Leveraging Domain Knowledge for Inclusive and Bias-aware Humanitarian Response Entry Classification – Appendices

A Hyperparameters

Table 1 reports the results of hyperparameter tunning.

B Additional Results

Figures 1-5 report the results of gender and country bias measurement over various backbone LLMs, and architectures, before after applying CDA bias mitigation.

Hyperparameters	Values
Number of Epochs	3
Initial Learning Rate	1e-4
Dropout Rate	0.2
Train Batch Size	8
Validation Batch Size	16
Optimizer	Adam Weight (with the standard Pytorch (https://pytorch.org/) hyperparameters)
Learning Rate Scheduler	Pytorch StepLR (with decay=0.4, step size=1)
LLM input text max length	200
Freezed LLM layers	LLM Embedding and first LLM layer
Decision boundary threshold	Finetuned differently for each training setup and tag on the best F1 score validation set after training (from 20 values ranging from the minimum to the maximum probability predicted for each tag).

Table 1: Hyperparameters used for finetuning Classification models and for Generating final predictions

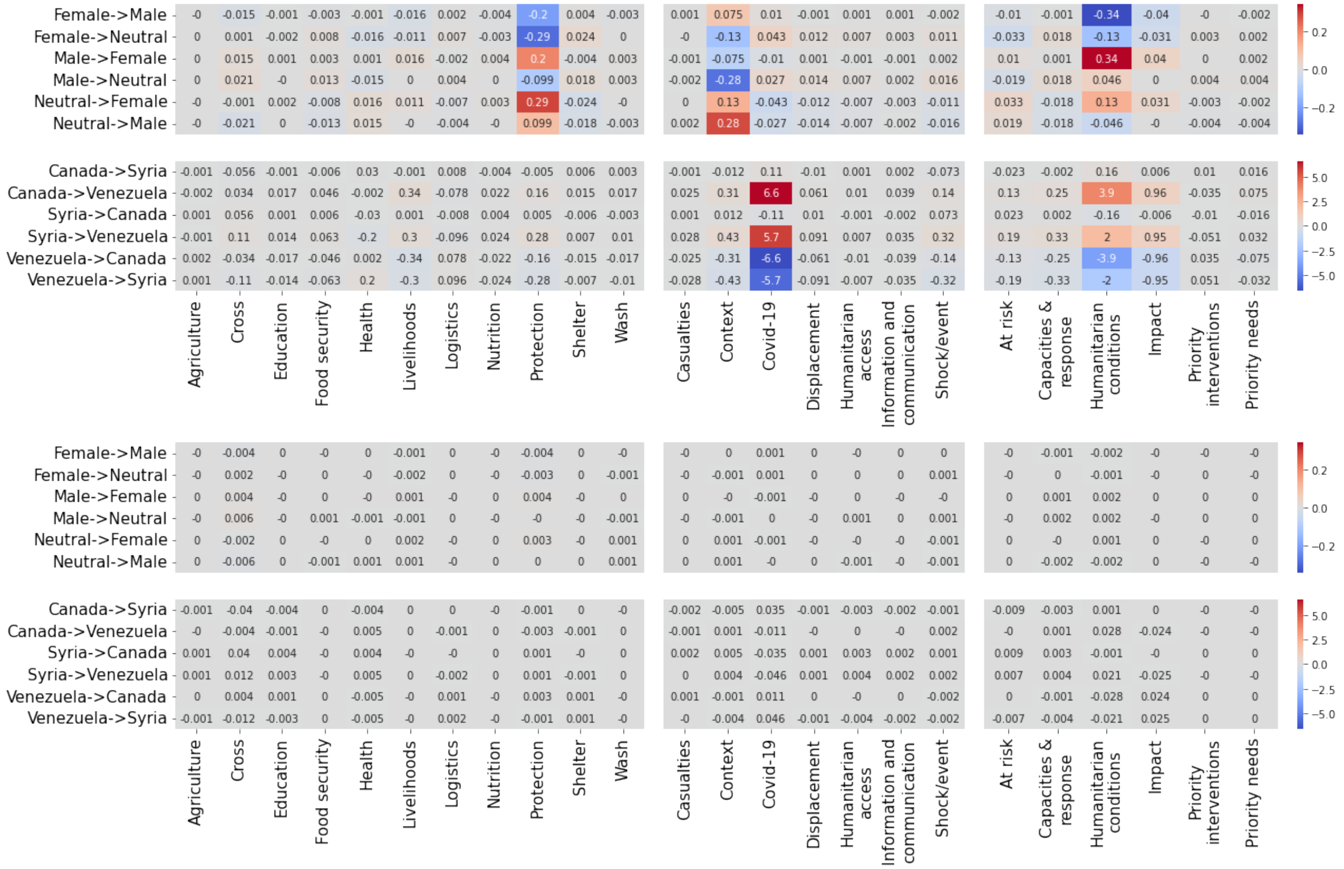

Figure 1: The results of Tag-Shift bias metric for BASE architecture using the HumBERT as the backbone. (Top) Original model without debiasing; (Bottom) Counterfactual debiasing.

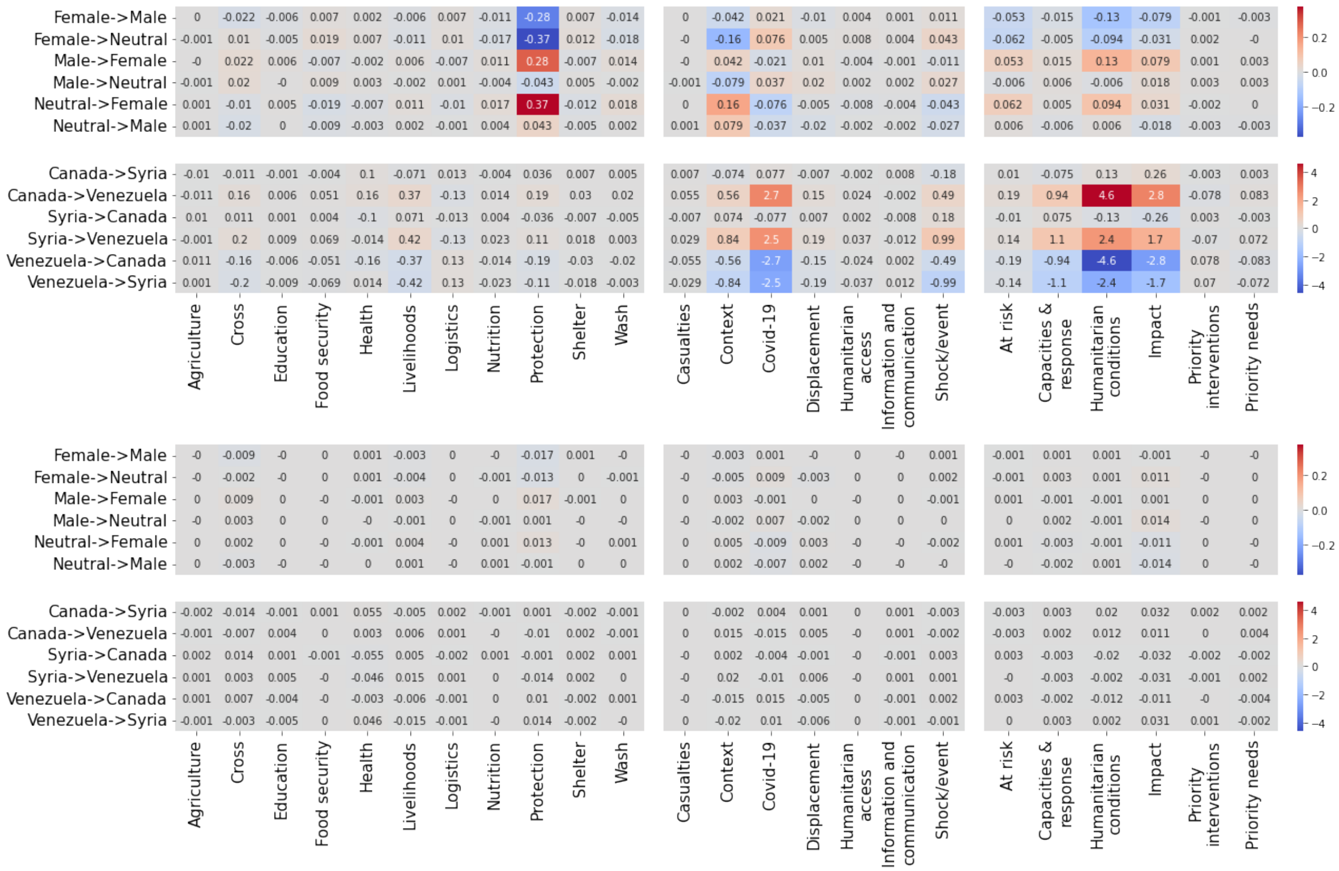

Figure 2: The results of Tag-Shift bias metric for BASE architecture using the XLM-R as the backbone. (Top) Original model without debiasing; (Bottom) Counterfactual debiasing.

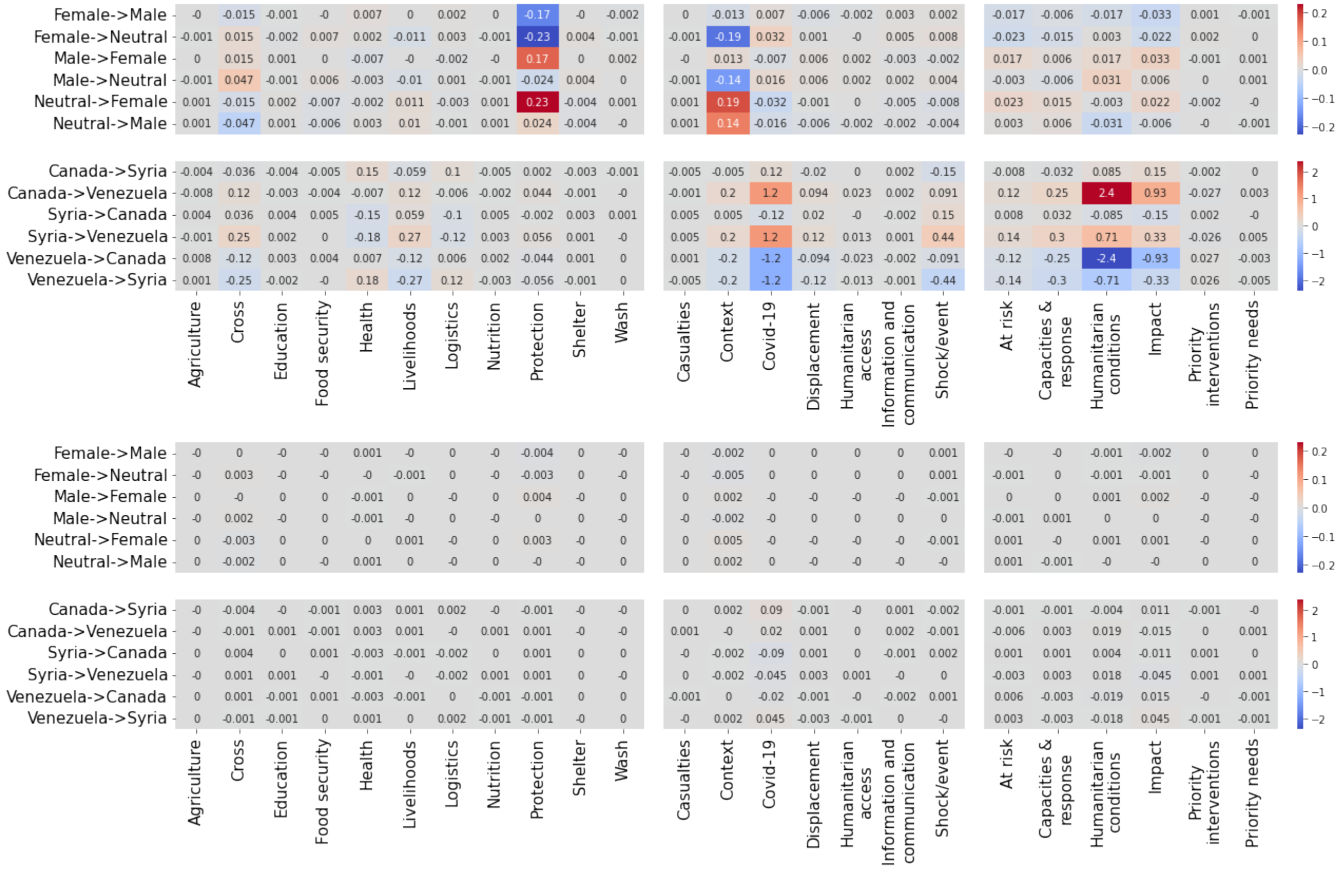

Figure 3: The results of Tag-Shift bias metric for OURS architecture using the XLM-R as the backbone. (Top) Original model without debiasing; (Bottom) Counterfactual debiasing.

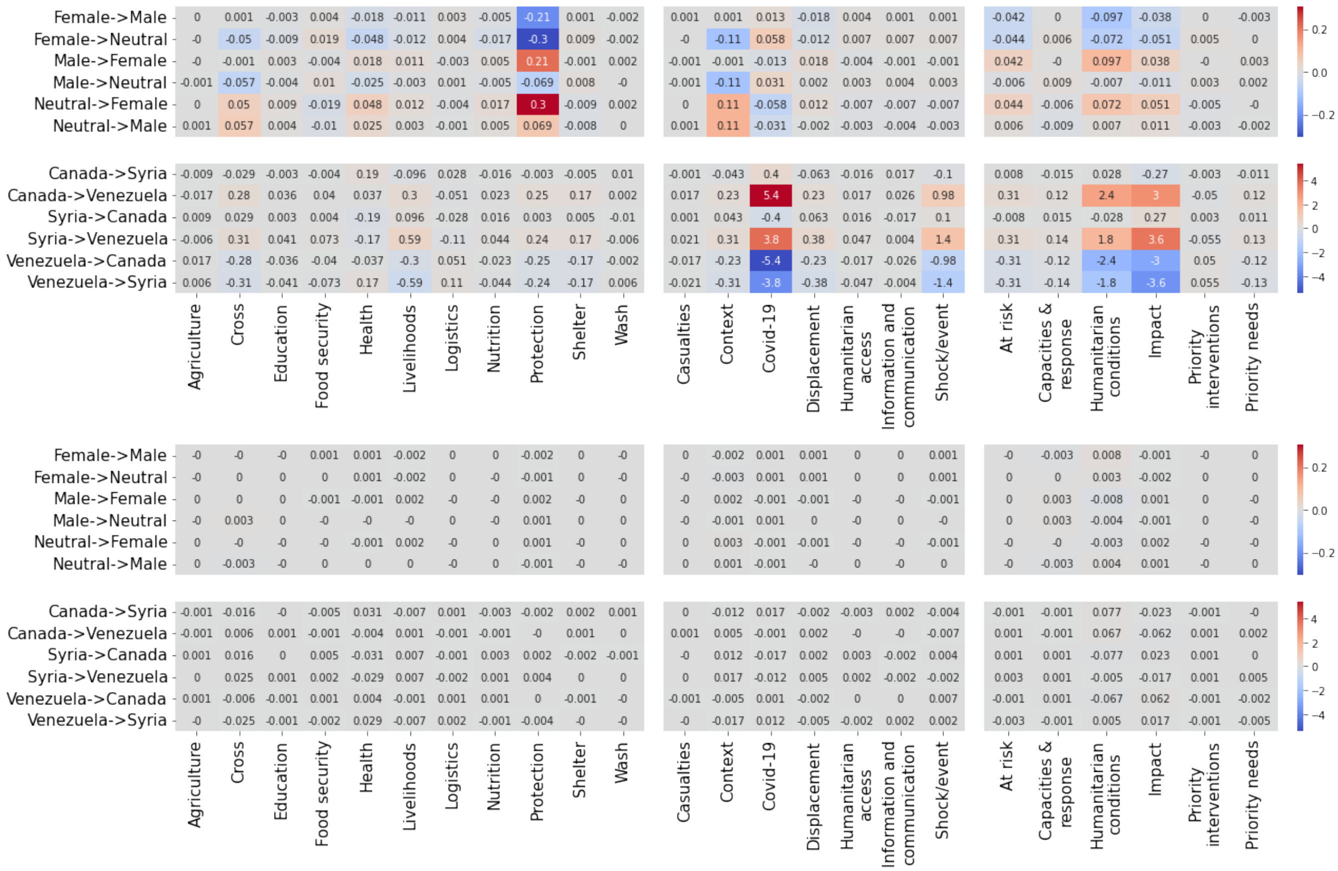

Figure 4: The results of Tag-Shift bias metric for BASE architecture using the m-BERT as the backbone. (Top) Original model without debiasing; (Bottom) Counterfactual debiasing.

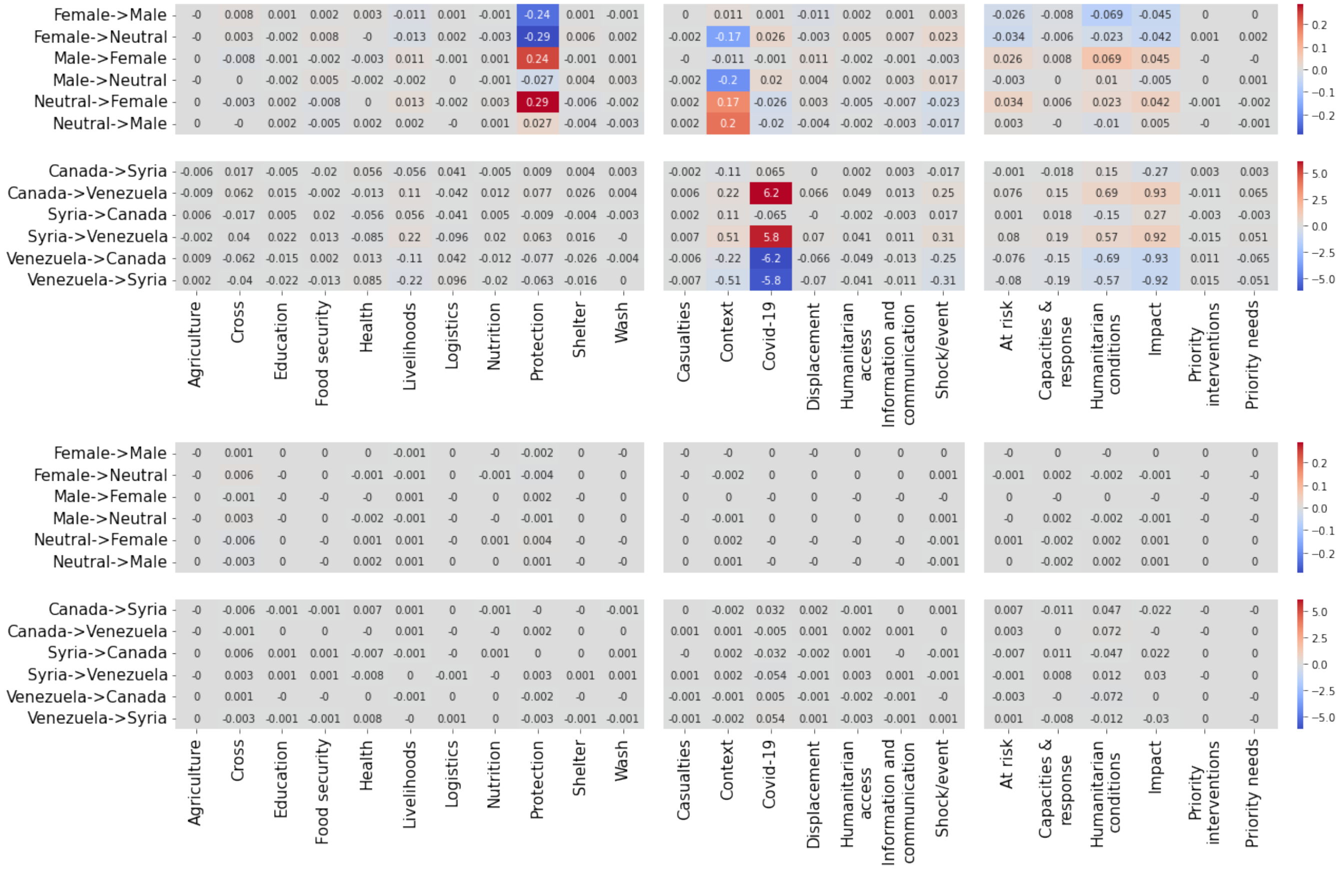

Figure 5: The results of Tag-Shift bias metric for OURS architecture using the m-BERT as the backbone. (Top) Original model without debiasing; (Bottom) Counterfactual debiasing.