Extending Prefix-Tuning to Multiple Languages for Bias Mitigation in Language Models

Maysara Al Jumaily, Yiduo Zhang, and Anuththara Lekamalage

***Abstract*—While pre-trained language models proved to undertand and generate natural languages, they also capture the bias and sterotypes of the data. In this work, we used NLLoss, cross entropy loss and focal loss to debias gendered words in sentences and compare earch other. We extended the dataset into French, also tested on English dataset that corresponds to the French dataset, found out that focal loss achieves good results among three dataset. The entire implementation has been hosted on GitHub through the following link: https://github.com/Aljumaily/COSC5P84FinalProject**

I. Introduction

Language models have revolutionized the way machines understand and generate human language. From powering virtual assistants and chatbots to enhancing machine translation and summarization tools, these models are now embedded in numerous applications that impact our daily lives. However, despite their impressive capabilities, language models are not free from flaws. A growing body of research has shown that these models often replicate and amplify societal biases found in their training data [1], [2], [3]. These biases can be harmful, leading to skewed representations, unfair treatment, or reinforcement of stereotypes.

One of the most pervasive types of bias observed in language models is gender bias [3], [4], [5]. For instance, when prompted with a sentence like "The nurse said," many models are more likely to continue the sentence with female pronouns, whereas "The doctor said" is more likely to be followed by male pronouns. Such associations, while subtle, can perpetuate stereotypes and influence user perceptions. In multilingual contexts, the issue becomes even more complex. Languages like French and Spanish use gendered grammatical structures, making the manifestation of bias both more nuanced and embedded in language rules themselves.

To address these challenges, researchers have explored various mitigation techniques [4]. Among the promising methods is prefix-tuning, a parameter-efficient approach that modifies only a small portion of the model while keeping the majority of the pre-trained parameters fixed [6]. This allows for efficient and targeted adaptation of language models without the high computational cost of full fine-tuning. Wang and Demberg (2024) demonstrated the efficacy of prefix-tuning in reducing gender bias in English language models by incorporating a multi-objective loss framework.

Building on their work, our project investigates whether this method can be effectively extended to multilingual scenarios, particularly focusing on French-English parallel corpora. Our primary aim is to explore how prefix-tuning can be adapted to languages with differing grammatical structures and to assess its potential in mitigating gender-based bias across culturally diverse contexts.

II. Research Objectives

The first objective of our research is to adapt and extend the prefix-tuning method, originally designed for English language models, to handle multiple languages. In particular, we focus on French-English models, leveraging parallel corpora to ensure that both languages are exposed to equivalent contextual information during training.

Another key goal is to investigate the impact of linguistic structure on model bias. French, with its inherently gendered grammatical system, presents unique challenges compared to English, which tends to be more gender-neutral. By comparing model outputs across these languages, we aim to understand how gender bias manifests differently and assess whether methods designed for English can be directly transferred or require adaptation.

Lastly, we aim to evaluate the efficacy of prefix-tuning in mitigating gender-based biases in a multilingual setting. This involves measuring changes in bias-related metrics, such as the frequency of gendered word associations, and analyzing whether prefix-tuned models can maintain linguistic fluency while producing fairer outputs. These insights could inform the development of future bias mitigation techniques across linguistically diverse settings.

III. Methodology

We began by collecting a French-English parallel corpus, ensuring that each pair of sentences conveyed the same context to minimize linguistic variability. This approach was essential for a fair comparison between a gendered language (French) and a relatively gender-neutral one (English). The corpus included common sentence structures and vocabulary typically found in real-world applications such as news articles.

Next, we created a gender-based term mapping. We extracted masculine and feminine word pairs from an online French dictionary and validated them using tools like Google Translate and Gemini 2.0 Flash. This process resulted in the generation of over 12,625 semantic gender-based pairs, significantly expanding upon the 223 pairs originally used in English-only datasets. We realized the output generated wasn't as reliable and thus, used a different approach that will be discussed in the next section.

For the debiasing process, we adopted prefix-tuning, a lightweight and efficient fine-tuning method. In this method, we appended small trainable prefix vectors to the inputs of a pre-trained language model. Rather than updating the entire model's parameters, this approach only adjusted the prefix, making it computationally efficient and suitable for multi-language adaptation.

Our optimization strategy was inspired by the work of Wang and Demberg (2024), who proposed a multi-objective loss framework to guide the training process. We incorporated four loss components: the language modeling loss , which ensures the preservation of general linguistic capabilities; the neutrality loss, which encourages the use of gender-neutral terms; and two equality losses, and, which balance gender representation at the token and sequence levels, respectively.

To further enhance the model's focus on underrepresented or harder-to-predict gender-neutral terms, we explored integrating a focal loss component. Originally introduced by Lin et al. (2017)[7] for dense object detection, focal loss addresses class imbalance by down-weighting easy examples and focusing more on hard, misclassified examples. The focal loss is defined as:

$$\text{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

where p_t is the model's estimated probability for the true class, is α_t a weighting factor, and γ is the focusing parameter. The new loss function that we are using is the focal loss function with gamma as 2. In our context, this loss can help prioritize learning on gender-neutral outputs which may be harder for the model to predict due to their infrequency in the training data.

These loss components were optimized simultaneously, steering the model away from biased outputs while maintaining fluency and coherence. The training was conducted over five epochs using a gender-swapped version of the original dataset to reinforce symmetric associations between masculine and feminine expressions. Prefix vectors were initialized randomly and updated via gradient descent, and loss values were monitored at each epoch to assess the progress of bias mitigation.

IV. The Dataset

We have five datasets that we generated: `en`, `fr`, `fr-en`, `es` and `es-en`. However, in terms of execution, we were able to execute the first three datasets (*i.e.,* `en`, `fr` and `fr-en`). The `en` represents the original dataset used by [6]. It contains 13995 sentences along with 222 male/female words (*e.g.,* king-queen, actor-actress, etc.) The `fr` dataset is the French dataset that is hosted through https://huggingface.co/datasets/bilalfaye/english-wolof-french-translation-bis. The dataset contains 11349 sentences. In terms of the gender-based semantic relationship pairs, a lot of work had to be done. First, we downloaded the entire French dictionary from https://kaikki.org/dictionary/French/index.html and scraped all words that contain the tag masculine and feminine. Then, we found the minimum word count between masculine and feminine to generate equal number of gender-based semantic relationship pairs. Finally, we fed the word list to GPT 4.0 (premium version) to find the equivalent words. We focused on context learning with GPT where prompts were requested to find the male word equivalent. In case a word doesn't have an equivalent, it is simply removed from the list. After sanitization, we generate 9675 pairs. The `fr-en` contains the same equivalent sentences that of the French dataset, but translated in English. For example, the French dataset might have the sentence 'Je suis ravi de vous rencontrer' and the `fr-en` would have the English equivalent, which is 'I'm pleased to meet you'. The same is applied on the Spanish dataset, which has the English equivalent `es-en`. The same process was used where we scraped the entire Spanish dictionary from https://kaikki.org/dictionary/Spanish/index.html. OpenAI's premium context learning was used to find the gender-based semantic relationship pairs. In total, there are 30888 pairs in the Spanish language.

V. The Results

We are able to only execute the `en`, `fr` and `fr-en` datasets. For each, we have used three different loss functions: `NLLLoss`, `CrossEntropyLoss` and `FocalLoss`. The original implementation only used `CrossEntropyLoss`. The motive of using these three loss functions is discussed in the subsequent sections. The hyper parameters are found in Table I. Each simulation used 5 as the number of epochs and `gelu_new` as the activation function. We see that as the number of epochs increased, so does the error. In all of the nine different simulations, this pattern followed through. The tables are found in the appendix section. Key highlights of the tables include:

- On English dataset, focal loss starts with the lowest loss with $8.179 \cdot 10^0$ in training losses and $6.055 \cdot 10^0$ in validation losses. After 5 epochs, cross entropy loss function achieves the best results with $4.492 \cdot 10^0$ in T otal and $3.954 \cdot 10^0$ in validation losses. Non-negative loss function is always the worst, it initials with $5.930 \cdot 10^1$ in training losses and $7.136 \cdot 10^0$ in validation losses.
- On French dataset, Focal Loss shows lower training loss ($7.674 \cdot 10^0$) compared to non-negative loss ($5.509 \cdot 10^1$) and cross entropy loss ($8.179 \cdot 10^0$). All 3 loss functions drop with the number of epochs. Training losses of Non-negative loss drop significantly from $5.509 \cdot 10^1$ to $9.949 \cdot 10^0$.
- Focal loss also shows good results on English dataset that correspond to the French dataset. Its training loss starts with $1.528 \cdot 10^1$ at epoch 1, and it decreases to $6.108 \cdot 10^0$, which is slightly lower than cross entropy ($6.365 \cdot 10^0$).

Validation metrics followed a similar trend, demonstrating the model's improved fairness while maintaining overall performance.

VI. Discussion and Future Work

Our approach proves that prefix-tuning can be extended to languages with gendered grammar. Here are some future work objectives:

- Current work involves improving contextual consistency in the dataset and expanding bias types beyond gender.
- Neutral words in French hasn't been implemented yet, however, the English language has this completed.
- The Spanish language has all of the data required. More text processing is needed to ensure the gender swapping mechanism is executed appropriately.
- From a programming perspective, the user should have the ability to pass in the name of the loss function as a parameter. Currently, the loss function is hardcoded and must be updated manually.
- Future directions include testing on other languages and using dynamic context-sensitive prefix generation.

VII. Negative Log Likelihood Loss (NLLoss)

Formula (Multi-Class Classification)

For a single sample, if y is the true class index and $\hat{y}_c$ is the predicted **log-probability** for class c, the NLLLoss is given by:

$$\text{NLLoss} = -\log(\hat{y}_y)$$

Note: $\hat{y}$ must be a log-probability (i.e., the output of a `log_softmax` function).

Relation to CrossEntropyLoss

NLLLoss is a core component of `CrossEntropyLoss`, which combines a `log_softmax` layer and the NLLLoss in one function.

Why we use it

NLLoss is convex, which helps optimization algorithms find a local minimum efficiently. It is also differentiable, making it suitable for gradient-based optimization methods . It uses maximum likelihood to adjust the model's predictions to be more equitable across different genders.

VIII. Cross-Entropy Loss

Cross-entropy loss measures the difference between the true distribution and the predicted distribution of probabilities.

- For **binary classification**:

  $$\text{Cross-Entropy Loss} = -\left(y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right)$$

- For **multi-class classification**:

  $$\text{Cross-Entropy Loss} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

  Where:
  - y is the true label (0 or 1).
  - $\hat{y}$ is the predicted probability.
  - C is the number of classes.
  - y_c is a binary indicator for class c (1 if it's the correct class, 0 otherwise).

IX. Focal Loss

Focal Loss is a loss function designed to address class imbalance in tasks like object detection [7]. It modifies the standard cross-entropy loss by adding a factor that reduces the relative loss for well-classified examples, focusing more on hard-to-classify examples.

Formula

For a single sample, the focal loss is defined as:

$$\text{Focal Loss} = -\alpha(1-\hat{y})^{\gamma} y \log(\hat{y})$$

Where:

- $\hat{y}$ is the predicted probability of the true class.
- y is the true label (0 or 1).
- α is a balancing factor for class imbalance.
- γ is a focusing parameter that adjusts the rate at which easy examples are down-weighted.

Focal Loss enhances the standard loss function by emphasizing difficult-to-classify examples.

Why we use it

Focal loss helps reduce gender bias by focusing training on misclassified or underrepresented gender examples. It down-weights easy, overrepresented samples and up-weights hard ones.

X. The Hyper Parameters Used

The main article published by Wang et al. used the hyper parameters found next. In terms of the `FocalLoss`, we set $\gamma = 2.0$.

Attribute	Value
Model Type	GPT2
Activation Function	GELU-New
Type	Prefix Tuning
Architecture	GPT2LMHeadModel
α_1	1
α_2	50
α_3	200
α_4	250
Batch Size	16
β	10
Epochs	5
Language Model Batch Size	16
Language Model Learning Rate	$5 \cdot 10^{-5}$
Learning Rate	$5 \cdot 10^{-5}$

TABLE I: The hyper parameters used in this study.

References

[1] K.-W. Chang, V. Prabhakaran, and V. Ordonez, "Bias and fairness in natural language processing," 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): Tutorial Abstracts*, 2019.

[2] S. Raza, M. Garg, D. J. Reji, S. R. Bashir, and C. Ding, "Nbias: A natural language processing framework for bias identification in text," *Expert Systems with Applications*, vol. 237, p. 121 542, 2024.

[3] K. Stanczak and I. Augenstein, "A survey on gender bias in natural language processing," *arXiv preprint arXiv:2112.14168*, 2021.

[4] T. Sun et al., "Mitigating gender bias in natural language processing: Literature review," *arXiv preprint arXiv:1906.08976*, 2019.

[5] M. Bartl, A. Mandal, S. Leavy, and S. Little, "Gender bias in natural language processing and computer vision: A comparative survey," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–36, 2025.

[6] Y. Wang and V. Demberg, "A parameter-efficient multi-objective approach to mitigate stereotypical bias in language models," in *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 2024, pp. 1–19.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

APPENDIX

The appendix contains all of the results of the nine simulations we executed using different datasets and loss functions.

Metric Loss	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Train Total	$1.123 \cdot 10^{1}$	$5.988 \cdot 10^{0}$	$5.455 \cdot 10^{0}$	$5.189 \cdot 10^{0}$	$4.924 \cdot 10^{0}$
Train Language Model	$5.730 \cdot 10^{0}$	$5.036 \cdot 10^{0}$	$4.571 \cdot 10^{0}$	$4.413 \cdot 10^{0}$	$4.149 \cdot 10^{0}$
Train Gender	$2.146 \cdot 10^{-2}$	$2.786 \cdot 10^{-3}$	$2.224 \cdot 10^{-3}$	$1.872 \cdot 10^{-3}$	$1.757 \cdot 10^{-3}$
Train Neutral	$1.415 \cdot 10^{-2}$	$1.002 \cdot 10^{-3}$	$9.583 \cdot 10^{-4}$	$6.145 \cdot 10^{-4}$	$5.643 \cdot 10^{-4}$
Train Other	$5.496 \cdot 10^{0}$	$9.519 \cdot 10^{-1}$	$8.834 \cdot 10^{-1}$	$7.765 \cdot 10^{-1}$	$7.750 \cdot 10^{-1}$
Train Bias	$3.760 \cdot 10^{-2}$	$5.166 \cdot 10^{-3}$	$4.745 \cdot 10^{-3}$	$3.972 \cdot 10^{-3}$	$3.903 \cdot 10^{-3}$
Validation Total	$5.206 \cdot 10^{0}$	$4.533 \cdot 10^{0}$	$4.892 \cdot 10^{0}$	$4.043 \cdot 10^{0}$	$3.954 \cdot 10^{0}$
Validation Language Model	$5.070 \cdot 10^{0}$	$4.376 \cdot 10^{0}$	$4.796 \cdot 10^{0}$	$3.881 \cdot 10^{0}$	$3.788 \cdot 10^{0}$
Validation Gender	$1.295 \cdot 10^{-4}$	$9.725 \cdot 10^{-5}$	$8.665 \cdot 10^{-5}$	$8.709 \cdot 10^{-5}$	$8.777 \cdot 10^{-5}$
Validation Neutral	$2.104 \cdot 10^{-5}$	$7.626 \cdot 10^{-6}$	$9.680 \cdot 10^{-7}$	$3.716 \cdot 10^{-6}$	$5.671 \cdot 10^{-6}$
Validation Bias	$5.880 \cdot 10^{-4}$	$6.528 \cdot 10^{-4}$	$4.044 \cdot 10^{-4}$	$6.680 \cdot 10^{-4}$	$6.829 \cdot 10^{-4}$

TABLE II: The cross entropy loss function was used on the English dataset

Metric Loss	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Train Total	$8.179 \cdot 10^{0}$	$6.833 \cdot 10^{0}$	$6.499 \cdot 10^{0}$	$6.269 \cdot 10^{0}$	$6.053 \cdot 10^{0}$
Train Language Model	$5.918 \cdot 10^{0}$	$5.747 \cdot 10^{0}$	$5.616 \cdot 10^{0}$	$5.493 \cdot 10^{0}$	$5.432 \cdot 10^{0}$
Train Gender	$7.487 \cdot 10^{-3}$	$3.548 \cdot 10^{-3}$	$2.754 \cdot 10^{-3}$	$2.322 \cdot 10^{-3}$	$1.613 \cdot 10^{-3}$
Train Neutral	0.0	0.0	0.0	0.0	0.0
Train Other	$2.261 \cdot 10^{0}$	$1.086 \cdot 10^{0}$	$8.829 \cdot 10^{-1}$	$7.757 \cdot 10^{-1}$	$6.211 \cdot 10^{-1}$
Train Bias	$1.054 \cdot 10^{-2}$	$5.055 \cdot 10^{-3}$	$4.082 \cdot 10^{-3}$	$3.567 \cdot 10^{-3}$	$2.807 \cdot 10^{-3}$
Validation Total	$6.055 \cdot 10^{0}$	$5.594 \cdot 10^{0}$	$5.448 \cdot 10^{0}$	$5.282 \cdot 10^{0}$	$5.277 \cdot 10^{0}$
Validation Language Model	$5.797 \cdot 10^{0}$	$5.434 \cdot 10^{0}$	$5.316 \cdot 10^{0}$	$5.154 \cdot 10^{0}$	$5.156 \cdot 10^{0}$
Validation Gender	0.0	0.0	0.0	0.0	0.0
Validation Neutral	0.0	0.0	0.0	0.0	0.0
Validation Bias	$1.029 \cdot 10^{-3}$	$6.401 \cdot 10^{-4}$	$5.244 \cdot 10^{-4}$	$5.099 \cdot 10^{-4}$	$4.859 \cdot 10^{-4}$

TABLE III: The focal loss function was used on the English dataset,

Metric Loss	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Train Total	$5.930 \cdot 10^{1}$	$1.234 \cdot 10^{1}$	$1.067 \cdot 10^{1}$	$9.883 \cdot 10^{0}$	$9.590 \cdot 10^{0}$
Train Language Model	$6.490 \cdot 10^{0}$	$7.191 \cdot 10^{0}$	$7.222 \cdot 10^{0}$	$7.162 \cdot 10^{0}$	$7.109 \cdot 10^{0}$
Train Gender	$3.097 \cdot 10^{-2}$	$3.259 \cdot 10^{-3}$	$2.061 \cdot 10^{-3}$	$1.660 \cdot 10^{-3}$	$1.494 \cdot 10^{-3}$
Train Neutral	$2.137 \cdot 10^{-2}$	$4.453 \cdot 10^{-3}$	$1.807 \cdot 10^{-3}$	$1.321 \cdot 10^{-3}$	$1.195 \cdot 10^{-3}$
Train Other	$5.281 \cdot 10^{1}$	$5.146 \cdot 10^{0}$	$3.448 \cdot 10^{0}$	$2.721 \cdot 10^{0}$	$2.481 \cdot 10^{0}$
Train Bias	$2.345 \cdot 10^{-1}$	$2.480 \cdot 10^{-2}$	$1.565 \cdot 10^{-2}$	$1.227 \cdot 10^{-2}$	$1.118 \cdot 10^{-2}$
Validation Total	$7.136 \cdot 10^{0}$	$7.196 \cdot 10^{0}$	$7.143 \cdot 10^{0}$	$7.043 \cdot 10^{0}$	$7.016 \cdot 10^{0}$
Validation Language Model	$6.896 \cdot 10^{0}$	$7.045 \cdot 10^{0}$	$7.024 \cdot 10^{0}$	$6.944 \cdot 10^{0}$	$6.921 \cdot 10^{0}$
Validation Gender	$6.175 \cdot 10^{-4}$	$4.444 \cdot 10^{-4}$	$3.234 \cdot 10^{-4}$	$2.500 \cdot 10^{-4}$	$2.316 \cdot 10^{-4}$
Validation Neutral	$2.549 \cdot 10^{-4}$	$3.207 \cdot 10^{-5}$	$1.488 \cdot 10^{-5}$	$1.313 \cdot 10^{-5}$	$1.208 \cdot 10^{-5}$
Validation Other	$2.404 \cdot 10^{-1}$	$1.513 \cdot 10^{-1}$	$1.189 \cdot 10^{-1}$	$9.972 \cdot 10^{-2}$	$9.479 \cdot 10^{-2}$
Validation Bias	$1.289 \cdot 10^{-3}$	$7.197 \cdot 10^{-4}$	$5.520 \cdot 10^{-4}$	$4.594 \cdot 10^{-4}$	$4.351 \cdot 10^{-4}$

TABLE IV: The non-negative loss function was used on the English dataset

Metric Loss	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Train Total	$8.179 \cdot 10^{0}$	$6.833 \cdot 10^{0}$	$6.499 \cdot 10^{0}$	$6.269 \cdot 10^{0}$	$6.053 \cdot 10^{0}$
Train Language Model	$5.918 \cdot 10^{0}$	$5.747 \cdot 10^{0}$	$5.616 \cdot 10^{0}$	$5.493 \cdot 10^{0}$	$5.432 \cdot 10^{0}$
Train Gender	$7.487 \cdot 10^{-3}$	$3.548 \cdot 10^{-3}$	$2.754 \cdot 10^{-3}$	$2.322 \cdot 10^{-3}$	$1.613 \cdot 10^{-3}$
Train Neutral	0.0	0.0	0.0	0.0	0.0
Train Other	$2.261 \cdot 10^{0}$	$1.086 \cdot 10^{0}$	$8.829 \cdot 10^{-1}$	$7.757 \cdot 10^{-1}$	$6.211 \cdot 10^{-1}$
Train Bias	$1.054 \cdot 10^{-2}$	$5.055 \cdot 10^{-3}$	$4.082 \cdot 10^{-3}$	$3.567 \cdot 10^{-3}$	$2.807 \cdot 10^{-3}$
Validation Total	$6.055 \cdot 10^{0}$	$5.594 \cdot 10^{0}$	$5.448 \cdot 10^{0}$	$5.282 \cdot 10^{0}$	$5.277 \cdot 10^{0}$
Validation Language Model	$5.797 \cdot 10^{0}$	$5.434 \cdot 10^{0}$	$5.316 \cdot 10^{0}$	$5.154 \cdot 10^{0}$	$5.156 \cdot 10^{0}$
Validation Gender	0.0	0.0	0.0	0.0	0.0
Validation Neutral	0.0	0.0	0.0	0.0	0.0
Validation Other	$2.572 \cdot 10^{-1}$	$1.600 \cdot 10^{-1}$	$1.311 \cdot 10^{-1}$	$1.275 \cdot 10^{-1}$	$1.215 \cdot 10^{-1}$
Validation Bias	$1.029 \cdot 10^{-3}$	$6.401 \cdot 10^{-4}$	$5.244 \cdot 10^{-4}$	$5.099 \cdot 10^{-4}$	$4.859 \cdot 10^{-4}$

TABLE V: The cross entropy loss function was used on the French dataset

Metric Loss	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Train Total	$1.551 \cdot 10^{1}$	$7.392 \cdot 10^{0}$	$6.792 \cdot 10^{0}$	$6.515 \cdot 10^{0}$	$6.365 \cdot 10^{0}$
Train Language Model	$5.713 \cdot 10^{0}$	$5.892 \cdot 10^{0}$	$5.714 \cdot 10^{0}$	$5.561 \cdot 10^{0}$	$5.466 \cdot 10^{0}$
Train Gender	$4.533 \cdot 10^{-2}$	$6.070 \cdot 10^{-3}$	$4.071 \cdot 10^{-3}$	$3.454 \cdot 10^{-3}$	$3.116 \cdot 10^{-3}$
Train Neutral	0.0	0.0	0.0	0.0	0.0
Train Other	$9.801 \cdot 10^{0}$	$1.501 \cdot 10^{0}$	$1.078 \cdot 10^{0}$	$9.538 \cdot 10^{-1}$	$8.996 \cdot 10^{-1}$
Train Bias	$4.827 \cdot 10^{-2}$	$7.217 \cdot 10^{-3}$	$5.128 \cdot 10^{-3}$	$4.506 \cdot 10^{-3}$	$4.222 \cdot 10^{-3}$
Validation Total	$5.851 \cdot 10^{0}$	$5.638 \cdot 10^{0}$	$5.485 \cdot 10^{0}$	$5.319 \cdot 10^{0}$	$5.255 \cdot 10^{0}$
Validation Language Model	$5.703 \cdot 10^{0}$	$5.566 \cdot 10^{0}$	$5.418 \cdot 10^{0}$	$5.250 \cdot 10^{0}$	$5.185 \cdot 10^{0}$
Validation Gender	$5.692 \cdot 10^{-4}$	$2.161 \cdot 10^{-4}$	$1.797 \cdot 10^{-4}$	$1.769 \cdot 10^{-4}$	$1.736 \cdot 10^{-4}$
Validation Neutral	0.0	0.0	0.0	0.0	0.0
Validation Other	$1.482 \cdot 10^{-1}$	$7.254 \cdot 10^{-2}$	$6.790 \cdot 10^{-2}$	$6.924 \cdot 10^{-2}$	$6.994 \cdot 10^{-2}$
Validation Bias	$7.067 \cdot 10^{-4}$	$3.334 \cdot 10^{-4}$	$3.075 \cdot 10^{-4}$	$3.123 \cdot 10^{-4}$	$3.145 \cdot 10^{-4}$

TABLE VI: The cross entropy function was used on the English dataset that corresponds to the French dataset.

Metric Loss	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Train Total	$1.528 \cdot 10^{1}$	$7.206 \cdot 10^{0}$	$6.599 \cdot 10^{0}$	$6.261 \cdot 10^{0}$	$6.108 \cdot 10^{0}$
Train Language Model	$5.696 \cdot 10^{0}$	$5.794 \cdot 10^{0}$	$5.568 \cdot 10^{0}$	$5.391 \cdot 10^{0}$	$5.279 \cdot 10^{0}$
Train Gender	$4.684 \cdot 10^{-2}$	$6.568 \cdot 10^{-3}$	$4.652 \cdot 10^{-3}$	$3.841 \cdot 10^{-3}$	$3.579 \cdot 10^{-3}$
Train Neutral	0.0	0.0	0.0	0.0	0.0
Train Other	$9.580 \cdot 10^{0}$	$1.412 \cdot 10^{0}$	$1.031 \cdot 10^{0}$	$8.695 \cdot 10^{-1}$	$8.281 \cdot 10^{-1}$
Train Bias	$4.769 \cdot 10^{-2}$	$6.962 \cdot 10^{-3}$	$5.055 \cdot 10^{-3}$	$4.246 \cdot 10^{-3}$	$4.028 \cdot 10^{-3}$
Validation Total	$5.772 \cdot 10^{0}$	$5.426 \cdot 10^{0}$	$5.256 \cdot 10^{0}$	$5.045 \cdot 10^{0}$	$4.955 \cdot 10^{0}$
Validation Language Model	$5.639 \cdot 10^{0}$	$5.366 \cdot 10^{0}$	$5.207 \cdot 10^{0}$	$4.993 \cdot 10^{0}$	$4.902 \cdot 10^{0}$
Validation Gender	$6.077 \cdot 10^{-4}$	$2.470 \cdot 10^{-4}$	$1.958 \cdot 10^{-4}$	$2.021 \cdot 10^{-4}$	$2.044 \cdot 10^{-4}$
Validation Neutral	0.0	0.0	0.0	0.0	0.0
Validation Other	$1.331 \cdot 10^{-1}$	$6.017 \cdot 10^{-2}$	$4.962 \cdot 10^{-2}$	$5.224 \cdot 10^{-2}$	$5.317 \cdot 10^{-2}$
Validation Bias	$6.540 \cdot 10^{-4}$	$2.901 \cdot 10^{-4}$	$2.377 \cdot 10^{-4}$	$2.494 \cdot 10^{-4}$	$2.535 \cdot 10^{-4}$

TABLE VII: The focal loss function was used on the English dataset that corresponds to the French dataset.

Metric Loss	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Train Total	$8.145 \cdot 10^{1}$	$1.460 \cdot 10^{1}$	$1.190 \cdot 10^{1}$	$1.095 \cdot 10^{1}$	$1.056 \cdot 10^{1}$
Train Language Model	$6.208 \cdot 10^{0}$	$7.044 \cdot 10^{0}$	$7.122 \cdot 10^{0}$	$7.106 \cdot 10^{0}$	$7.097 \cdot 10^{0}$
Train Gender	$3.734 \cdot 10^{-2}$	$3.743 \cdot 10^{-3}$	$2.470 \cdot 10^{-3}$	$1.983 \cdot 10^{-3}$	$1.666 \cdot 10^{-3}$
Train Neutral	0.0	0.0	0.0	0.0	0.0
Train Other	$7.525 \cdot 10^{1}$	$7.553 \cdot 10^{0}$	$4.775 \cdot 10^{0}$	$3.845 \cdot 10^{0}$	$3.460 \cdot 10^{0}$
Train Bias	$3.085 \cdot 10^{-1}$	$3.096 \cdot 10^{-2}$	$1.959 \cdot 10^{-2}$	$1.578 \cdot 10^{-2}$	$1.417 \cdot 10^{-2}$
Validation Total	$6.964 \cdot 10^{0}$	$7.102 \cdot 10^{0}$	$7.080 \cdot 10^{0}$	$7.059 \cdot 10^{0}$	$7.049 \cdot 10^{0}$
Validation Language Model	$6.722 \cdot 10^{0}$	$6.986 \cdot 10^{0}$	$6.985 \cdot 10^{0}$	$6.972 \cdot 10^{0}$	$6.966 \cdot 10^{0}$
Validation Gender	$7.063 \cdot 10^{-4}$	$4.482 \cdot 10^{-4}$	$3.798 \cdot 10^{-4}$	$3.461 \cdot 10^{-4}$	$3.313 \cdot 10^{-4}$
Validation Neutral	0.0	0.0	0.0	0.0	0.0
Validation Other	$2.421 \cdot 10^{-1}$	$1.160 \cdot 10^{-1}$	$9.545 \cdot 10^{-2}$	$8.704 \cdot 10^{-2}$	$8.255 \cdot 10^{-2}$
Validation Bias	$1.110 \cdot 10^{-3}$	$5.538 \cdot 10^{-4}$	$4.578 \cdot 10^{-4}$	$4.174 \cdot 10^{-4}$	$3.965 \cdot 10^{-4}$

TABLE VIII: The non-negative loss function was used on the English dataset that corresponds to the French dataset.

Metric Loss	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Train Total	$7.674 \cdot 10^{0}$	$6.460 \cdot 10^{0}$	$6.414 \cdot 10^{0}$	$6.175 \cdot 10^{0}$	$5.897 \cdot 10^{0}$
Train Language Model	$5.985 \cdot 10^{0}$	$5.694 \cdot 10^{0}$	$5.594 \cdot 10^{0}$	$5.491 \cdot 10^{0}$	$5.440 \cdot 10^{0}$
Train Gender	$7.260 \cdot 10^{-3}$	$3.124 \cdot 10^{-3}$	$3.494 \cdot 10^{-3}$	$2.756 \cdot 10^{-3}$	$1.743 \cdot 10^{-3}$
Train Neutral	0.0	0.0	0.0	0.0	0.0
Train Other	$1.689 \cdot 10^{0}$	$7.665 \cdot 10^{-1}$	$8.201 \cdot 10^{-1}$	$6.843 \cdot 10^{-1}$	$4.577 \cdot 10^{-1}$
Train Bias	$8.207 \cdot 10^{-3}$	$3.691 \cdot 10^{-3}$	$3.979 \cdot 10^{-3}$	$3.288 \cdot 10^{-3}$	$2.180 \cdot 10^{-3}$
Validation Total	$5.594 \cdot 10^{0}$	$5.338 \cdot 10^{0}$	$5.283 \cdot 10^{0}$	$5.190 \cdot 10^{0}$	$5.174 \cdot 10^{0}$
Validation Language Model	$5.508 \cdot 10^{0}$	$5.258 \cdot 10^{0}$	$5.210 \cdot 10^{0}$	$5.118 \cdot 10^{0}$	$5.104 \cdot 10^{0}$
Validation Gender	0.0	0.0	0.0	0.0	0.0
Validation Neutral	0.0	0.0	0.0	0.0	0.0
Validation Other	$8.633 \cdot 10^{-2}$	$7.958 \cdot 10^{-2}$	$7.263 \cdot 10^{-2}$	$7.196 \cdot 10^{-2}$	$6.976 \cdot 10^{-2}$
Validation Bias	$3.453 \cdot 10^{-4}$	$3.183 \cdot 10^{-4}$	$2.905 \cdot 10^{-4}$	$2.879 \cdot 10^{-4}$	$2.790 \cdot 10^{-4}$

TABLE IX: The focal loss function was used on the French dataset

Metric Loss	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Train Total	$5.509 \cdot 10^{1}$	$1.427 \cdot 10^{1}$	$1.119 \cdot 10^{1}$	$1.033 \cdot 10^{1}$	$9.949 \cdot 10^{0}$
Train Language Model	$6.537 \cdot 10^{0}$	$6.903 \cdot 10^{0}$	$6.890 \cdot 10^{0}$	$6.896 \cdot 10^{0}$	$6.881 \cdot 10^{0}$
Train Gender	$5.080 \cdot 10^{-3}$	$6.192 \cdot 10^{-4}$	$3.063 \cdot 10^{-4}$	$4.072 \cdot 10^{-4}$	$2.188 \cdot 10^{-4}$
Train Neutral	0.0	0.0	0.0	0.0	0.0
Train Other	$4.855 \cdot 10^{1}$	$7.364 \cdot 10^{0}$	$4.296 \cdot 10^{0}$	$3.432 \cdot 10^{0}$	$3.067 \cdot 10^{0}$
Train Bias	$1.952 \cdot 10^{-1}$	$2.958 \cdot 10^{-2}$	$1.725 \cdot 10^{-2}$	$1.381 \cdot 10^{-2}$	$1.231 \cdot 10^{-2}$
Validation Total	$6.942 \cdot 10^{0}$	$6.852 \cdot 10^{0}$	$6.813 \cdot 10^{0}$	$6.820 \cdot 10^{0}$	$6.799 \cdot 10^{0}$
Validation Language Model	$6.526 \cdot 10^{0}$	$6.635 \cdot 10^{0}$	$6.654 \cdot 10^{0}$	$6.679 \cdot 10^{0}$	$6.666 \cdot 10^{0}$
Validation Gender	0.0	0.0	0.0	0.0	0.0
Validation Neutral	0.0	0.0	0.0	0.0	0.0
Validation Other	$4.159 \cdot 10^{-1}$	$2.170 \cdot 10^{-1}$	$1.587 \cdot 10^{-1}$	$1.412 \cdot 10^{-1}$	$1.335 \cdot 10^{-1}$
Validation Bias	$1.664 \cdot 10^{-3}$	$8.681 \cdot 10^{-4}$	$6.350 \cdot 10^{-4}$	$5.650 \cdot 10^{-4}$	$5.339 \cdot 10^{-4}$

TABLE X: The non-negative loss function was used on the French dataset