokens in their sentence. This

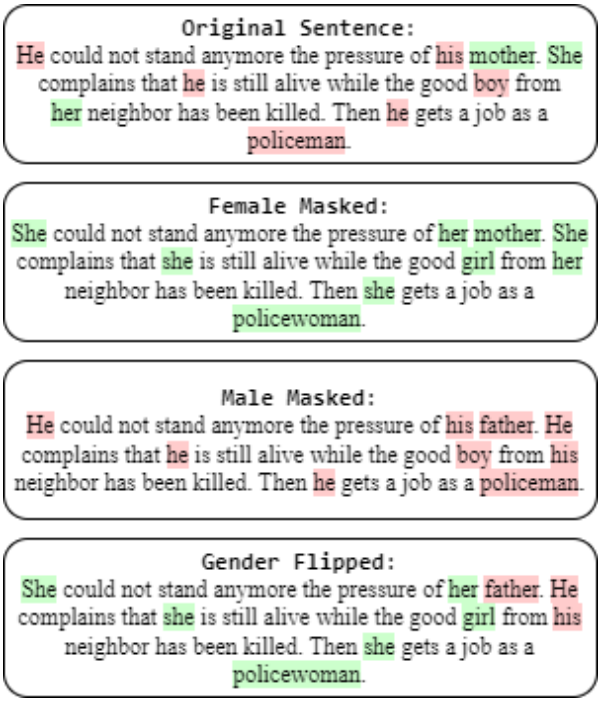

Figure 1. Masking and Flipping strategies for gendered sentences

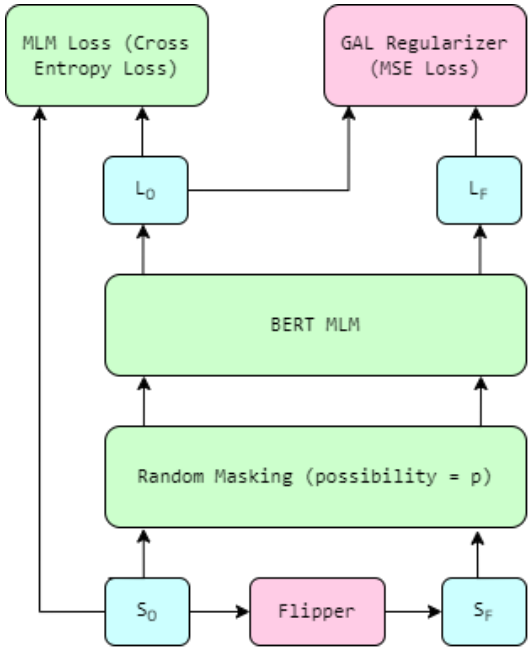

Figure 2. Model Architecture for fine-tuning with GAL_1

results in two masked versions of one original sentence - Male Masked and Female Masked. We hypothesize that their masking strategy is sub-optimal, as they convert all gendered words into the same gender, thereby losing out on cross-gender relationships. To tackle this, we present a unique flipping strategy - for each sentence in the corpus, we create its *Gender Flipped* version by flipping all its gendered tokens. Our strategies are compared in Figure 1

To create the Flipper, we used a public dictionary of gendered words in English, available at https://github.com/ecmonsen/gendered_words. The dictionary was created by manually tagging a large set of words as male, female, or neutral. We form a frozen mapping of male-to-female and female-to-male correspondences using this dictionary. For example, `actress` is mapped to `actor`, `brother` is mapped to `sister`, and so on.

The first GAL Regularizer can be used with the MLM objective of the BERT MLM, which is typically the Cross-Entropy Loss between the original sentence and its prediction after masking. The regularizing term is introduced using the prediction of the flipped sentence. Let S_O be the original sentence and S_F be its flipped version. We create a mask which masks each token independently with probability p and mask both sentences using it. Let L_0 and L_F be their corresponding predicted logits on the masked sentences. Then, the final fine-tuning objective is given by

$$\mathcal{L}_1 = \text{CE}(S_O, L_O) + \lambda \, \text{MSE}(L_O, L_F) \tag{1}$$

The idea is to reduce the difference/*distance* between the predictions for both the original and flipped sentences. λ is a hyperparameter to influence the strength of regularization.

The second GAL Regularizer can also be used with the MLM objective to create the final loss function. This time, we mask all the gendered tokens in the sentence. Then, we define the final fine-tuning objective as

$$\mathcal{L}_2 = \text{CE}(S_O, L_O) + \lambda \, \text{CE}(S_F, L_O) \tag{2}$$

The hope for this idea is to ensure that the model learns to predict both genders under appropriate contexts with approximately equal probability given any sentence.

Concisely, our contributions are two GAL regularizers

$$\text{GAL}_1 = \text{MSE}(L_O, L_F) \tag{3}$$
$$\text{GAL}_2 = \text{CE}(S_F, L_O) \tag{4}$$

which can be used appropriately with the MLM objective to de-bias the MLM's outputs. The final model architectures are given in Figure 2 and Figure 3.

One final idea is to use a randomized loss regularizer, GAL_R, in which we omit including the regularizer itself with some probability. As per our experiments, we were motivated to only introduce this technique with GAL_2.

$$\text{GAL}_R = \mathbf{I}_p \, \text{GAL}_2 = \mathbf{I}_p \, \text{CE}(S_F, L_O) \tag{5}$$

where $\mathbf{I}_p \sim \text{BERNOULLI}(p)$ is a variable which switches to 1 with probability p. This way, with probability $1 - p$,

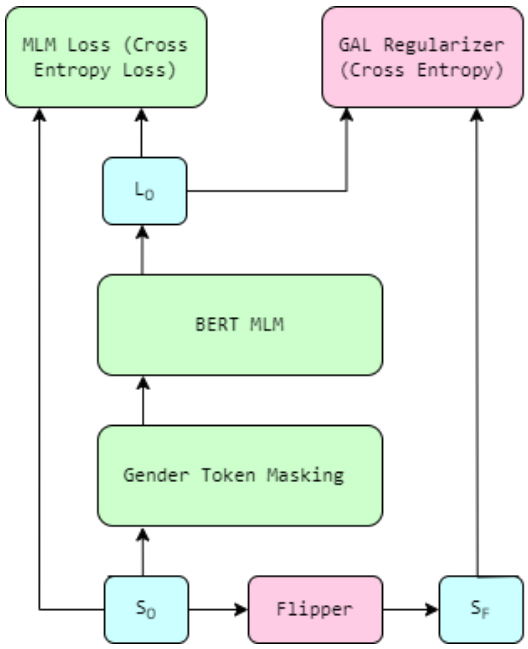

Figure 3. Model Architecture for fine-tuning with GAL$_2$

the model only uses the MLM loss for loss calculation, gradient propagation, and parameter updates. This ideally ensures that the model does not throw away its already learned representations fast enough in the pursuit to minimize the regularizing term.

4. Dataset

Noticing the prevalence of gender-based bias in neural language models (specifically in machine translation and coreference resolution), Levy et al. [6] crafted a large-scale gender bias dataset called BUG. The dataset was created with more than 105,000+ diverse real-world English sentences, including stereotypical gender roles like female nurses and non-stereotypical gender roles like male dancers. The dataset is an unlabelled corpus of data relating to gender roles in society, which is perfect for our application. For example, the dataset consists of sentences like *"He was an American tap dancer, educator, and choreographer."* which are non-stereotypical, and sentences like *"Her ballerina skills were unmatched, but still questioned by many."*.

The dataset was collected semi-automatically from different real-world corpora, designed to be challenging regarding societal gender role assignments using techniques strategized to combat stereotypes as discussed in the Literature Review in 2.3. They used 14 lexical-syntactic patterns with two main anchors - a pronoun and a profession and performed co-reference resolution to indicate which anchors relate. The dataset is complete with labels indicating whether each sentence in the corpus is stereotypical (+1),

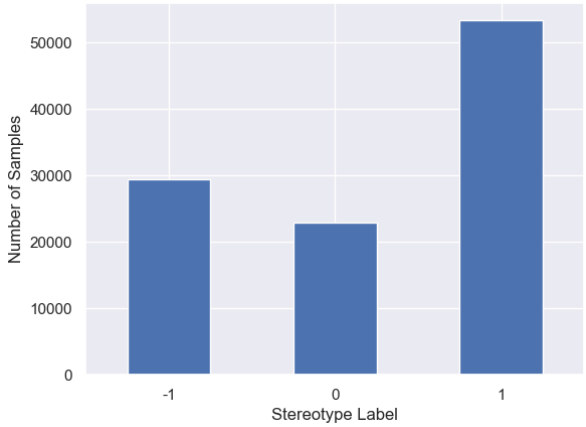

Figure 4. Distribution of stereotype labels in BUG

neutral (0), or anti-stereotypical (-1) and the professions and pronouns prevalent in the sentence.

The dataset is publicly available on GitHub, https://github.com/SLAB-NLP/BUG.

4.1. Dataset Preparation and Preprocessing

To suit our training pipeline, we simply utilize the unlabeled sentences. We ignore the stereotype labels in the dataset as we must fine-tune our model on a balanced mix of neutral, stereotypical, and anti-stereotypical sentences. Figure 4 displays the distribution of sentences in each class. Finally, we use the flipping strategy described in 3 to obtain the gender-flipped versions of each sentence.

4.2. Alternate Datasets

We also adopted an alternate strategy to extend the dataset. Under this, we picked up large volumes of text corpora that we expected to be relevant to our problem statement. Using our sentence gender flipper, we identify gendered sentences in the corpora and append them to our dataset. However, a large fraction of the sentences were found to be irrelevant to the problem statement. Corpora such as wikitext-103-v1 also contained gendered sentences relating to biology and botany. Such sentences may hinder our progress as the model may lose discriminative ability against factual truths. For example, it is known that pregnant male seahorses give birth to live young [9]. Our training procedure may penalize the model for producing factually correct outputs.

5. Experimental Setup

We set up our experiments under the following constraints. The pre-trained BERT MLM checkpoint, available through Hugging Face, and the BERT base uncased tokeniser were used for this task. We used the Adam optimizer with a learning rate of $5e-5$ for all experiments,

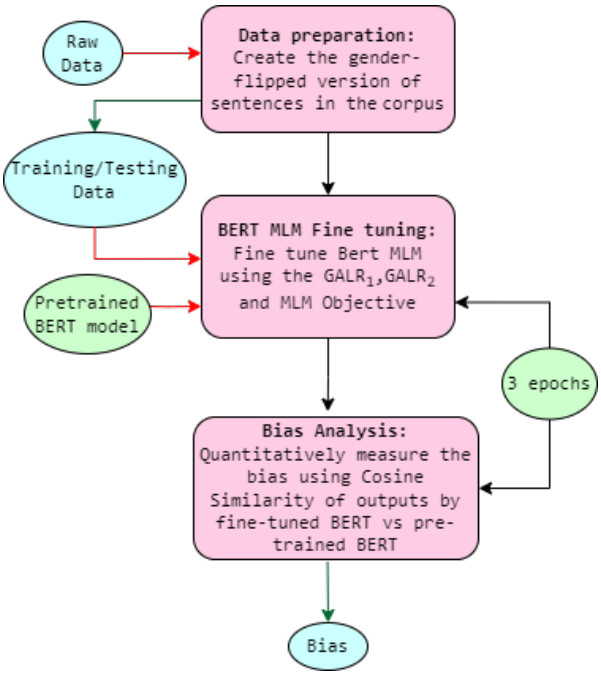

Figure 5. Experiment Pipeline

and the batch size was set to 16. The model was trained on A-100 GPUs. We run the fine-tuning for a total of 3 epochs, as suggested by [3]. Our experimental pipeline is shown in Figure 5.

To test GAL_1, we performed several experiments. We tested regularizing strengths of $\lambda = 2e-5, 0.1, 1, 10, 100$. We observed that with higher values of the regularizing strengths, the model prioritizes minimizing the gender loss term more than the MLM loss, resulting in meaningless outputs. It tends to produce extremely neutral words for every predicted token, such as `the`, `and`, `on`, etc., entirely losing the sentence coherence.

To test GAL_2, we tested regularizing strengths of 0.5 and 1. We found that with $\lambda = 1$, the model MLM is de-biased to the best possible extent. However, we note that even in this case, the model invariably produces some outputs we cannot reason for.

Finally, the pre-trained BERT MLM was fine-tuned on the Randomized Ensemble Regularizer, GAL_R. In the ensemble method of fine-tuning, We observed that the hyper-parameter setting of $\lambda = 1$ and $p = 0.5$ works well for our use case. This way, we ensure that the MLM retains its most useful predictions while focussing on overcoming the bias.

6. Results

Our qualitative and quantitative analysis identifies the model fine-tuned with the GAL_R regularizer as the best model. We visualize the loss per step of the best model

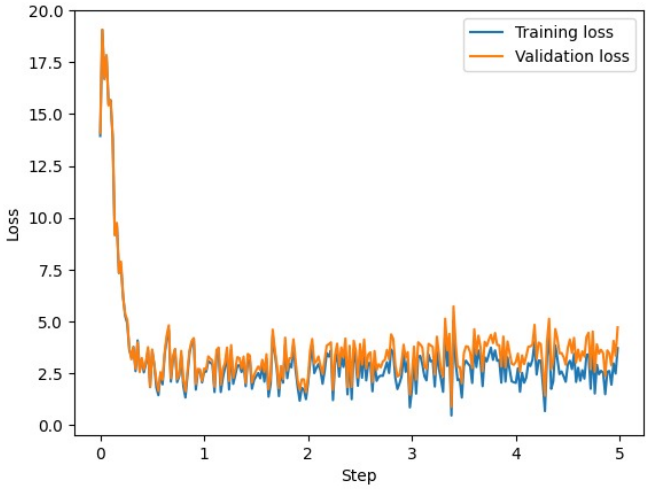

Figure 6. Loss per step of fine-tuning

Model	Word	Probability
Pre-trained Base	`Mom`	0.36
Pre-trained Base	`Mother`	0.26
Pre-trained Base	`Dad`	0.10
GAL_1	`Friend`	0.22
GAL_1	`Mother`	0.12
GAL_1	`Chef`	0.10
GAL_2	`[PAD]`	0.23
GAL_2	`Mother`	0.11
GAL_2	`Father`	0.08
GAL_R	`Mother`	0.20
GAL_R	`Father`	0.12
GAL_R	`Friend`	0.08

Table 1. Prediction Probabilities for the Test Sentence

in Figure 6.

We first qualitatively assess the performance of our models. We use a dataset by Nangia et al. [7], explicitly curated to analyze social biases in masked language models. The dataset contains pairs of male and female-dominated sentences in professional and social environments. It is publicly available at https://github.com/nyu-mll/crows-pairs. Consider a sentence from this dataset `"My [MASK] spent all day cooking food for dinner."`. The predictions with probabilities are given in Table 1.

To quantitatively analyze bias, we perform sentence similarity matching using cosine similarity. On the test dataset, we calculate the cosine similarity between the predicted sentences for each original sentence and its flipped version. We hypothesize that after the fine-tuning steps, the MLMs are de-biased enough to produce semantically meaningful yet similar outputs for both the original and gender-flipped

Model	Cosine Similarity
Pre-trained Base	0.9522 ± 0.0342
GAL_1	0.9996 ± 0.0005
GAL_2	0.9976 ± 0.0078
GAL_R	0.9972 ± 0.0077

Table 2. Cosine Similarity between Original and Flipped Sentences

sentences. As the literature review suggests, the pre-trained BERT MLM simply produces varying outputs depending on gender-based context. So, our final bias quantifier is

$$\text{BIAS}(M) = \frac{1}{N}\sum_{i=1}^{N} \text{COS}\left(M\left(S_O^{(i)}\right), M\left(S_F^{(i)}\right)\right) \quad (6)$$

where $S_O^{(i)}$ and $S_F^{(i)}$ are the original and flipped versions of the i^{th} sentences, M is the model, and COS represents the cosine similarity. As opposed to just producing higher similarity scores, a sensible de-biased model must find a balance between producing an extremely high similarity and predicting meaningful outputs. The test results are summarized in Table 2.

We find that the model fine-tuned on the probabilistic regularizer GAL_R, produced the most semantically meaningful and gender-neutral outputs.

7. Discussion

Through the strategies presented in this paper, we aim to work towards a lingering issue in all of machine learning, especially in the natural language field. We hope the ideas shared in this paper spark further discussion and interest in the intersection of gender studies and natural language processing since all models must represent all groups fairly. The problem is simple - biased data will produce biased models. Its solution, however, turns out to be tricky.

7.1. Limitations

Our methodology revolves around *flipping* the gendered words in sentences and analyzing and de-biasing the model results. However, the ideas in this paper may turn back on themselves if proper datasets are not used for fine-tuning. For example, if one chooses to fine-tune for gender-based bias minimization using our technique with a dataset of fauna information, such as a subset of the Wikitext, the models may lose their discriminative ability towards biological truths.

7.2. Possibilities for Exploration

Since the regularizers work favourably, we hypothesize that even more representative datasets can boost the performance we intend to achieve. For example, the bookcorpus dataset from Hugging Face, a corpus of text from the book, contains more than 74 million entries, which must embody the sexism present in conservative written literature. Such large datasets are well-suited to the problems we intend to solve.

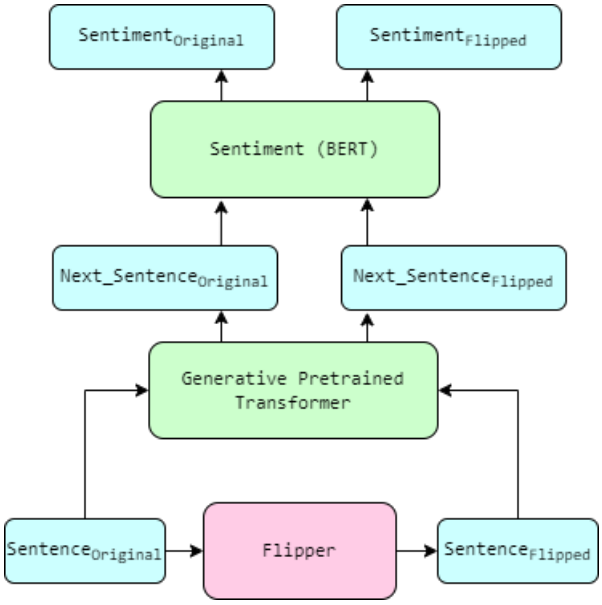

Figure 7. Proposed model architecture to de-bias GPT-2

8. Conclusion and Future Work

Through this paper, we have introduced two novel Gender-Stereotype Aware Loss Regularizers. We have demonstrated how incorporating these regularizers in the training objective of the Masked Language Models reduces gender-based bias and favours the model to produce more gender-neutral outputs while remaining semantically meaningful under appropriate contexts. We show that the BERT-based masked language model when fine-tuned using the randomized GAL regularizer, results in the most gender-neutral model.

We believe the same idea can be transferred from Masked Language Modeling to the Next Sentence Prediction Problem. Next sentence predictors, even those trained on humongous datasets, are prone to gender-based bias due to the bias inherent to their training datasets. This is due to the very nature of societal bias - larger datasets amplify them [8]. Hence, we propose a technique to de-bias GPT-2 as a future plan. The basic model architecture is shown in Figure 7. The basic low-level idea is to generate the next sentence from the original and gender-flipped versions of the sentences in the corpus and perform sentiment-based matching of the outputs. This sentiment matching can then act as a regularizing term, where we hope to produce sentences with similar sentiments for either version of the sen-

tence. We may also use metrics such as cosine similarity to perform a matching.

References

[1] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating gender bias in bert, 2020. 1

[2] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016. 1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 5

[4] Sophie Jentzsch and Cigdem Turan. Gender bias in bert - measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, 2022. 1, 2

[5] Thibaud Leteno, Antoine Gourru, Charlotte Laclau, and Christophe Gravier. An investigation of structures responsible for gender bias in bert and distilbert. In Bruno Crémilleux, Sibylle Hess, and Siegfried Nijssen, editors, *Advances in Intelligent Data Analysis XXI*, pages 249–261, Cham, 2023. Springer Nature Switzerland. 1

[6] Shahar Levy, Koren Lazar, and Gabriel Stanovsky. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2, 4

[7] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models, 2020. 5

[8] Elena Pesce and Eva Riccomagno. Large datasets, bias and model oriented optimal design of experiments, 2018. 6

[9] Kai Stölting and Anthony Wilson. Male pregnancy in seahorses and pipefish: Beyond the mammalian model. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 29:884–96, 09 2007. 4