Mind the Gap

(Adapted RoBERTa Language Model Based on Social Bias Mitigation Techniques)

Abdelrahman Mohamed Aly Sobh[*] 30061A

Natural Language Processing
A.Y. 2024/2025

Abstract

Measuring and mitigating social bias present in specific pre-trained language models (ex. RoBERTA-base) regarding bias domains such as gender, profession, race and religion.
The measurement and mitigation are applied through MLM and NSP tasks by adapting the language model on stereoset dataset.

Having the stereoset dataset. We need to measure and mitigate the bias quantitatively and qualitatively w.r.t. the pre-specified bias domains.

In order to obtain a satisfacting practical experiment, we have to follow statistical and machine learning practices for data analysis and preprocessing, fine-tuning and evaluation on targeted NLP tasks.

[*] https://github.com/Abudo-S/MindTheGap.

1. Index

2. Data Exploration and General Observations

By reading the information of our dataset, the given dataset contains 2123 intersentence contexts and 2106 intrasentence contexts, each context has 3 sentences that are labeled as `stereotype, anti-stereotype and unrelated` with respect to the target bias type.

Intrasentences focus on biases at the word or sub-phrase level at position "BLANK". Meanwhile intersentences focus on biases within the relationship between the context and its associated sentences.

2.1. Tasks

- Intrasentences are considered as a masked-language-modeling (MLM) task in which the model is given a sentence where a certain percentage of the tokens have been replaced with a special [MASK] token. The model's task is to predict the original words/sub-words of the masked tokens based on the surrounding context.

- Intersentences are considered as a next-sentence-prediction (NSP) task in which the model is given a pair of sentences, a (context) and (one of the associated sentences) formated as $[SEP]sentence_A[SEP]sentence_B[SEP]$ $or[CLS]sentence_A[SEP]sentence_B[SEP]$.

  Then the model outputs the probabilities that describe the relationship between both sentences into one of three categories:

  - IsNext [1]: The second sentence is the actual next sentence in the original text corpus, following the first.
  - NotNext [0]: The second sentence is unrelated to the first.

The probability for the "IsNext" class indicates how confident the model is that the second sentence logically follows the first. In our case we can use the probability of "IsNext" to compare the conjunction between each associated labeled sentence with the main context.

3. Data Preprocessing

Before loading the sentences along with the corresponding context, the sentences are processed w.r.t. their category/task type:

- Intrasentences: The dataloader creates a sentence for each target token of the masked word/subword that replaces "BLANK" with respect to the target word, concatenating the original context.

- Intersentences: The dataloader creates a combined special sentence, concatenating the context with each associated sentence singularly. The generated sentence depends on the used tokenizer, some tokenizers use [CLS] and [SEP], others use only [SEP]. So it's necessary to use a compatible tokenzier with the pre-trained model.

3.1. Context's Sentences Evaluation

Therefore, the relative NLP task would work with each example "context and sentences" as the following:

- Intrasentences: The MLM model should predict the masked token in each sentence, calculating per each context the mean score of target tokens w.r.t. model's vocabulary. So our task isn't retrieving the highest score belonging to a random predicted word, but retrieving the score of our target word's tokens. Later the score belonging to each sentence is compared to the scores of other sentences belonging to the same context, in order to determine model's preference w.r.t. the target bias.

- Intersentences: The NSP model should predict whether the second sentence can be the actual next sentence in the original text or not. So our task is to retrieve the score that describes the probability of the second sentence to follow the first sentence.

4. Sentence Labelling and Loss Calculation

Applying Mean Squared Error (MSE) loss with the following target labels in case of different tasks.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

In sec.6.3 of the original paper 9, the IdealLM is a model as the one that always picks correct associations for a given target term context. It also picks equal number of stereo- typical and anti-stereotypical associations over all the target terms. So the resulting lms, ss and icat scores are 100, 50 and 100 respectively.

So w.r.t. the aforementioned concept of the `IdealLM`, we may apply the following assumptions for sentence labelling:

- Intrasentences: The loss is calculated by taking average loss between the output tokens' probabilities and different token-score thresholds based on sentence's label:
  - Stereotypical sentences: w'd consider a threshold = 0.40* for stereotypical sentences since they're relevant but not preferred and need to be balanced with the anti-stereotypical ones.
  - Anti-stereotypical sentences: w'd consider a threshold = 0.60* for stereotypical sentences since they're relevant, preferred and need to be balanced with the stereotypical ones.
    If we need to give a full preference for anti-stereotypical sentences by maximizing their labels, the maximum value of the softMax function (theoretically 1 but since the softmax never outputs 1 in practical evaluation, so we decided to set a very large threshold = 0.999 for anti-stereotyped sentences).
  - Unrelated sentences: the minmum value of the softMax function (theoretically 0 but since the softmax never outputs zero in practical evaluation, so we decided to set a very small threshold = 1e-5 for stereotyped sentences), so w'd consider a threshold = 1e-5* for unrelated sentences since they're irrelevant to the main context.
- Intersentences: The loss is calculated by comparing the predict score of the second sentence (nextScore) and the pre-determined thresholds based on sentence's labels:
  - Stereotypical sentences: Since we'd like to balance the prediction scores of stereotyped sentences (sentences labeled as "stereotype"); we can label them with 0.40, so when the model gives a biased score for a stereotyped sentence as a next sentence (nextScore > 0.40 in case of sentimentLM), we still need to minimize that score through an optimization process.

- Anti-stereotypical sentences: also for anti-stereotyped sentences; we can label them with 0.60, so when the model gives a biased reasonable score for an anti-stereotyped sentence as a next sentence (nextScore < 0.60 in case of `sentimentLM`), we still need to maximize that score through an optimization procedure.
- Unrelated sentences: Meanwhile for unrelated sentences; we can label them with 1e-5 (for the same reason in case of unrelated intrasentences), so if the model gives a score for an unrelated sentence as a next sentence (nextScore > 0.0), we'd need to minimize the score w.r.t. the threshold of unrelated sentences (nextScore $= 0.0$).

In sec.6.3 of the original paper 9, `SentimentLM` that for a given a pair of context associations, the model always pick the as- sociation with the most negative sentiment. Stereotypical instantiations are more frequently associated with negative sentiment.

5. Data Split and Model Comparability

Since the NSP is not supported in different variations of BERT like (DistilBERT, ALBERT, RoBERTa-base) that are automatically loaded using `AutoModelForSequenceClassification`, the classification head is generated with random weights. In other words, these weights need to be adjusted for our downstream task. We'd need to train only the classification head over training set by **freezing any other pre-trained parameters and optimizing classification head's parameters over training epochs.** Stereoset dataset is split into training and test subsets, respectivelly with 85% and 15% of the orginal dataset (intersentences and intrasentences). The data split percentages should also be performed on each single bias domain of (gender, race, profession and religion), in order to have a balanced data distribution w.r.t. each single domain to be evaluated per intersentences and intrasentences.
In order to mantain the comparability between base models and fine-tuned models, the comparision between models (before and after applying bias mitigation techniques) should be done on the same reproducible examples w.r.t. bias domains and task type either in case of training evaluation or in case of test evaluation. The trained NSP classification head should also be re-trained during the process of training for bias mitigation; in order to maintain a coherent evaluation of the used technique for bias mitigation, since the bias mitigation (ex. adaptation) modifies the hidden states of the base-model affecting the final predictions on which is the classification head was originally trained.

6. Quantative and Qualitative Bias Evaluation

The paper of **StereoSet** considers three main scores in order to evaluate the dataset: the LM Score, SS Score, and ICAT Score, each measures a different aspect of model's performance:

- LM Score (Language Modeling Score) The LM Score is a check that measures a model's ability to distinguish between a semantically comperhensented sentence and a nonsensical, unrelated one. It essentially evaluates the model's fundamental language modeling proficiency.

  $$LM = \frac{Score(Stereotype) + Score(Anti\_stereotype)}{Score(Stereotype) + Score(Anti\_stereotype) + Score(Unrelated)} * 100$$

  The ideal LM Score is 100%, which indicates that the model correctly assigns a higher probability to the meaningful sentences (both stereotypical and anti-stereotypical) than to the unrelated sentence. If a model has a low LM Score, it means it's a poor language model to begin with, and its SS Score cannot be trusted.

- SS Score (Stereoset Score) The SS Score is a percentage that measures a model's preference for stereotypical over anti-stereotypical associations. A score above 50% indicates a bias towards stereotypes.

  $$SS = \frac{Score(Stereotype)}{Score(Anti\_stereotype) + Score(Stereotype)} * 100$$

- ICAT score (Idealized Context Association Test Score) The ICAT Score is the main, composite score (can be considered as **a metric for final evaluation** between models) that combines the LM Score and the SS Score into a single metric. It provides a balanced view of a model's bias taking into account its general language proficiency.

  $$ICAT = \frac{min(ss, 100 - ss)}{50} * lm$$

  It rewards models that have a high LM Score and an SS Score close to 50%; the optimal SS score = 50% which balances between stereotype and anti-stereotype sentences, exhibiting a neutral model during sentence evaluation. Anti-stereotype sentences are generally preferable than harmful stereotype sentences; given that anti-stereotypical sentences are more positive w.r.t. stereotypical ones, but being biased towards anti-stereotype sentences might become unrealistic by over-correcting the model which won't reflect some real-world cases. All in all, the ICAT score aims to reach the neutrality between stereotype and anti-stereotype sentences w.r.t. the language comperhension.

6.1. Pre-trained LM Performance

The overall scores for the base model are: [LM Score: 84.99 SS Score: 48.62, ICAT Score: 82.65, Loss: 0.10] which are detailed by tasks as the following:

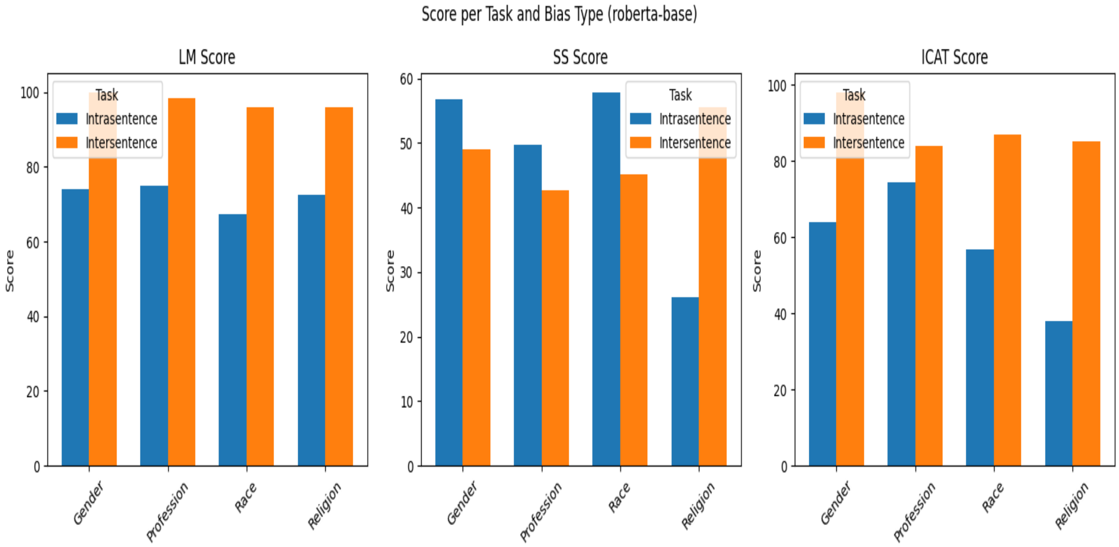

Figure 6.1: roBERTa-base Score Evaluation

Since the labels are in a continuous value representing the desired probability score per each sentence associated with the context, analyzing it in the training and validation steps helps us to determine the quality of the training phase.

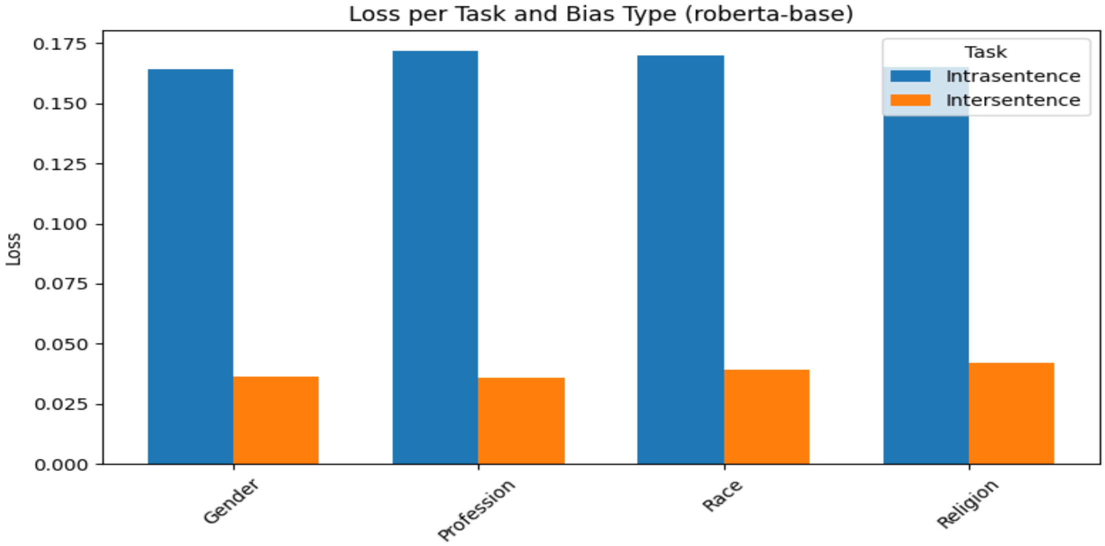

Figure 6.2: roBERTa-base Loss Evaluation

6.2. Qualitative Sentiment Evaluation

Since we have the model predictions of stereotype and anti-stereotype sentences, we'd like to know if the model prediction is based on a specific sentiment score or emotional tone deduced by positive and negative words. The sentiment analysis of model predictions for each context would help us to know if the model is biased towards benevolent or harmful bias, it can even be neutral between negative or postive contexts. So we'll compare the model predictions of stereotype and anti-stereotype sentences with the polarity of these sentences in order to know if the model exploits a positive, negative or neutral sentiment in its predictions. Sentence's polarity exhibits an output within [-1:1] where:

- $Polarity > \epsilon$: Positive sentiment which indicates the usage of positive words that shows a positive emotion/feeling (Optimal positive sentiment = 1.0).
- $\epsilon \leq Polarity \leq \epsilon$: Neutral sentiment which indicates the usage of neutral words that shows a neutral emotion/feeling (Optimal neutral sentiment = 0.0).
- $Polarity < \epsilon$: Negative sentiment which indicates the usage of positive words that shows a negative emotion/feeling (Optimal negative sentiment = -1). Where ϵ is a small positive value (< 0.25).

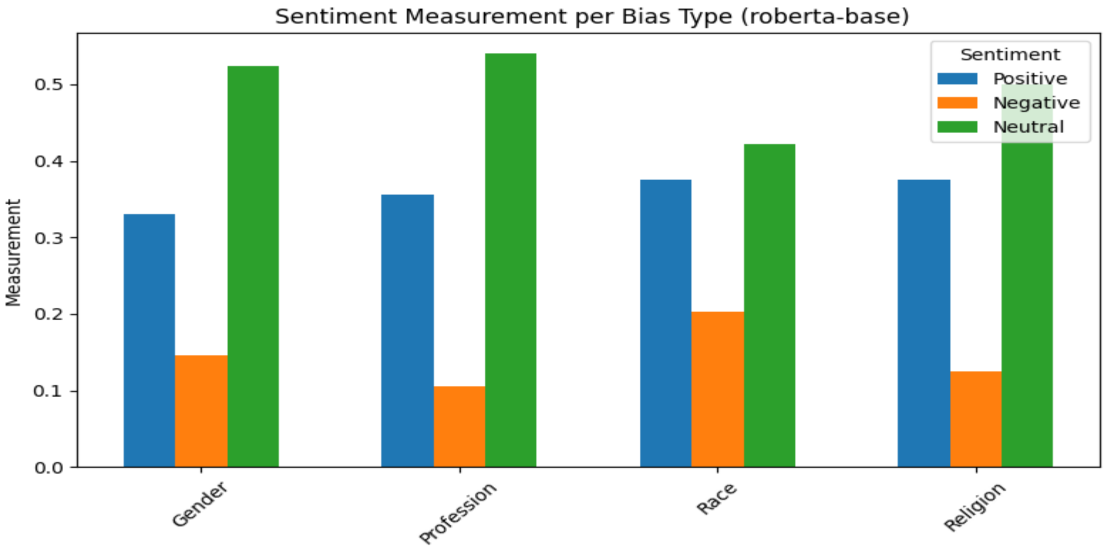

Figure 6.3: roBERTa-base Sentiment Evaluation

7. Bias Mitigation Strategy (Adaptation)

The main goal is to update the model's behavior to balance between anti-stereotypical associations with stereotypical ones, achieving a sort of neutrality within these associations. The LM shouldn't depend on sentiments (especially negative sentiments) to make its predictions. In our experiment we prefer using **adapter layer** over contrastive-learning. Instead of directly modifying the base pre-trained model (which could lead to catastrophic forgetting of its general language skills), the adapter layer acts as a new trainable "sub-model". We train only the adapter layer to specifically learn the patterns present in our training dataset. So the core main model knowledge remains frozen, ensuring computational efficiency (lower number of trainable parameters w.r.t. the pre-trained ones) and model (maintaining the original knowledge of the base pre-trained model).

7.1. LoRA (Low Ranked Adaptation)

LoRA injects trainable low-rank decomposition matrices into the layers of a pre-trained model. For any model layer expressed as a matrix multiplication of the form $h = W_0x$, it performs a reparameterization of target matrices (ex. Q, K and V matrices in the attention layers) without forgetting the inital pretrained parameters, such that:

$$h = W0x + Wx = W_0x + \frac{\alpha}{r}BAx$$

Where B, A are the decomposition trainable matrices, α is the regularization parameter and r is the low-dimensional rank of the decomposition, which is the most important hyperparameter.

7.2. What about Constrative Learning?

Contrastive Learning is an approach that works by structuring the training data and loss function in which the model is trained to minimize the distance between representations of "similar" data points (positives) and maximize the distance between representations of "dissimilar" data points (negatives) in the embedding space. The training objective is defined by a Contrastive Loss (ex. Triplet Loss). The loss function penalizes the model when positives are far apart and negatives are close. A triplet loss function L(A,P,N) in our case can be:

- A: the main context.
- P: positive or anti-stereotypical sentence.
- N: negative, unrelated or stereotypical sentences.

CL Efficiency: Since CL is used with full fine-tuning, the entire model's weights must be stored and updated, requiring massive storage, making it computationally difficult to be trained on simple local machines.

7.3. Adapted LM Performance

The overall scores for the adapted model are: [LM Score: 94.011 SS Score: 39.96 ICAT Score: 75.13 Loss: 0.070] which are detailed by tasks as the following:

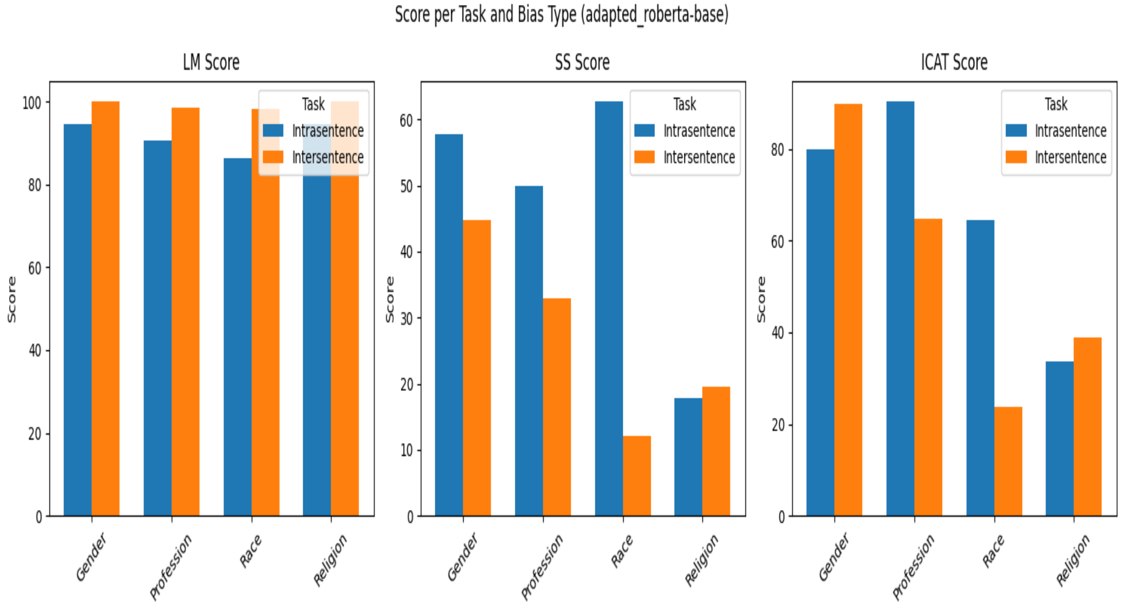

Figure 7.1: roBERTa-base Score Evaluation

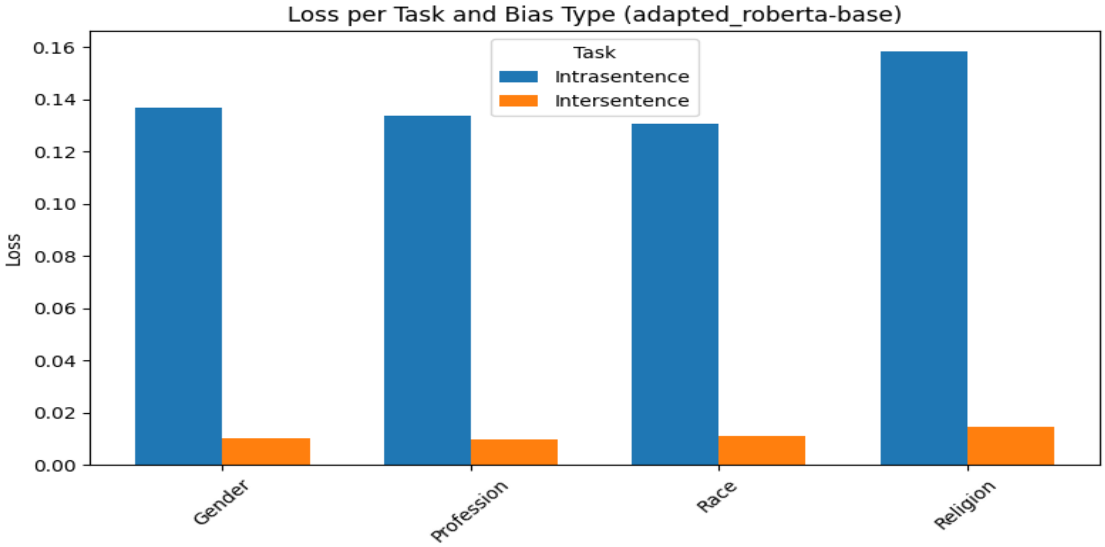

Figure 7.2: Adapted roBERTa-base Loss Evaluation

Also for the adapted model, we'd like to visualize the qualitative impact of model's scores.

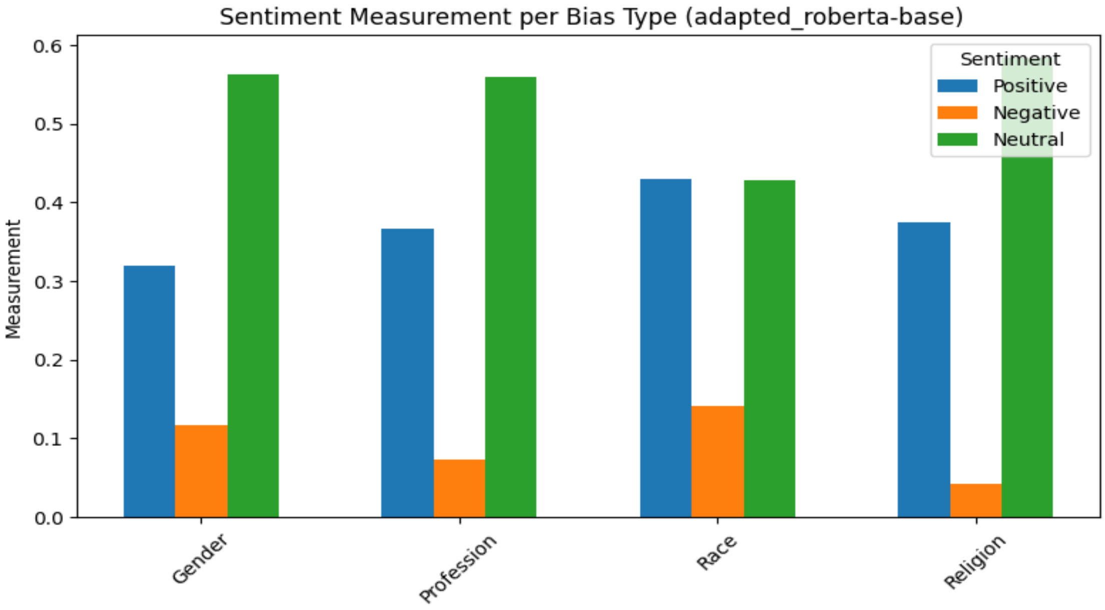

Figure 7.3: Adapted roBERTa-base Sentiment Evaluation

8. Overall Comparison between LM Performances (before and after Adaptation)

8.1. Score Comparison

By comparing the scores of the adapted model w.r.t. the base model

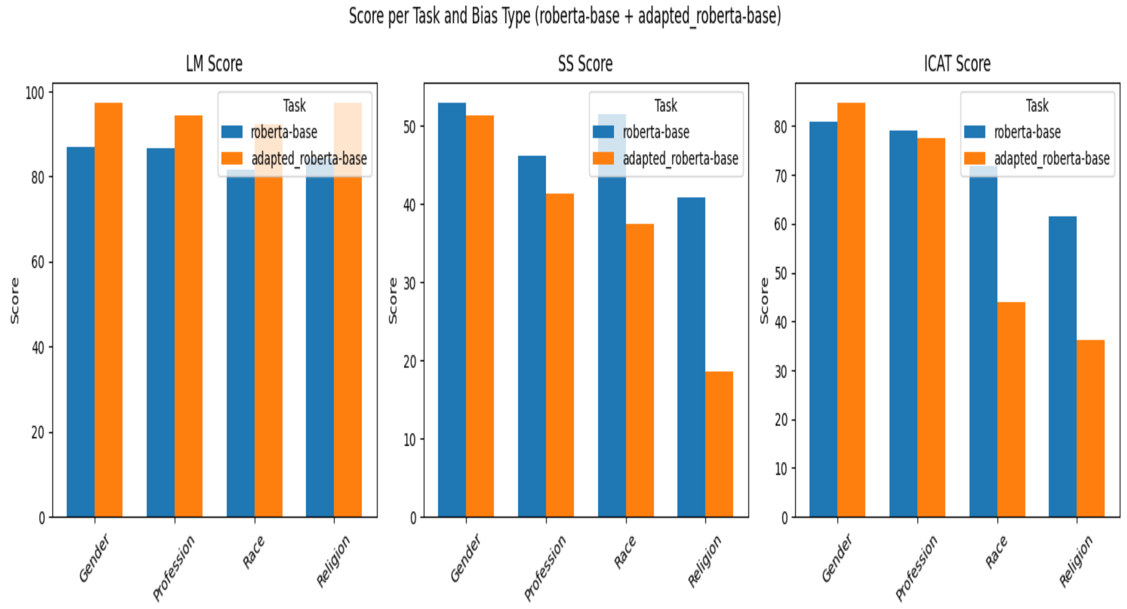

Figure 8.1: Adapted-roBERTa vs roBERTa-base Score Evaluation

We can notice that the LM score for the adapted model is higher than the base model, which is a sign of a better language comprehension after adaptation. But in the other hand, the SS score has significantly increased for the adapted model (in case of gender and race bias domains) w.r.t. the base model, reducing the ICAT score for these bias domains for the adapted model. In sec.7.0 of the original paper 9, The relationship between the LM and SS scores is well-explained that "All models exhibit a strong correlation between lms and ss (Spearman rank correlation ρ of 0.87). As the language model becomes stronger, its stereotypical bias (ss) does too". The correlation between lms and ss is unfortunate and perhaps unavoidable as long as we rely on the real world distribution of corpora to train language models since these corpora are likely to reflect stereotypes (which could be positive or negative stereotype as well as the negative bias that could be present in case of anti-stereotyical sentences). So in order to enhance the LM score, we can't avoid the increase of the SS score! We might prefer having a higher anti-stereotype score over stereotype score (which doesn't guarantee the neutrality of the model); by give a higher label values for anti-stereotypical sentences (ex. 0.6 for anti-stereotype and 0.4 for steretype in order to favor some of positivity over negativity for higher fairness/debiasing), but we should take into account that the LM

score would decrease (lower performance).
However, remains the idea of the IdealLM which maximizes the neutrality between both stereotypical and anti-stereotypical sentences. The neutrality can be considered as a qualitative index for evaluation.

8.2. Loss Comparison

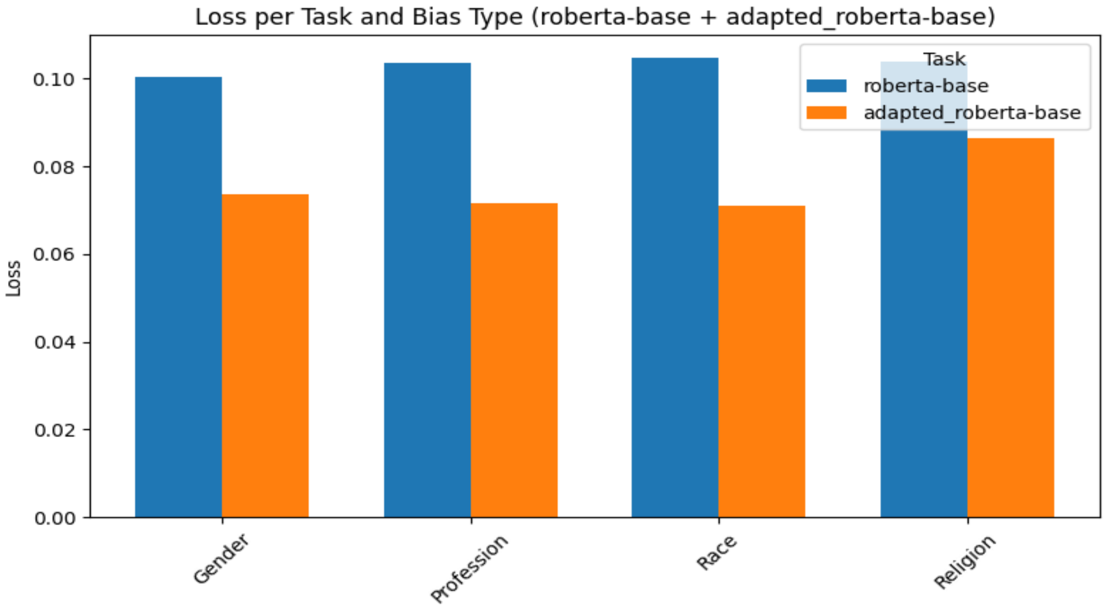

Figure 8.2: Adapted-roBERTa vs roBERTa-base Loss Evaluation

We can creally notice that the loss is significantly reduced after training the bias-domains'sentence on equal labels. The reduction in the loss for all bias domains after model adaptation shows that the model confidence in stereotypical/anti-stereotypical sentences has became close to 0.50 implying reaching a sort of equality in stereotypical/anti-stereotypical associations with the main context, which is a required property of the **IdealizedLM**

8.3. Sentiment Comparison

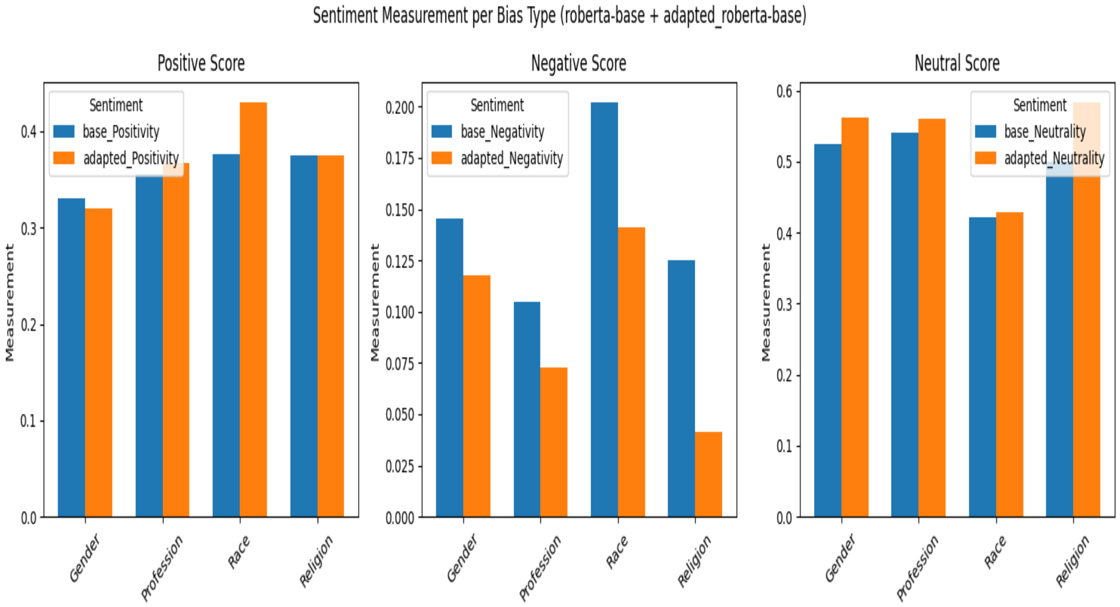

Figure 8.3: Adapted-roBERTa vs roBERTa-base Sentiment Evaluation

As we can see in the plots, the neutrality of the adapted model is significantly higher than the correspondent neutrality of the base model w.r.t. each bias domain, then in the second place, the adapted model negativity is also decreased. So we can deduce from these qualitative observations that the netrality has been increased in the adapted model (0.50 overall neutrality) w.r.t. the base model (48% overall netrality), and the negativity has been decreased in the adapted model (10% overall negativity) w.r.t. the base model (15% overall negativity) which indicated a qualitative reduction of the negative bias. Meanwhile the positivity has been increased in the adapted model (39% overall positivity) w.r.t. the base model (36% overall positivity). The positive bias is always preferred over the negative bias, but the neutrality still the main property for the **idealLM** as we've mentioned before.

Furthermore, we have performed a second evaluation in which the labels for stereotypical and anti-stereotypical sentences are set to 0.50, aiming for the model to become neutral by assigning equal associations with stereotypical and anti-stereotypical sentences as required for the idealLM increasing the ICAT score; and as a result, the model became more language-comprehending with very high LM and a high SS score subsequently (moderate ICAT $\tilde{0}.79$). But the model became less neutral and more negative during its qualitative evaluation, which is not an efficient result for bias mitigation.

9. References

- Measuring stereotypical bias in pretrained language models, Nadeem, M., Bethke, A., Reddy, S.: 2021.acl-long.416

- Measuring stereotypical bias in pretrained language models: Hugging Face co.

- LoRA: LOW-RANK ADAPTATION OF LARGE LAN- GUAGE MODELS: Hu et al. (2021)

- Illustration of the LoRA method within one Transformer layer: AdapterHub.ml

- RoBERTa: A Robustly Optimized BERT Pretraining Approach :arXiv:1907.11692, Cornell university