Gender and Societal Bias Measurement and Mitigation for Humanitarian Text Classification

Student: Selim Fekih: 288330
Supervisor: Assistant Professor Robert West
External Advisor: Dr. Navid Rekabsaz

Contents

Abstract

As Artificial Intelligence (AI) and Natural Language Processing (NLP) are increasingly used in the humanitarian sphere to improve the humanitarian response, it is paramount to ensure that the AI models do not contain societal and demographical stereotypes that can deteriorate the quality of the aid. This comes under one of the core principles of the humanitarian sphere: the Leave No One Behind (LNOB) principle. This work aims to measure and mitigate the societal biases encoded in the humanitarian classification models for an ethical humanitarian response. We tackle this problem by (1) creating and releasing a new classification dataset used for both bias measurement and bias mitigation, (2) proposing four bias measurement techniques, including extrinsic and intrinsic metrics, (3) testing in-processing and pre-processing bias mitigation methods, and (4) conducting experiments and detailed analysis on the biases encoded in models. Our experiment results show an improvement in classification results and a decay in the bias contained in the classification models, reflected in all the bias measurement metrics proposed. Finally, we see that the amount of bias differs depending on the tag and the population.

A partial part of this research has been published on the proceedings of the International Joint Conference of Artificial Intelligence 2023 (IJCAI 2023), on which the author of this thesis (Selim Fekih) is a co-author[Tamagnone et al. 2023].

1 Introduction

Machine Learning (ML), Deep Learning (DL) for Natural Language Processing (NLP) and the more recently released Large Language Models (LLM) have produced a lot of real-world applications. From translation to question answering, by text classification and summarization, NLP has significantly impacted different sectors, such as insurance [Medium n.d.(b)] and social media [Medium n.d.(a)]. The humanitarian sector also benefited from these advancements [Rocca et al. 2023], thanks to the various projects in this domain, one of which is the Data Entry and Exploration Platform (DEEP)[1]. This platform is the largest repository of humanitarian annotated data. It also contributes to the advancement of NLP for the humanitarian sector by releasing and maintaining an entry classification dataset: the HUMSET [Fekih et al. 2022]. This dataset was manually annotated and verified by humanitarian analysts and is based on past humanitarian crises (occurring between 2018 and 2021). These humanitarian crises are of different types [Iberdola n.d.]. Some crises are socio-economical and include displacement of population. Others can be linked to hunger, lack of essential services, environmental and many other causes. Different demographic groups and countries and geographical areas in the world have not been impacted equally by various disasters, thus producing potentially biased training datasets. Knowing that the LLM backbones finetuned for classification tasks already contain intrinsic bias from the initial training [Rekabsaz and Schedl 2020; Rekabsaz, Kopeinik, and Schedl 2021], we see the potential bias problem encoded in the humanitarian classification models. As part of the neutrality principle of humanitarian interventions, it is primordial to ensure models don't respond differently depending on the geographical area mentioned or the specific population impacted. In other words, ensuring that the humanitarian classification models do not contain any societal biases and stereotypes is vital. Biased results can be translated into a wrong perception of the situations where humanitarian organisations work and cause an inadequate humanitarian response.

The objective of the research work conducted during this master thesis is *to define, measure, and mitigate possible societal and harmful biases reflected in the Humanitarian Classification models while preserving a good classification performance.*

To this end, this work contains four contributions. We first introduce HUMSETBIAS, an extension to HUMSET, which has labelled examples of Gender and Country attributes, focusing on three bias labels each (Female, Male, Neutral and Canada, Venezuela, and Syria, respectively). HUMSETBIAS is very helpful for studying and mitigating societal biases in the Humanitarian domain.

[1] https://thedeep.io/

The second contribution is the measurement of bias inside the LLM classification models with four metrics: intrinsic metrics (the LLM Embeddings Distance and the Salience Discrepancy) and extrinsic metrics (the Prediction Flips and the Probabilities Discrepancy). Thirdly, we present two techniques to address the bias contained in models: a preprocessing-based method (the Counterfactual Data Augmentation CDA [Zhao et al. 2018; Lu et al. 2018]) and an in-processing approach (the Focal Loss FL [Lin et al. 2018] based training). Finally, we conduct experiments to evaluate and study the biases and bias-mitigation process and analyse the classification and bias results on the system level and tag level. Our full implementation can be found in **https://github.com/sfekih/bias-measurement-mitigation-humanitarian-text-classification**

Our work is structured as follows: Firstly, we start by giving a general overview of the HUMSET followed by the presentation of HUMSETBIAS, a dataset which consists of an extension of HUMSET, used to measure and mitigate the bias present in the classification models. Secondly, we present two extrinsic metrics (Tag Flips, Predictions Discrepancy) and two intrinsic metrics (Salience Discrepancy, Embeddings Distance) used for bias measurement and give three bias mitigation techniques (Focal Loss training, Counterfactual Data augmentation and their combination). Fourthly, we present our experimental setup and assess each mitigation method on the classification performance. Finally, we conduct a detailed analysis of the bias contained in the models with each training setup, analyse how bias measurement techniques relate to one another, and conduct a specific study on the particular biases to be handled in a production-like format.

2 Background and Related Work

2.1 NLP for the Humanitarian Domain

Previous work to improve the humanitarian response includes NLP models research, training datasets creation and, lately, a Large Language Model creation.

Datasets Different datasets focused on humanitarian topics have been released to promote the training of models and performing analyses. The HUMSET enters this line of work but is not the only dataset released and publicly available. Other datasets are based on annotated social media data. Some examples include [Alam et al. 2021; Imran, Mitra, and Castillo 2016] on Crises related datasets, and [Adel and Wang 2020; Alharbi and M. Lee 2021] on Arabic Twitter based datasets for classification

Models Different NLP techniques for different use cases have been developed to facilitate the humanitarian response process. First, [Yela-Bello, Oglethorpe, and Rekabsaz 2021] develops Deep Learning-based extractive summarization models to extract relevant information from humanitarian documents. Then, [Lai et al. 2022] created Named Entity Recognition models to extract flooding information and reduce their inducted risks. Lastly and most recently, [Tamagnone et al. 2023] proposes a zero-shot classification pipeline and new classification architectures to improve the classification results on labelled and unlabelled data.

LLM Backbone: HumBERT From the first emergence of LLMs with BERT [Devlin, M. Chang, et al. 2018], many new transformers have appeared with many possible usages and specialisations. Nowadays, LLMs can be trained and finetuned on any corpus of data, and HumBERT[1] has emerged as a new language model explicitly trained on a humanitarian corpus. This new model has proven to improve the performance on downstream humanitarian classification tasks compared to other models of similar size and complexity [Tamagnone et al. 2023].

2.2 Bias Measurement and mitigation in NLP

Awareness of the biases encoded in models is critical. It ensures their correct usage and consciousness of their potential problems, mainly when deployed and subject to real-world scenarios and

[1] https://huggingface.co/nlp-thedeep/humbert

data. Over the past years, many bias measurement and debiasing techniques have emerged for all Machine Learning related tasks, NLP being no exception.

Bias Measurement Bias measurement for NLP tasks consists of Extrinsic and Intrinsic (also referred to as Explicit and Implicit, respectively). The first refers to the bias measurable by the models' outputs, while the latter measures the biases encoded within the models. An example of an intrinsic bias measurement metric is given by [Liu et al. 2021], which used explainability methods for text classification tasks as a bias measurement technique. Their work focuses mainly on analysing how specific keywords in a sentence can affect the Salience score of other keywords and the overall impact on the classification score.

Demographical bias measurement and models debiasing Debiasing NLP Models has been challenging, especially for the LLMs, which are trained on a large dataset and encode much information. Previous research has been focused on analysing demographical and societal biases encoded into various NLP tasks. For example, similar to our task, [Elazar and Goldberg 2018] measures different demographic group biases for sentiment and Mention-Detection tasks using data Leakage and proposes adversarial training as a debiasing method.

2.3 Focal Loss as a technique to improve classification results in NLP

The Focal Loss [Lin et al. 2018] training consists in an adaptation of the Binary Cross Entropy loss [Zhang and Sabuncu 2018]. It was first designed to tackle the class imbalanceness in computer vision. This loss function has been used since then for different NLP tasks. [Rajič et al. 2022] test this loss function for fighting shallow heuristics on Natural Language Inference tasks and [Usuga-Cadavid et al. 2021] use it to handle the data imbalanceness problem on NLP classification tasks and improve classification performance.

2.4 Explainability for Deep Learning Models

Deep Learning models are usually very complex, with a very high number of parameters. Their interpretation and understanding is not always evident, so they're often referred to as black boxes. Research has been conducted to understand these models and their behaviour. Such models understanding techniques are called model explainability techniques. Different explainability methods have emerged and some research has focused on analysing them. [Ancona et al. 2018; Gholizadeh

and Zhou 2021; Nielsen et al. 2022] present and compare the most relevant explainability methods, with the pros and cons of each one and their applicability depending on the use case.

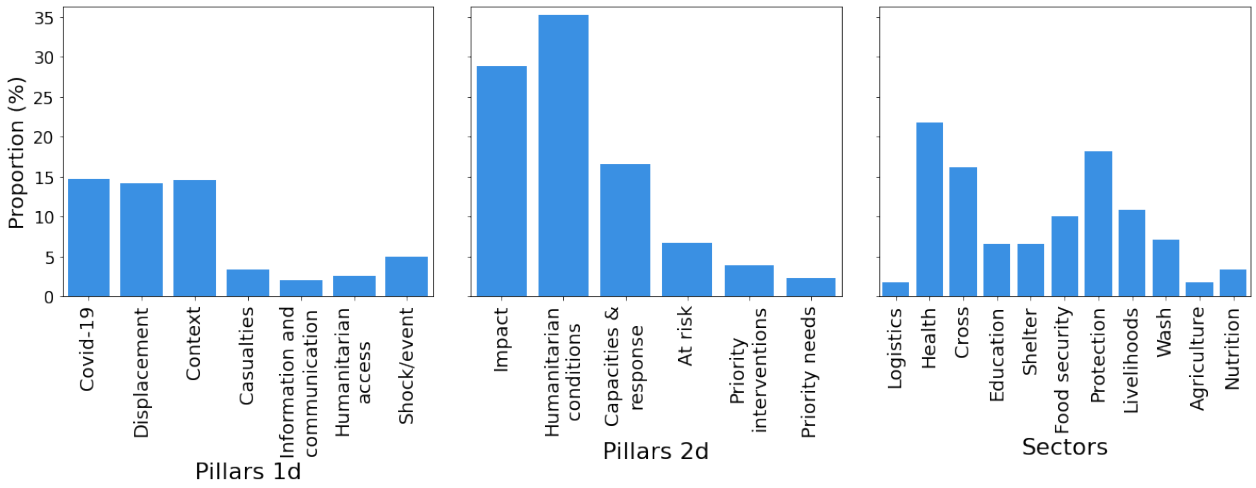


Figure 1: Tags proportions in HUMSET

3 Data

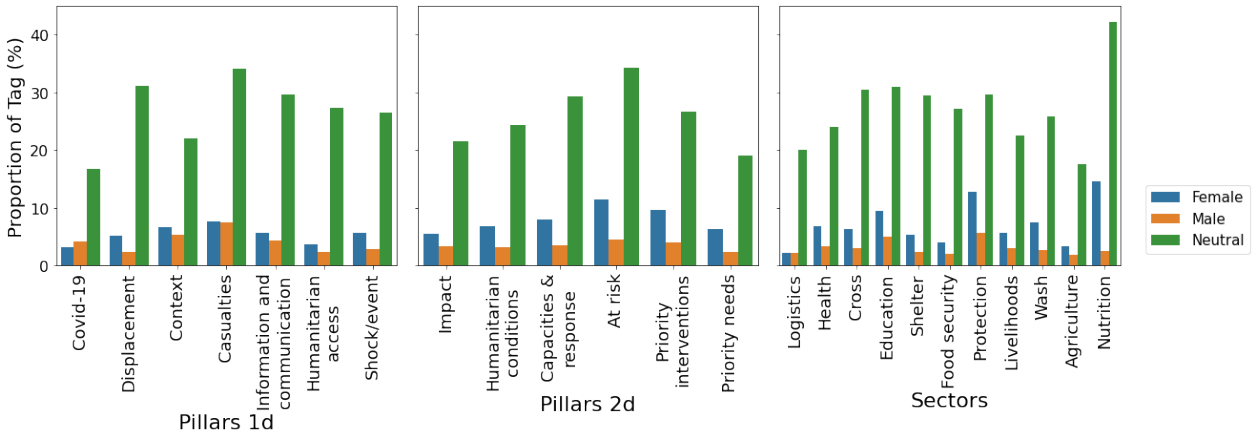
We base our research work on the HUMSET, a dataset publicly available on HuggingFace[1]. We start by doing exploratory Data Analysis on the tags distribution and the bias labels. Then we move to introduce our new dataset, the HUMSETBIAS, which will be used for bias measurement and mitigation.

3.1 HUMSET Analysis

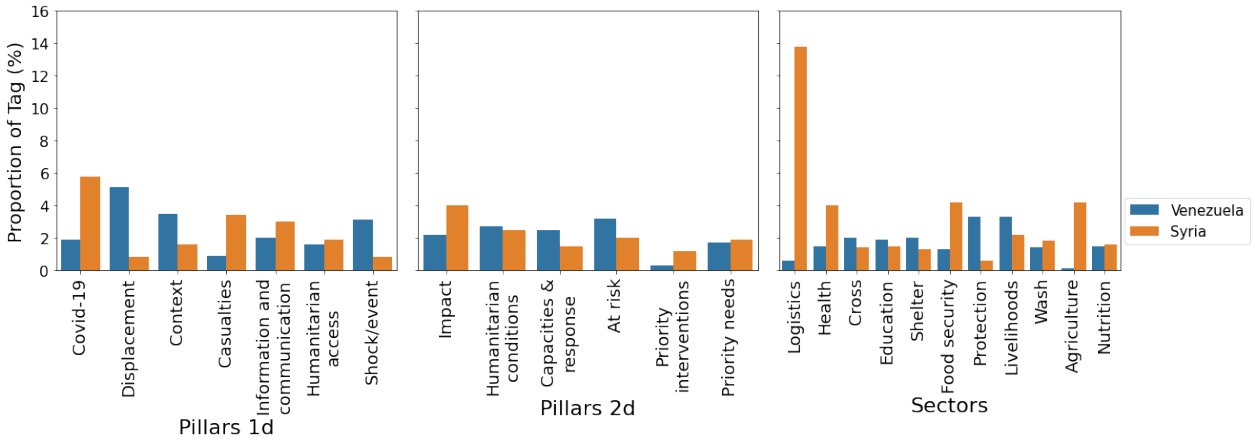
3.1.1 General Analysis

We start by performing a high-level analysis of HUMSET. The dataset contains approximately 154k entries from 46 projects, each generally treating a different humanitarian crisis. It is composed of three languages: English (59.5%), Spanish (20.9%), and French (19.6%). The dataset is split into three categories: train, validation, and test (with a number of entries of 129,268, 11,940, and 12,896 respectively). The dataset comprises five tasks (Sectors, Pillars 1d, Pillars 2d, Subpillars 1d, Subpillars 2d). Each task contains a set of classification tags. For this work, we choose to focus on three tasks only: Sectors (11 tags), Pillars 1d (7 tags), and Pillars 2d (6 tags). Each tag is considered independent of others, and one entry can have any number of tags for each task (including no tags). We start by visualising the tags' proportions in the dataset (Fig 1).

[1]The link to the dataset is the following: `https://huggingface.co/datasets/nlp-thedeep/humset`. We use the second released version in this work (v2.0.0)



(a) Gender



(b) Country

Figure 2: Proportion of entries containing each bias label for each tag

A quick look at Fig. 1 indicates the imbalanced distribution of tags. Proportions range from 1.8% (Agriculture and Logistics Sectors) to 35.3% (Humanitarian Conditions Pillars 2d).

3.1.2 Bias labels analysis

To gain insight into each bias label's presence in the dataset, we visualize the proportion of entries containing each bias label for each tag (Fig 2). A general overview of proportions can indicate the types of crises associated with different attributes.

A quick view of the Gender barplots (Fig 2a) indicated that the Neutral excerpts always have the highest proportions for each tag. Proportions also vary depending on the tag. For example, we see that the Neutral Gender proportions for the Covid-19 Pillar 1d are inferior to 20% but are superior to 40% for the Nutrition sector. Similar insights can be drawn for the Male and Female genders. Each gender is differently present for each tag, with a general tendency of a higher proportion of the Female gender.

On the other side, Countries are more differently mentioned depending on tags. A look at Fig. 2b gives evidence that Syria and Venezuela have different tags mentioned: We see the apparent tendency for the Covid-19 and Casualties, Logistics, Food Security and Agriculture tags for Syria, against the Displacement, Shock/Event and Protection tags for Venezuela.

3.2 HUMSETBIAS

As the distribution of gender and countries is different in the HUMSET and depends on the tag, we suspect that the trained models can themselves contain gender and societal biases. To measure and mitigate the bias encoded in the classification models, we create the HUMSETBIAS. It is focused on two bias attributes: Gender and Country. We start by extracting the excerpts that can be used for the dataset creation. Then we create a set of counterfactual excerpts from the original ones. All the work on the HUMSETBIAS is restricted to English.

The HUMSETBIAS can be found under this link **`https://huggingface.co/datasets/nlp-thedeep/humsetbias`** .

3.2.1 Dataset creation

Gender bias is one of the most centered on biases in literature. In addition to this, we assess the models' biases when the mentioned country is different. For this purpose, we choose two countries that are present enough in the dataset and that were subjected to different humanitarian crises: Syria (for the civil war crisis[1]) and Venezuela (political and socioeconomic crisis[2]).

One possible results falsing scenario we can potentially face is when two bias labels contain the same amount of bias, the final result being a null shift (For example, having the shift from Female to Male null). Concluding that these bias labels don't contain biases would be false and induce errors in interpreting the model. To benchmark the important bias labels (Female, Male for Gender and Syria, Venezuela for Countries), we introduce neutral bias labels (Neutral for Gender and a country not affected by humanitarian crises for Countries, Canada in our example). These new bias labels represent the baselines to which the other bias labels are compared. For this work, we have after all bias attributes with three bias labels each: Female, Male, Neutral for Gender and Canada, Syria, Venezuela for Country).

The set of keywords used for the countries is trivial, being only the country name. For Gender, on the other hand, many keywords can be used. Table 1 shows the list of keywords used for each

[1] `https://en.wikipedia.org/wiki/Syrian_civil_war`
[2] `https://en.wikipedia.org/wiki/Crisis_in_Venezuela`

Female	Male	Neutral
she	he	they
woman	man	person, individual
women	men	persons, individuals
mother	father	person, individual
mothers	fathers	persons, individuals
girl	boy	child
girls	boys	children
her	his	their
her	him	them
female	male	person, individual
females	males	persons, individuals
wife	husband	person, individual
wives	husbands	persons, individuals

Table 1: Keywords indicating gender information and the corresponding mappings for creating counterfactual samples.

Bias Attribute	Bias label	Train	Val.	Test	All
Gender	Female	1,173	92	80	1,345
	Male	604	57	65	726
	Neutral	15,271	1,443	1,524	18,238
	Sum	17,048	1,592	1,669	20,309
Country	Venezuela	261	20	13	294
	Syria	1,686	182	187	2,055
	Canada	0	0	0	0
	Sum	1,947	202	200	2,349

Table 2: Statistics of the HUMSETBIAS dataset for each bias type, as subsets of the train/validation/test sets of HUMSET.

gender bias label. After getting the most important keywords for each bias label, we ensure that the excerpt is context-free apart from the selected keywords. In fact, we disregard excerpts that contain any geolocation other than the original country (cities, other countries, nationalities...) and keywords that refer to only one type of Gender (pregnant, lactating...). This cleaning work keeps the HUMSETBIAS clean and avoids model confusion while training. The total number of excerpts extracted for each bias label are given in Table 3.2.1.

This dataset creation process is rule-based and done automatically. with a large part of manual validation of excepts. Much of it was manually validated but might still contain some errors.

3.2.2 Targeted Counter-factual Samples

From the extracted excerpts, we move to create a counterfactual dataset. This dataset can be used as an extension of HUMSET for training or as a test set for measuring the biases encoded in models.

Bias Attribute	Bias label	Excerpt	Excerpt Type
Gender	Female	In Syria, girls are victims of violence.	Counterfactual
	Male	In Syria, boys are victims of violence.	Counterfactual
	Neutral	In Syria, children are victims of violence.	Original
Country	Venezuela	In Venezuela, children are victims of violence.	Counterfactual
	Syria	In Syria, children are victims of violence.	Original
	Canada	In Canada, children are victims of violence.	Counterfactual

Table 3: Example of counterfactual dataset creation. Original sentence: "In Syria, children are victims of violence."

The main idea is to create, from each excerpt and bias attribute new excerpts where the original bias label is replaced by the other ones.

The counterfactual excerpt creation for countries is straightforward as we only replace one country by another. For the Gender attribute, we apply a row-bases mapping from Table 1. Table 3 gives an example of the creation of counterfactual entries. We choose an excerpt containing both bias attributes: "In Syria, children are victims of violence"[1]. For the Country attribute, we replace the original country (Syria) with the two others (Canada and Venezuela) and apply the same for the Gender attribute.

Finally, we must note that we use all bias label combinations for this work, even the transitions from a counterfactual excerpt to another counterfactual one. The total number of excerpts for each bias attribute is, therefore, the same for each bias labels and equals the Sum row in Table 3.2.1. For example, for the gender attribute, the number of excerpts for each of the Female, Male and Neutral labels is the same and equal to 17,048 , 1,592 and 1,669 for the train, validation and test sets respectively.

[1]This excerpt is not present in the dataset and is an example for understanding the counterfactual dataset creation.

4 Methodology

Before explaining our methodology for bias measurement and mitigation, we first describe the classification model architecture used for training, depicted in Fig. 3 (next page). We use the standard architecture for classification tasks [Devlin, M.-W. Chang, et al. 2019]. The text entry is encoded into a set of input ids using a tokenizer fed to the LLM. An MLP is then initialised at the top of the LLM layers and acts as a linear classification head, each head corresponding to a different tag. We use all the tags from the 3 presented tags together and end with 24 output tags. The loss function used for the base training is the Binary Cross Entropy (BCE).

This section is organised into two parts. First, we present the bias measurement techniques used to evaluate the amount of bias in the classification models. Then, in the second part, we give the debiasing methods we experiment with.

4.1 Bias Measurement

We propose and formulate four metrics for evaluating the bias encoded in the classification models, two extrinsic and two intrinsic metrics. We start by defining x as the input excerpt. We define two different bias labels denoted by m and n, and x_m and x_n the excerpts with the respective label. The set $\{\text{m, n}\}$ can be any (non-ordered) combination of the protected labels. These labels according to the HUMSET dataset are defined as follows:

$$\begin{aligned}\Big\{\big\{\text{Female, Male}\big\}, \big\{\text{Female, Neutral}\big\}, \big\{\text{Neutral, Male}\big\}\Big\}\\ \Big\{\big\{\text{Syria, Canada}\big\}, \big\{\text{Syria, Venezuela}\big\}, \big\{\text{Venezuela, Canada}\big\}\Big\}\end{aligned} \tag{1}$$

Additionally, X_m refers to the set of data points for the attribute m, and T to the set of all tags from all different classification tasks, which in our case $|T| = 24$. We use the superscript $\text{m} \rightarrow \text{n}$ to refer to excerpts changed from the bias label m to the bias label n. We refer to C_G as the set of all possible counterfactual transitions for the Gender attribute, and C_C as the set of all possible counterfactual transitions for the Country attribute with $|C_G| = |C_C| = 6$, as shown below:

$$\begin{aligned}\text{C}_G = \Big\{&\text{Female} \rightarrow \text{Male}, \text{Female} \rightarrow \text{Neutral}, \text{Male} \rightarrow \text{Neutral},\\ &\text{Male} \rightarrow \text{Female}, \text{Neutral} \rightarrow \text{Female}, \text{Neutral} \rightarrow \text{Male}\Big\}\\ \text{C}_C = \Big\{&\text{Canada} \rightarrow \text{Syria}, \text{Canada} \rightarrow \text{Venezuela}, \text{Syria} \rightarrow \text{Venezuela},\\ &\text{Syria} \rightarrow \text{Canada}, \text{Venezuela} \rightarrow \text{Canada}, \text{Venezuela} \rightarrow \text{Syria}\Big\}\end{aligned} \tag{2}$$

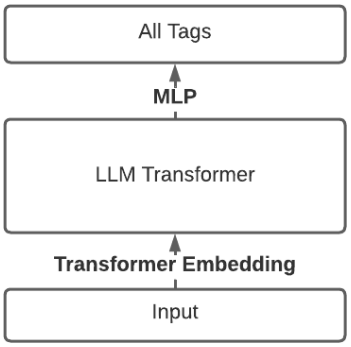


Figure 3: Model Architecture

Finally for brevity, we refer to average with "Avg", to input text excerpt/data with "Xpt", probability with "Prob", Salience with "Sal", and Embedding with "Embed".

We now show the metrics used for bias measurement. Our metrics are four in total, two being extrinsic and two intrinsic.

4.1.1 Extrinsic Metrics

Bias extrinsic metrics refer to metrics based on the models' output on the classification task. We use in our work two extrinsic metrics: the Tag Flip measurement, based on the models' predicted tags and the Probabilities Discrepancy metric, constructed from the models' output probabilities.

4.1.1.1 Tag Flips

The first bias metric we propose is the number of prediction changes when a base excerpt is replaced by a counterfactual one. This metric is very useful to estimate the models' performance on a real-world setup, and to have concrete metrics on the extent of prediction change when a base keyword is replaced by a counterfactual. To this aim, we first define the following binary value to capture whether or not a tag is predicted for the counterfactual excerpt and not the base excerpt.

$$\text{Xpt-Pred-Flips}(x, t, m \to n) = \begin{cases} 1 & \text{if } \begin{cases} \text{Tag not predicted for } x_n \\ \text{Tag predicted for } x_m \end{cases} \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

This metric captures the number of prediction changes per excerpt and the direction of change. It is, therefore, useful to understand which bias labels would be more predicted for each tag compared to others. Moreover, it is essential to note that the Xpt-Pred-Flips metric depends on the direction of source and target attribute, and for a specific x and t, the result of $m \to n$ and $n \to m$ are not

equal.

Using the value defined above, we calculate the Average Tag Prediction Flips metric (Tag-Pred-Flips) by averaging the excerpt tag prediction flip binary over the list of excerpts, defined below.

$$\text{Tag-Pred-Flips}(t, m \to n) = \frac{1}{|X_m|} \sum_{x \in X_m} \text{Xpt-Pred-Flips}(x, t, m \to n) \tag{4}$$

Next, the bias label prediction flips are defined as the sum of the tag prediction flips over the set of tags, formulated as:

$$\text{Bias-label-Pred-Flips}(m \to n) = \sum_{t \in T} \text{Tag-Pred-Flips}(t, m \to n) \tag{5}$$

The average number of prediction flips consists of the number tag changes per excerpt averaged over the set of all possible counterfactual transitions:

$$\begin{aligned} \text{G-Avg-Numb-Pred-Changes} &= \frac{1}{|C_G|} \sum_{m \to n \in C_G} |\text{Bias-label-Pred-Flips}(m \to n)| \\ \text{C-Avg-Numb-Pred-Changes} &= \frac{1}{|C_C|} \sum_{m \to n \in C_C} |\text{Bias-label-Pred-Flips}(m \to n)| \end{aligned} \tag{6}$$

Finally, the model's Average number of prediction flips is the average change for all bias types:

$$\text{T-Avg-Numb-Pred-Changes} = \frac{\text{C-Avg-Numb-Pred-Changes} + \text{G-Avg-Numb-Pred-Changes}}{2} \tag{7}$$

4.1.1.2 Probabilities Discrepancy

The Probability Discrepancy metric captures the shift in probability from the excerpt with bias label m to the one with bias label n. The excerpt probability shift from attribute m to attribute n is shown in Fig. 4 and formulated with:

$$\text{Xpt-Prob-Shift}(x, t, m \to n) = (Prob(x_n, t) - Prob(x_m, t)) \times 100 \tag{8}$$

The 100 multiplication factor's purpose is to support the scores' readability. Then, to analyse the tag-wise probability shift for each pair of attributes $\{$m, n$\}$, we define the Tag Probability Shift between attributes m and n as the median of all excerpt probability shifts from the tag t and attributes

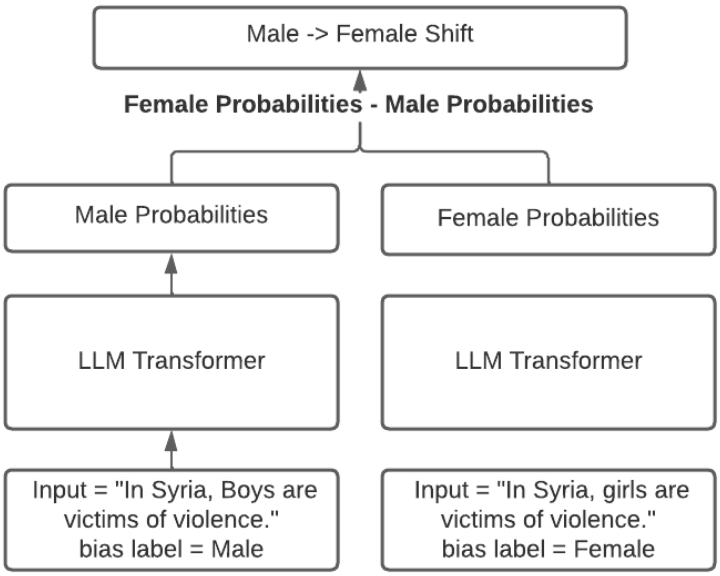


Figure 4: Probabilities discrepancy Computation between two bias labels for one excerpt and one tag. The original Excerpt is: "In Syria, boys are victims of violence." and the Bias labels given for this example are: Male, Female

{m, n}, defined below:

$$\text{Tag-Prob-Shift}(t, m \to n) = \underset{x \in X_m}{Median} \Big\{ \text{Xpt-Prob-Shift}(x, t, m \to n) \Big\} \tag{9}$$

In both Eq. 8 and 9, the values are negated when the direction of m to n is flipped. These metrics can therefore be helpful to determine which attributes the models tend to give higher output probabilities to for different tags. Consequently, to measure the total amount of probability changes from an attribute to another, we sum the corresponding values for tags:

$$\text{Bias-label-Prob-Shift}(m \to n) = \sum_{t \in T} \text{Tag-Prob-Shift}(t, m \to n) \tag{10}$$

Finally, the model Average probability shift consists of the average attribute probability shift over all bias label combinations, shown below.

$$\begin{aligned} \text{G-Avg-Prob-Shift} &= \frac{1}{|C_G|} \sum_{m \to n \in C_G} |\text{Bias-label-Prob-Shift}(m \to n)| \\ \text{C-Avg-Prob-Shift} &= \frac{1}{|C_C|} \sum_{m \to n \in C_C} |\text{Bias-label-Prob-Shift}(m \to n)| \end{aligned} \tag{11}$$

$$\text{T-Avg-Prob-Shift} = \frac{\text{G-Avg-Prob-Shift} + \text{C-Avg-Prob-Shift}}{2} \tag{12}$$

4.1.2 Intrinsic Metrics

Intrinsic metrics, as opposed to extrinsic ones, focus on the internal properties of the models. We research and present two intrinsic metrics: the Salience Discrepancy and the Embeddings Distance.

4.1.2.1 Salience Discrepancy

We present a new bias-measurement method: the Token-Salience Discrepancy Measurement. Similarly to the probability discrepancy method presented earlier, this method consists in comparing two tag-wise metrics, with the difference that we don't use the output probability but the Salience score of the bias-label-specific tokens for each tag.

Choosing a pertinent Explainability method We have a set of requirements for our use case that the Salience computation method needs to satisfy. First, it needs to apply to the LLMs and not be designed only for a different types of model architectures (1). Secondly, the method must handle all the most commonly used activation functions (ReLU, LeakyReLU, Tanh, Sigmoid, Softplus ...) without diverging (2). Thirdly, the method must be proven to yield good interpretability results for other tasks (3). Finally, it needs to be of a cheap computational cost in order to be able to work with limited resources (4).

A literature analysis of the most commonly used explainability methods shows that the following methods cannot be used: Firstly, LIME [Ribeiro, Singh, and Guestrin 2016]; SHAP [Lundberg and S.-I. Lee 2017]; Integrated Gradients [Sundararajan, Taly, and Yan 2017] are computationally expansive [Gholizadeh and Zhou 2021] and do not satisfy the requirement (4). Secondly, the LRP method [Bach et al. 2015] isn't applicable for all types of activation functions [Ancona et al. 2018]. Therefore it doesn't satisfy the requirement (2). Finally, Grad-CAM [Selvaraju et al. 2019] was mainly designed for Convolutional Neural Networks and thus doesn't content the condition (1).

From the methods satisfying the listed conditions, we focus on and analyze the the DeepLift method [Shrikumar, Greenside, and Kundaje 2017]. It is an attribution propagation method which relies on the usage of a baseline input. DeepLift assigns contribution scores as the difference between the activation of each neuron and its reference activation. For implementation, we define the reference token input here as the padding [PAD] token. The final reference input entry will therefore be $\Big[$[CLS] [PAD] [PAD] .... [PAD] [SEP]$\Big]$[1]. The layer for which the attributions are computed is the LLM embedding layer.

[1] as inspired by the `https://captum.ai/tutorials/Bert_SQUAD_Interpret` tutorial

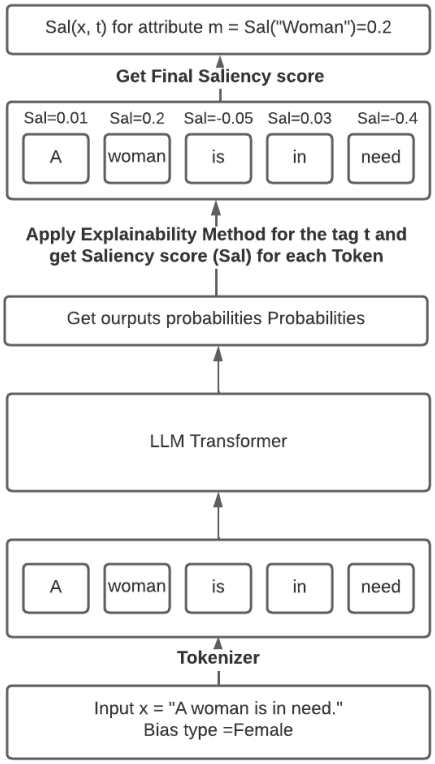


Figure 5: Salience score computation (One excerpt, One tag, One attribute). For simplification, we do not put the special tokens after applying the tokenizer and use an excerpt with only one keyword per bias label. The Salience scores set for each token are examples and not based on any model. Finally, for the cases where more than one keyword exists per bias label, we define the final Salience score for the excerpt as the mean of relevancies of all keywords.

Salience Measurement After choosing the Salience computation algorithms, we propose a method to measure the Tokens Salience score for each excerpt and each bias label. To illustrate our method, we use an example for one excerpt[1] and two specific bias labels, shown in Fig. 5. The methodology used can be generalised for all attributes, excerpts and tags.

Metrics The Token-Relevance discrepancy metrics are similar to those presented in the Probabilities Discrepancy Measurement section (Section 4.1.1.2). We replace the Probability Prob(x, t) with the Salience Sal(x, t), obtained using the mentioned explainability method, and get the following same bias-measurement pipeline and equations:

$$\text{Xpt-Sal-Shift}_{m \to n}(x, t) Sal(x_n, t) - Sal(x_m, t) \tag{13}$$

$$\text{Tag-Sal-Shift}(t, m \to n) = \underset{x \in X_m}{Median} \{\text{Xpt-Sal-Shift}(x, t, m \to n)\} \tag{14}$$

[1]The example excerpt is not present in the dataset, it is only created for a matter of explanation.

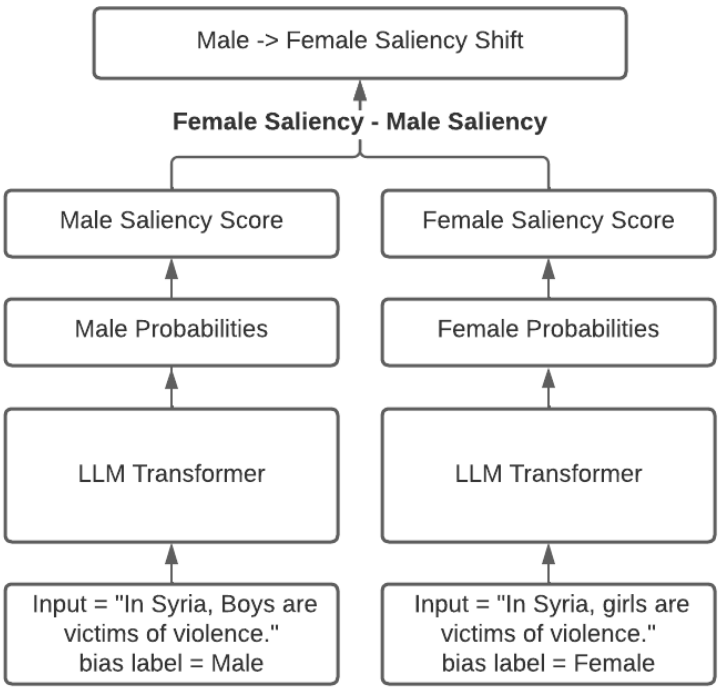


Figure 6: Salience discrepancy Computation between two bias labels for one excerpt and one tag. Original Excerpt: "In Syria, boys are victims of violence." Bias labels: Male, Female.

$$\text{Bias-label-Sal-Shift}(m \to n) = \sum_{t \in T} \text{Tag-Sal-Shift}(t, m \to n) \tag{15}$$

$$\begin{aligned} \text{G-Avg-Sal-Shift} &= \frac{1}{|C_G|} \sum_{m \to n \in C_G} |\text{Bias-label-Sal-Shift}(m \to n)| \\ \text{C-Avg-Sal-Shift} &= \frac{1}{|C_C|} \sum_{m \to n \in C_C} |\text{Bias-label-Sal-Shift}(m \to n)| \end{aligned} \tag{16}$$

$$\text{T-Avg-Sal-Shift} = \frac{\text{G-Avg-Sal-Shift} + \text{C-Avg-Sal-Shift}}{2} \tag{17}$$

4.1.2.2 Embeddings Distance

This metric is based on the Euclidean distance between the embedding outputs of an excerpt with bias labels m and n, as illustrated in Fig. 7. This metric is intrinsic as it doesn't rely on the tags used for training but on the LLM model embedding output. The excerpt embedding distance for attributes m and n is defined as following:

$$\text{Xpt-Embed-Dist}(x, m \to n) = \text{Euclidean-Distance}\left(Embed(x_n), Embed(x_m)\right) \tag{18}$$

We define the embedding distance between two attributes m and n as the average of all excerpt

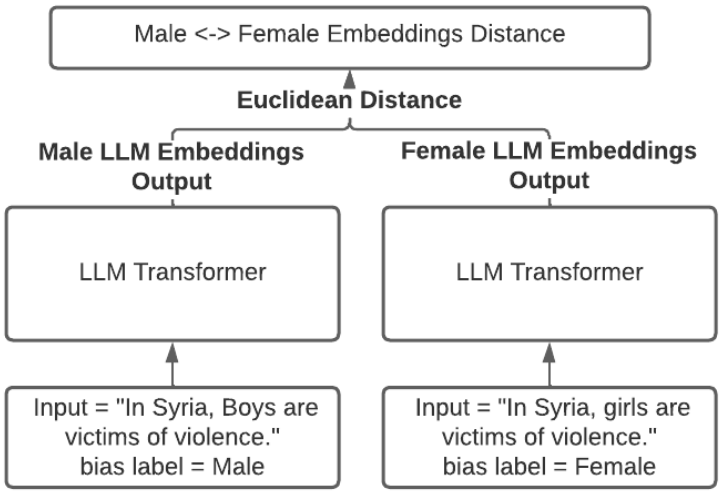


Figure 7: Embeddings Distance Computation between two bias labels for one excerpt. Original Excerpt: "In Syria, boys are victims of violence." Bias labels: Male, Female.

embedding distances between m and n.

$$\text{Bias-label-Embed-Dist}(m \to n) = \underset{x \in X_m}{Med}\Big\{\text{Xpt-Embed-Dist}(x, m \to n)\Big\} \tag{19}$$

Since the Euclidean distance measurement stays the same when the direction of m to n is flipped, the same applies to the excerpt embedding distance and attribute embedding distance. Similarly to previous equations, we get the average embedding distance for one model as the mean of all combinations of attributes embedding distances.

$$\begin{aligned} \text{G-Avg-Embed-Dist} &= \frac{1}{|C_G|} \sum_{m \to n \in C_G} |\text{Bias-label-Embed-Dist}(m \to n)| \\ \text{C-Avg-Embed-Dist} &= \frac{1}{|C_C|} \sum_{m \to n \in C_C} |\text{Bias-label-Embed-Dist}(m \to n)| \end{aligned} \tag{20}$$

$$\text{T-Avg-Embed-Dist} = \frac{\text{G-Avg-Embed-Dist} + \text{C-Avg-Embed-Dist}}{2} \tag{21}$$

4.1.3 Metrics Summary

We summarise all the extrinsic and intrinsic metric equations into the Tables 4 and 5. We ensure here that all metrics are normalised to the number of excerpts so the results of both bias attributes (Gender and Country) can be later compared to another. No normalisation is done for the tags as their number remains the same for both attributes and would only remove readability from results by making them lower.

Name & Variables	Prob-Disc	Pred-Flip
Xpt(x, t, m→n)	$(Prob(x_n, t) - Prob(x_m, t)) \times 100$ ♣	$\begin{cases} 1 & \text{if } \begin{cases} \text{Tag not predicted for } x_n \\ \text{Tag predicted for } x_m \end{cases} \\ 0 & \text{Otherwise} \end{cases}$
Tag(t, m→n)	$\underset{x \in X_m}{Med}\Big\{\text{Xpt(x, t, m→n)}\Big\}$ ♣	$\frac{1}{\lvert X_m \rvert} \sum_{x \in X_m} \text{Xpt(x, t, m→n)}$
B-Lab(m→n)	$\sum_{t \in T} \text{Tag(t, m→n)}$ ♣	$\sum_{t \in T} \text{Tag(t, m→n)}$
B-Attr(i), $i \in \{C, G\}$	$\frac{1}{\lvert C_i \rvert} \sum_{m \to n \in C_i} \lvert \text{B-Lab(m→n)} \rvert$	
B-Tot	$\frac{1}{2} \sum_{i \in \{C,G\}} \text{B-Attr(i)}$	

Table 4: Extrinsic bias metrics. The sign ♣ refers to metrics that are negated when the direction of m to n is flipped. : Metric(m→n) = - Metric(n→m)

4.2 Model Debiasing

We propose two model debiasing techniques, the Focal loss training and the CDA.

4.2.1 Focal Loss Training

We first explore another loss function's impact on the classification performance and bias results: the non-balanced variant of the focal loss. This loss is based on adding a focusing factor to the standard Cross Entropy. Referring to the original formula for one data point and one tag, we get the following equation:

$$p_t = \begin{cases} \text{p} & y = 1 \\ \text{1-p} & \text{otherwise} \end{cases} \tag{22}$$

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

With p the output probability and y the groundtruth value. Knowing that the Cross Entropy CE can be written as CE $= -\log(p_t)$, we see that the main contribution of this new loss compared to the CE loss is the $(1-p_t)^{\gamma}$ factor, which will give more training weight to the hard to classify examples rather than the easy ones.

4.2.2 CDA Training

This debiasing is a pre-processing method, relying mainly on another training dataset. The model architecture and hyperparameters are the same as the BASE one, the only difference being that we

Name & Variables	Sal-Disc	Emb-Dist
Xpt(x, t, m→n)	$Sal(x_n, t) - Sal(x_m, t)$ ♣	Eucl-Dist $(\text{Emb}(x_n), \text{Emb}(x_m))$ ♠
Tag(t, m→n)	$\underset{x \in X_m}{Med}\Big\{$Xpt(x, t, m→n)$\Big\}$ ♣	-
B-Lab(m→n)	$\sum_{t \in T}$ Tag(t, m→n) ♣	$\underset{x \in X_m}{Med}\Big\{$Xpt(x, m→n)$\Big\}$ ♠
B-Attr(i), $i \in \{C, G\}$	$\frac{1}{\lvert C_i \rvert} \sum_{m \to n \in C_i}$ |B-Lab(m→n)|	
B-Tot	$\frac{1}{2} \sum_{i \in \{C,G\}}$ B-Attr(i)	

Table 5: Intrinsic bias metrics. The Embedding Metric is Tag-independent and the Excerpt metric is therefore not tag-dependent. The sign ♣ refers to metrics that are negated when the direction of m to n is flipped. : Metric(m→n) = - Metric(n→m). The sign ♠ refers to metrics that stay the same when the direction of m to n is flipped: Metric(m→n) = Metric(n→m)

apply training not on the base HUMSET, to which we add the counterfactual examples created in Section 3.2 for the Country and Gender attributes. This training setup benefits from the counterfactual samples added to the training data to try and reduce the amount of bias encoded in models.

4.2.3 CDA+FL Combination Training

As the two mentioned debiasing methods work differently (the CDA method on the training data and the FL on the loss function), it is also relevant to analyse how these methods behave when combined. The last debiasing method we propose, therefore consists of a training relying on the adversarial training data from the CDA training setup, with the difference that the loss function used is the Focal Loss (as opposed to the BCE loss function for the CDA setup).

5 Experiment Setup

5.1 Training Setup

For this research work, we use the HumBERT as the base LLM for bias measurement and mitigation. The hyperparameters for model finetuning are reported in Table 6.

Hyperparameters	Values
Number of Epochs	3
Initial Learning Rate	3e-5 if focal loss loss function, 8e-5 if BCE loss function
Dropout Rate	0.2
Train Batch Size	8
Validation Batch Size	16
Gamma Focal Loss	2
Optimizer	Adam Weight (Pytorch implementation[1])
Learning Rate Scheduler	Pytorch StepLR (with decay=0.4, step size=1)
LLM input text max length	256
Freezed LLM layers	LLM Embedding and first LLM layer
Decision boundary threshold	Finetuned separately for each tag on the validation F1 score.

Table 6: Hyperparameters used for finetuning models and Generating final predictions

5.2 Classification Performance on Standard Test Set

To evaluate the standard classification performance, we use two metrics: the Precision[2] and F1[3] score. The following standard calculation metric defines the tag metric

$$\text{Tag-Metric}(t) = \text{Metric}\big(\text{Predictions}(t), \text{Groundtruth}(t)\big) \tag{23}$$

Then, the metric of the ith Task is defined as the average Tag metric for all tags included in the ith Task T_i, with i = 1 to the total number of tasks (in our case, the total number of tasks is three: Sectors, Pillars 1D, Pillars 2D).

$$\text{Task-Metric}_i = \frac{1}{|T_i|} \sum_{t \in T_i} \text{Tag-Metric}(t) \tag{24}$$

Finally, we define the Total Average for each metric as the macro average of the tasks metric.

$$\text{Avg-Metric} = \frac{1}{\#\text{ Tasks}} \sum_{i=1}^{\#\text{ Tasks}} \text{Task-Metric}_i \tag{25}$$

[2] https://en.wikipedia.org/wiki/Precision_and_recall
[3] https://en.wikipedia.org/wiki/F-score

Model	Sectors		Pillars 1D		Pillars 2D		Avg.	
	Prec.	F1	Prec.	F1	Prec.	F1	Prec.	F1
BASE	0.76	**0.78**	**0.69**	0.69	0.69	0.68	0.71	0.72
FL	**0.77**	**0.78**	**0.69**	**0.70**	0.69	**0.71**	**0.72**	**0.73**
CDA	0.76	0.76	0.68	0.69	0.66	0.68	0.70	0.71
CDA+FL	**0.77**	**0.78**	**0.69**	**0.70**	**0.70**	**0.71**	**0.72**	**0.73**

Table 7: Classification Metrics ↑

6 Results and Discussion

We start by reporting the classification results, then we move to the bias measurement outputs.

6.1 Classification Performance on Standard Test Set

Table 7 shows the classification results for models. We see that all training setups yield close classification performances. The FL training slightly improves the baseline results (0.01 amelioration for the Precision and F1 score) while the CDA slightly worsens them (0.01 deterioration for both metrics). Finally, we see that the combination of both debiasing techniques yields results equivalent to those obtained with the FL method, and therefore an improvement in classification results compared to the baseline model.

6.2 Biases Analysis

After evaluating the classification results for all experiment setups, we measure the bias contained in models and compare all bias measurement methods. We start by measuring the bias encoded into models on system level, then we move to compare the different bias measurement methods on tag level.

6.2.1 System-level Results

We start by analysing the bias in the trained models for the four proposed metrics. In Table 8, we report the bias results for all bias-labels combinations as well as the average results.

Debiasing Methods comparison We see that the CDA training setup yields a significant improvement in bias results for all bias metrics. For example, the total amount of bias contained in

[3] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html#torch.optim.AdamW

Metric	Model	Gender-Avg	Country-Avg	Total-Avg
Tag Flips ↓	BASE	0.11	0.48	0.29
	FL	0.12	0.48	0.32
	CDA	**0.04**	**0.08**	**0.06**
	CDA+FL	**0.04**	0.13	0.09
Probablities Discrepancy ↓	BASE	0.48	12.12	6.30
	FL	5.71	42.85	24.28
	CDA	**0.01**	**0.24**	**0.13**
	CDA+FL	0.55	2.80	1.60
Salience Discrepancy ↓	BASE	0.54	3.28	1.91
	FL	0.84	4.18	2.51
	CDA	**0.10**	**1.06**	**0.58**
	CDA+FL	0.17	1.45	0.81
Embeddings Distance ↓	NO-FT	**0.30**	**0.58**	**0.44**
	BASE	1.15	6.38	3.77
	FL	1.22	4.05	2.63
	CDA	0.40	1.29	0.84
	CDA+FL	0.45	1.72	1.08

Table 8: All Bias Metrics. We use the following abbreviations.

models with the CDA training compared to the BASE setup is divided by ∼5 for the Tag Flips metric (0.29 to 0.06) and by ∼5 for the Salience Discrepancy (1.91 to 0.58).

On the other side, the FL training method yields mixed results. We notice a slight improvement in the Embedding Distance metrics (3.77 to 2.63 compared to the BASE setup). However, the other metrics are worsened, sometimes significantly as for the Probability Discrepancy metric (6.30 for the BASE to 24.80 for the FL).

Finally, the combination of the debiasing methods yields low bias results, even if they remain higher than those of the CDA method for all metrics. For example, the Embedding distance goes from 2.63 to 1.08 when going from the FL setup to the CDA+FL one, and the same can be said about the three other metrics.

HUMSET and HumBERT bias contributions The non-null Embeddings Distance for the NO-FT setup shows that the HumBERT LLM also contains country and nationality biases. They however remain low compared to the amount of bias contained in the HUMSET dataset (0.44 Total Average Embeddings distance for the NO-FT setup vs 3.77 for the BASE training setup). For both the HUMSET and the HumBERT, we also notice that the amount of bias for the Country attribute is higher than for the Gender one.

Bias-attributes comparison We notice that the amount of bias contained for the Country attribute is higher than the Gender one. This statement is verified for all the proposed metrics and training architectures. This difference in the bias amounts is present both in the initial NO-FT backbone model and the training data. In fact, we find a value of 0.58 for the Country attribute vs 0.30 for the Gender attribute for the Embeddings Distance for the NO-FT setup, giving a ratio of $\sim$2. This ratio is accentuated for the other setups (for example to a ratio of $\sim$3 for the CDA training setup).

6.2.2 Tag-wise Results

In order to have a good understanding of the specific behaviour of each metric in a production-like setup, we perform a tag-wise bias analysis. The Embedding distance metric being non-tag-based, we focus on the three other bias metrics we proposed. We center our comparison on one training setup, the CDA+FL Training setup as it is the setup the most likely to be used in a production-wise setup for Gender and Countries by using Fig. 8 and 9 respectively. We start by comparing the results obtained by different bias-measurement techniques then move to discerning patterns and biases to be aware of when using the models. A similar metrics comparison methodology can be applied to other bias attributes and training setups, using the visualisations in the Appendix.

We start by comparing the Salience discrepancy results to the extrinsic metrics. Looking at the bias results for countries (Fig. 9), we see that there is no clear pattern we can extricate. For example, when replacing the Canada country by the Syria one, we see that the Salience discrepancy is positively correlated to the probabilities discrepancy for the "displacement" tag (0.18 and 0.19 respectively), negatively correlated for the "Covid-19" tag (-0.14 and 0.75 respectively) and very lowly correlated for the "Humanitarian Accesss" tag (0.11 and -0.024 respectively).

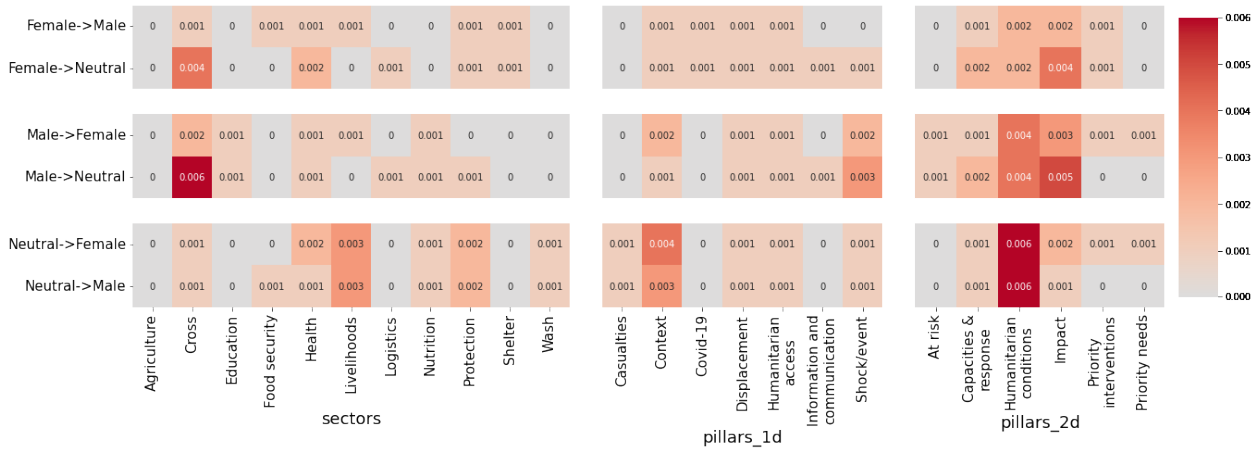
Then, we move analyzing and comparing the extrinsic results, the Average tag flips count nd the probabilities discrepancy. We notice a general pattern, which shows that the higher the probability median shift between two bias labels for one tag, the higher the amount of average tag flips. This can be seen especially for the "Cross" and "Impact" tags when switching the Male bias label by the Neutral on Fig. 8. It is also important to note that for some cases, even though the median probability discrepancy is shifted from one bias label to another, there can be some excerpts where there is a tag shift on the other direction. For example, on Fig. 8, for the "Protection" tag, even though the median probability for Female is higher than the one of Male (as seen in Fig. **??** with a discrepancy of 0.047), Fig. 9a also shows that there exist cases where the average tag-flip count is not null (0.001). On the other side, having a positive median shift doesn't always cause prediction

flips. In fact, even though the median shift from Male to Neutral for the "Covid-19" tag is positive (0.039), we see no prediction flips.

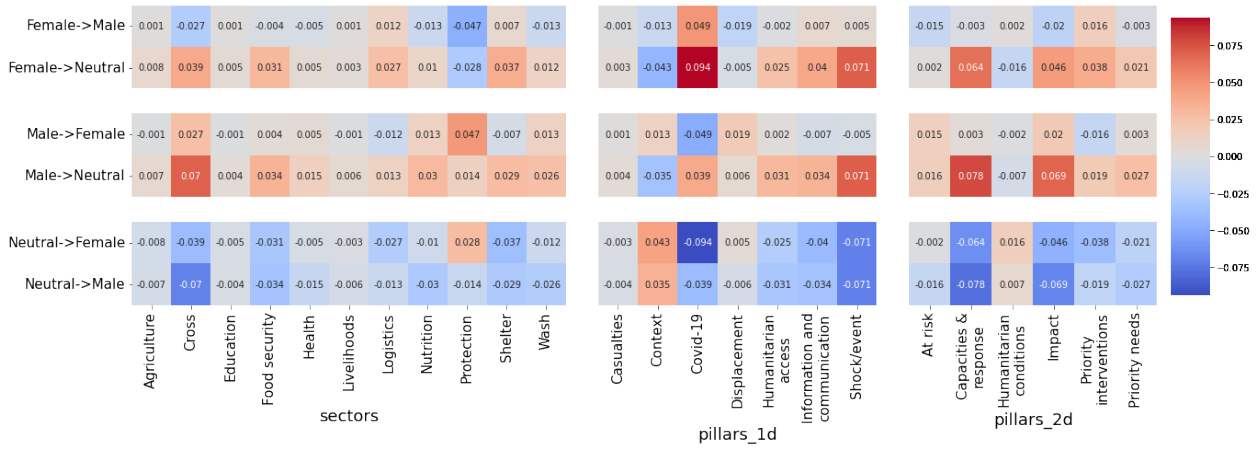
Next, we perform a thorough analysis of the Tag Flips results, as it is the most important one to understand the model's real-world behaviour and the biases potentially present when receiving predictions. As noted in Section 6.2.1, the amount of bias contained for the Country attribute is higher than the Gender one. As a matter of fact, the maximum value for the country attribute is of 0.02, $\sim$3 times more than the Gender one (0.006). These values however are low and show that the debising pipeline worked well on all tags and for all bias attributes. A look at Fig. 11a also shows a significant decrease in bias outputs for the most biased tags (the maximum value of bias for Countries decreased by 4 times, from 0.08 to 0.02).

Finally, we use Fig. 11a to understand the biases contained in the BASE model. On this figure, we see that Venezuela is highly biased towards "Covid-19" tags (values of 0.08 and 0.085 when used as a counterfactual of Canada and Syria respectively). This shows that the training dataset is biased, causing the BASE classification models to contain these biases. We also see that Syria and Venezuela are both negatively biased towards the "Impact" tag, as the neutral keyword, Canada, contains more shifts than these two countries.

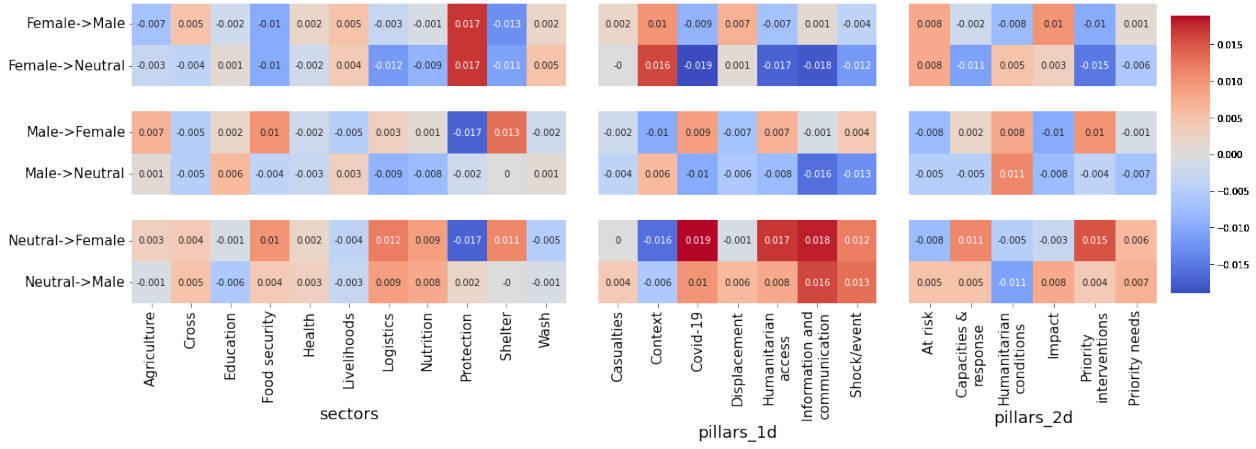
In conclusion, we see that even though the macro-averages of biases generally behave the same (as seen in Section 6.2.1), the tag-wise plots give results that contain some similarities but also many differences. While the extrinsic metrics can show real interpretability on tag-wise results, the Salience metric works well for the system-level result (with an overall amount of bias decreasing for bias-label-specific tokens) but isn't helpful for analyzing tag-wise biases.



(a) Average Tags Flips count

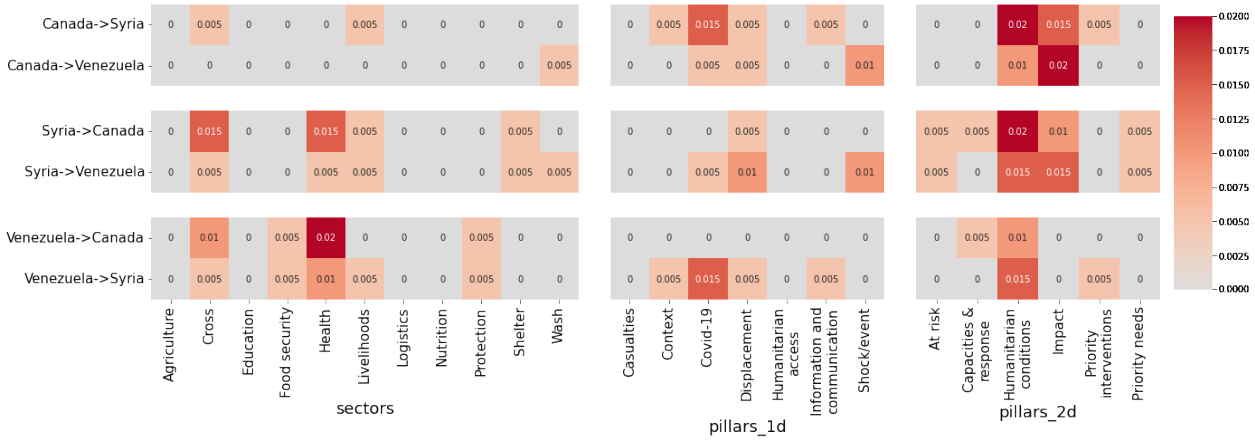


(b) Probabilities Discrepancy

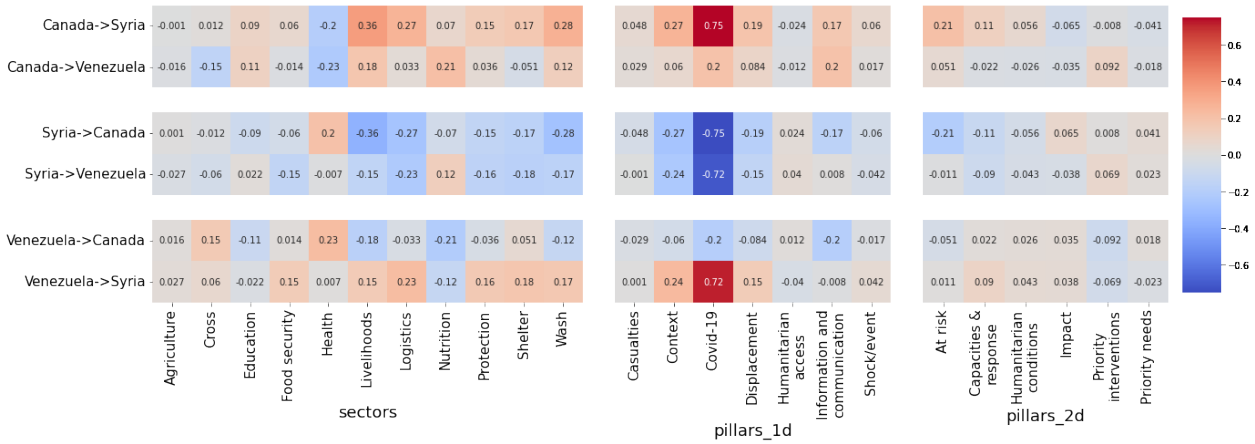


(c) Salience Discrepancy

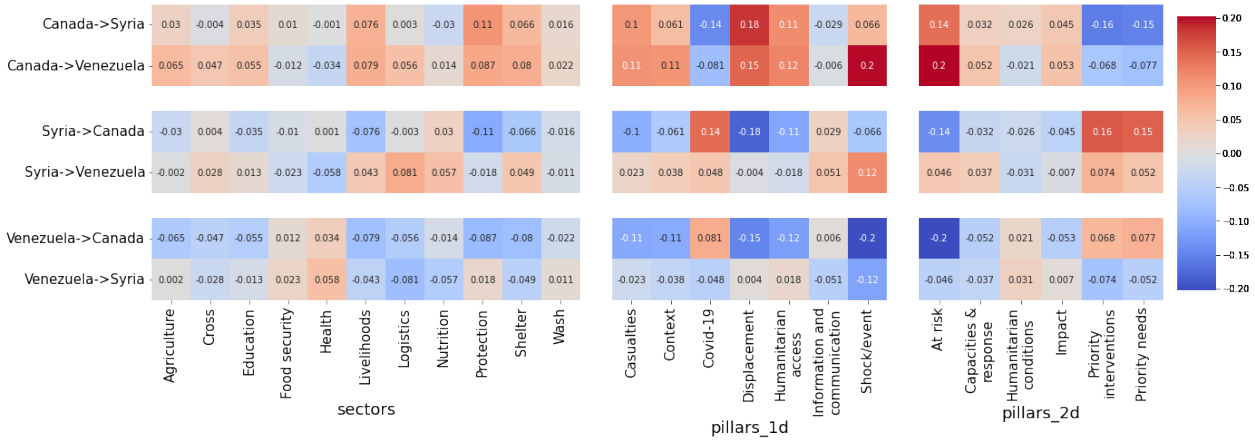
Figure 8: CDA+FL Training Setup Gender Tag-wise results



(a) Average Tags Flips count



(b) Probabilities Discrepancy



(c) Salience Discrepancy

Figure 9: CDA+FL Training Setup Country Tag-wise results

7 Conclusion

As AI and NLP are having more and more impact on day-to-day applications, ensuring an ethical utilisation and unbiased outputs is particularly pertinent when applied to Humanitarian crisis response with people directly affected. For this bias measurement and mitigation work, we create a new dataset, the HUMSETBIAS, from the released HUMSET. This new dataset contains two bias attributes, Gender and Country, with three bias labels each. Then, we propose a pipeline to measure the bias encoded in the classification models by four different methods. These metrics behave similarly on system level but reveal different results when used to find the tag and attribute-wise specific biases. Results show that the Country bias attributes contain more bias than the Gender. Finally, we present and compare different debiasing methods, effectively reducing the bias encoded in the classification models. On the one hand, the CDA training setup improves all the bias metrics but slightly worsens the classification results. On the other hand, the FL improves the classification results but is less effective as a debiasing setup, with an improvement in the intrinsic metrics but a deterioration for the extrinsic ones. Finally, combining both debiasing setups maintains the high classification performance attained with the FL training while maintaining a low bias amount in models.

8 Limitations and Ethical Considerations

Even if this work provides several contributions, it is important to note different limitations. Firstly, HUMSETBIAS is restrained only to Gender biases and two countries (Syria and Venezuela). As many countries as possible should complete this dataset and all groups affected by humanitarian crises (displaced people, migrants, refugees, minorities...). This dataset is also rule-based, with manual sanity checks. Still, it probably contains some errors, which could be grammatical because of some keywords switching or context linking (for example, some keywords related to specific countries could not be detected). Then, other bias measurement and mitigation techniques could have been explored. Some of the most known bias measurement metrics include data leakage computation with probing [Mendelson and Belinkov 2021]. Moreover, we restricted this work to only one backbone for computational resource optimisation: HumBERT. Future work can include running the proposed experiments or other language models and verifying whether the results obtained with HumBERT are reproduced. Finally, while the debiasing techniques helped reduce bias, the classification models still contain multiple flaws from both the classification and the bias standpoints. They should therefore be used mainly to assist analysts and subjected to quality control rather than fully automatic.

Acknowledgements

Academically I want to start by thanking all my professors and project supervisors during my years here at EPFL and in my exchange year in Paris. I would also like to thank my master section supervisors, Simone Deparis and Orane Pouchon. They always were understanding and caring and let me gain the specific knowledge I need for my professional work. Finally, I would like to thank my thesis supervisor, Robert West, for accepting my thesis proposal and giving me time and energy during these months.

Professionally I would like to address a big thanks to Data Friendly Space DFS[1], as this organisation contributed a lot to my professional, technical, and personal growth. I started at DFS two years ago as an intern with very basic coding skills and knowledge in NLP. I worked with beautiful humans, from my coworkers in Italy and Nepal to the organisation leaders. Moreover, I worked on the DEEP Platform with people from a lot of different backgrounds and skills and am happy to have lived this experience. I'm genuinely thankful to have been in this position and in this work environment, which gave me everything I needed to pursue my master studies and my thesis, from the time and flexibility to the computation power. Finally, I would like to address a special thanks to Navid Rekabsaz for all the supervision over the past 2 years and during this master thesis. He has been an awesome technical supervisor and an amazing person. He always pushed me to improve, valuing my work and giving me the confidence to continue.

Personally On a personal level, I would like to start by thanking my family, my parents, and my brother for always standing up by my side during these six years in university. They were very understanding and patient with me even when I wasn't acting in the best possible way. My friends here at EPFL also had a huge impact on my master's success. They carried me, were patient with me, and gave me a lot of help to pass the courses and the projects.

[1] https://datafriendlyspace.org/

References

Adel, Ghadah and Yuping Wang (2020). "Detecting and Classifying Humanitarian Crisis in Arabic Tweets". In: *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. DOI: 10.1109/ICAIBD49809.2020.9137480.

Alam, Firoj et al. (2021). "CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing". In: *Proceedings of the International AAAI Conference on Web and Social Media*.

Alharbi, Alaa and Mark Lee (Apr. 2021). "Kawarith: an Arabic Twitter Corpus for Crisis Events". In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, pp. 42–52. URL: https://aclanthology.org/2021.wanlp-1.5.

Ancona, Marco et al. (2018). *Towards better understanding of gradient-based attribution methods for Deep Neural Networks*. arXiv: 1711.06104 [cs.LG].

Bach, Sebastian et al. (July 2015). "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLOS ONE* 10.7, pp. 1–46. DOI: 10.1371/journal.pone.0130140. URL: https://doi.org/10.1371/journal.pone.0130140.

Devlin, Jacob, Ming-Wei Chang, et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

— (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Elazar, Yanai and Yoav Goldberg (Oct. 2018). "Adversarial Removal of Demographic Attributes from Text Data". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 11–21. DOI: 10.18653/v1/D18-1002. URL: https://aclanthology.org/D18-1002.

Fekih, Selim et al. (2022). *HumSet: Dataset of Multilingual Information Extraction and Classification for Humanitarian Crisis Response*. arXiv: 2210.04573 [cs.CL].

Gholizadeh, Shafie and Nengfeng Zhou (2021). "Model Explainability in Deep Learning Based Natural Language Processing". In: *CoRR* abs/2106.07410. arXiv: 2106.07410. URL: https://arxiv.org/abs/2106.07410.

Iberdola (n.d.). *What are the largest humanitarian crises in the planet today?* URL: https://www.iberdrola.com/social-commitment/humanitarian-crises-causes-effects-solutions.

Imran, Muhammad, Prasenjit Mitra, and Carlos Castillo (2016). "Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages". In: *arXiv preprint arXiv:1605.05894*.

Lai, Kelvin et al. (2022). "A natural language processing approach to understanding context in the extraction and geocoding of historical floods, storms, and adaptation measures". In: *Information Processing & Management* 59.1, p. 102735.

Lin, Tsung-Yi et al. (2018). *Focal Loss for Dense Object Detection*. arXiv: 1708.02002 [cs.CV].

Liu, Haochen et al. (2021). *The Authors Matter: Understanding and Mitigating Implicit Bias in Deep Text Classification*. arXiv: 2105.02778 [cs.CL].

Lu, Kaiji et al. (2018). "Gender bias in neural natural language processing". In: *arXiv preprint arXiv:1807.11714*.

Lundberg, Scott and Su-In Lee (2017). *A Unified Approach to Interpreting Model Predictions*. arXiv: 1705.07874 [cs.AI].

Medium (n.d.[a]). *Natural Language Processing in the Social Media Age*. URL: https://medium.com/geekculture/natural-language-processing-in-the-social-media-age-23e1e10235bd.

— (n.d.[b]). *NLP use cases in the insurance industry*. URL: https://medium.com/ubiai-nlp/nlp-use-cases-in-the-insurance-industry-7eab6de4baae.

Mendelson, Michael and Yonatan Belinkov (Nov. 2021). "Debiasing Methods in Natural Language Understanding Make Bias More Accessible". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1545–1557. DOI: 10.18653/v1/2021.emnlp-main.116. URL: https://aclanthology.org/2021.emnlp-main.116.

Nielsen, Ian E. et al. (July 2022). "Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks". In: *IEEE Signal Processing Magazine* 39.4, pp. 73–84.

DOI: 10.1109/msp.2022.3142719. URL: https://doi.org/10.1109%5C%2Fmsp.2022.3142719.

Rajič, Frano et al. (2022). *Using Focal Loss to Fight Shallow Heuristics: An Empirical Analysis of Modulated Cross-Entropy in Natural Language Inference*. arXiv: 2211.13331 [cs.CL].

Rekabsaz, Navid, Simone Kopeinik, and Markus Schedl (2021). "Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 306–316.

Rekabsaz, Navid and Markus Schedl (2020). "Do Neural Ranking Models Intensify Gender Bias?" In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2065–2068.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). *Model-Agnostic Interpretability of Machine Learning*. arXiv: 1606.05386 [stat.ML].

Rocca, Roberta et al. (2023). "Natural language processing for humanitarian action: Opportunities, challenges, and the path toward humanitarian NLP". In: *Frontiers in Big Data* 6. ISSN: 2624-909X. DOI: 10.3389/fdata.2023.1082787. URL: https://www.frontiersin.org/articles/10.3389/fdata.2023.1082787.

Selvaraju, Ramprasaath R. et al. (Oct. 2019). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2, pp. 336–359. DOI: 10.1007/s11263-019-01228-7. URL: https://doi.org/10.1007%2Fs11263-019-01228-7.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). "Learning Important Features Through Propagating Activation Differences". In: *CoRR* abs/1704.02685. arXiv: 1704.02685. URL: http://arxiv.org/abs/1704.02685.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic Attribution for Deep Networks". In: *CoRR* abs/1703.01365. arXiv: 1703.01365. URL: http://arxiv.org/abs/1703.01365.

Tamagnone, Nicolò et al. (2023). *Leveraging Domain Knowledge for Inclusive and Bias-aware Humanitarian Response Entry Classification*. arXiv: 2305.16756 [cs.CL].

Usuga-Cadavid, Juan Pablo et al. (2021). "Exploring the Influence of Focal Loss on Transformer Models for Imbalanced Maintenance Data in Industry 4.0". In: *IFAC-PapersOnLine* 54.1. 17th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2021, pp. 1023–1028. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2021.

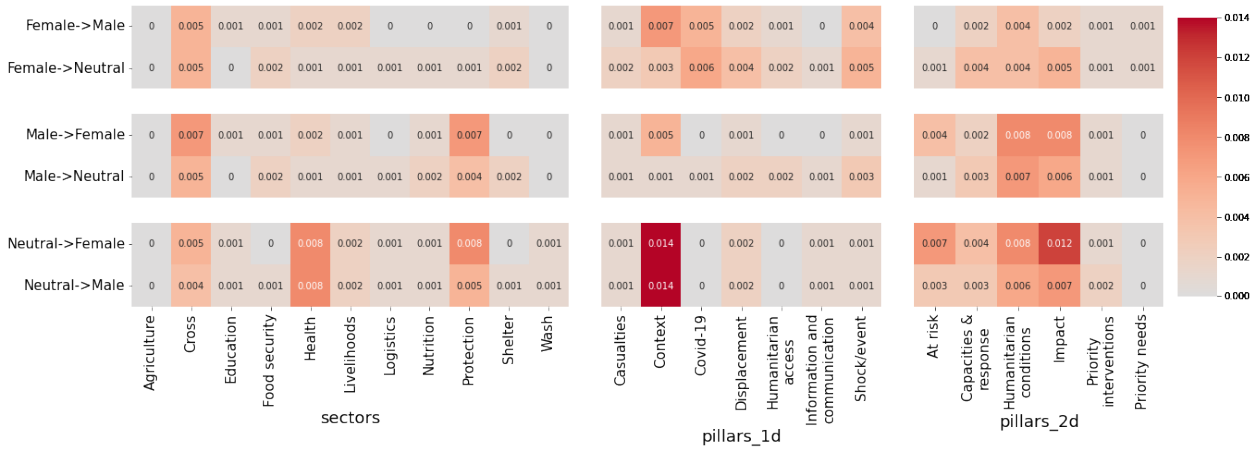
08.121. URL: https://www.sciencedirect.com/science/article/pii/S2405896321008776.

Yela-Bello, Jenny Paola, Ewan Oglethorpe, and Navid Rekabsaz (Apr. 2021). "MultiHumES: Multilingual Humanitarian Dataset for Extractive Summarization". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1713–1717. DOI: 10.18653/v1/2021.eacl-main.146. URL: https://aclanthology.org/2021.eacl-main.146.

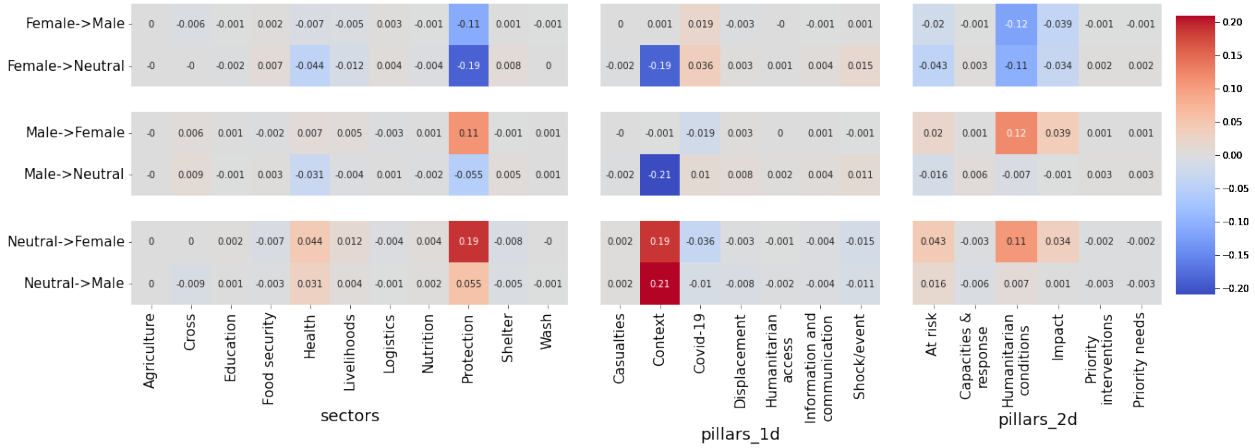
Zhang, Zhilu and Mert R. Sabuncu (2018). *Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels*. arXiv: 1805.07836 [cs.LG].

Zhao, Jieyu et al. (2018). "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods". In: *Proceedings of North American Chapter of the Association for Computational Linguistics*.

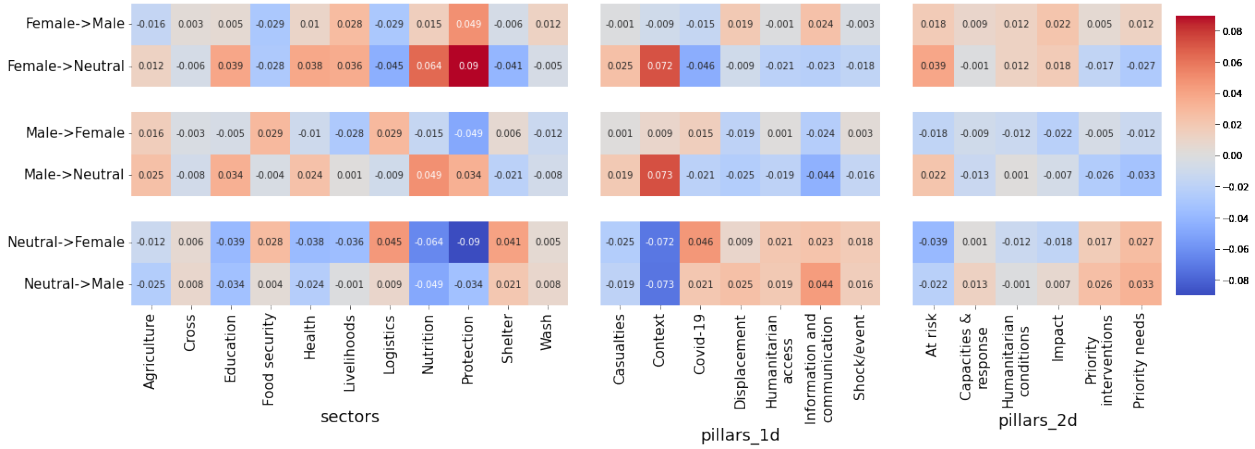
Appendix



(a) Average Tags Flips count



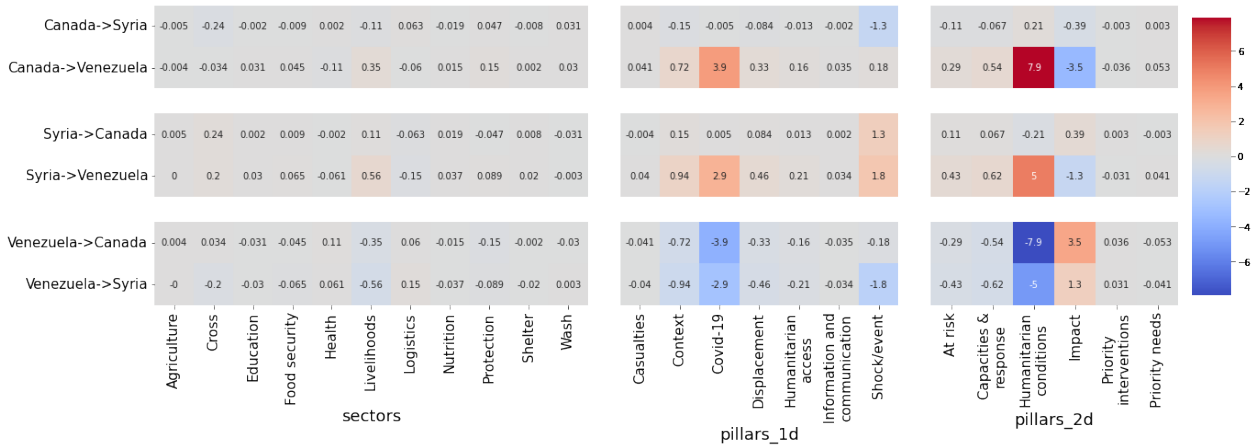
(b) Probabilities Discrepancy



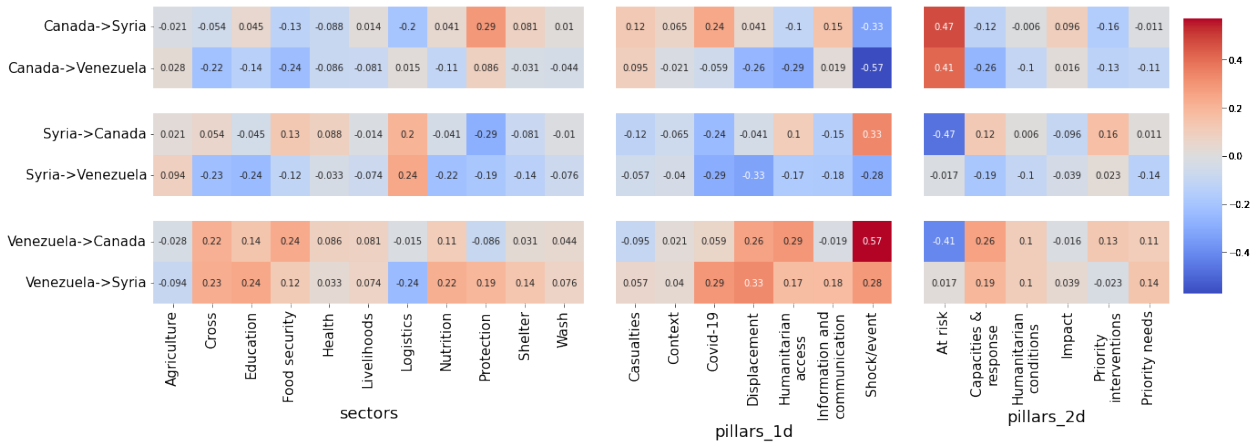
(c) Salience Discrepancy

Figure 10: BASE Training Setup Gender Tag-wise results

(a) Average Tags Flips count

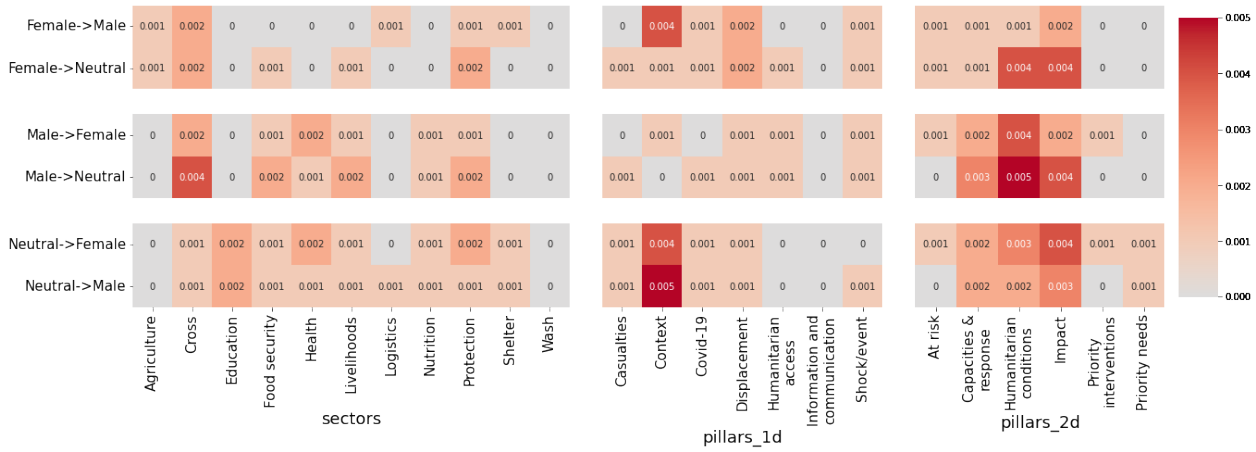


(b) Probabilities Discrepancy

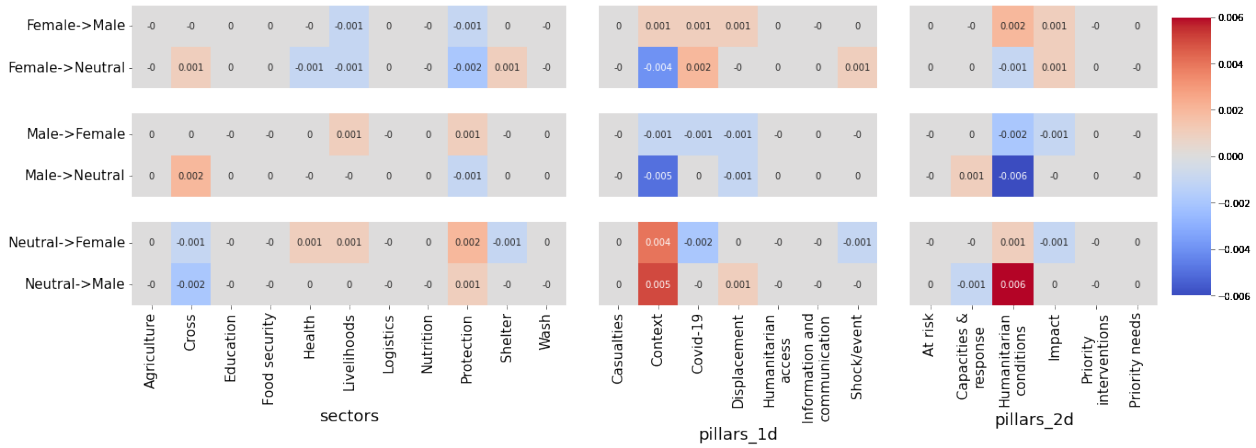


(c) Salience Discrepancy

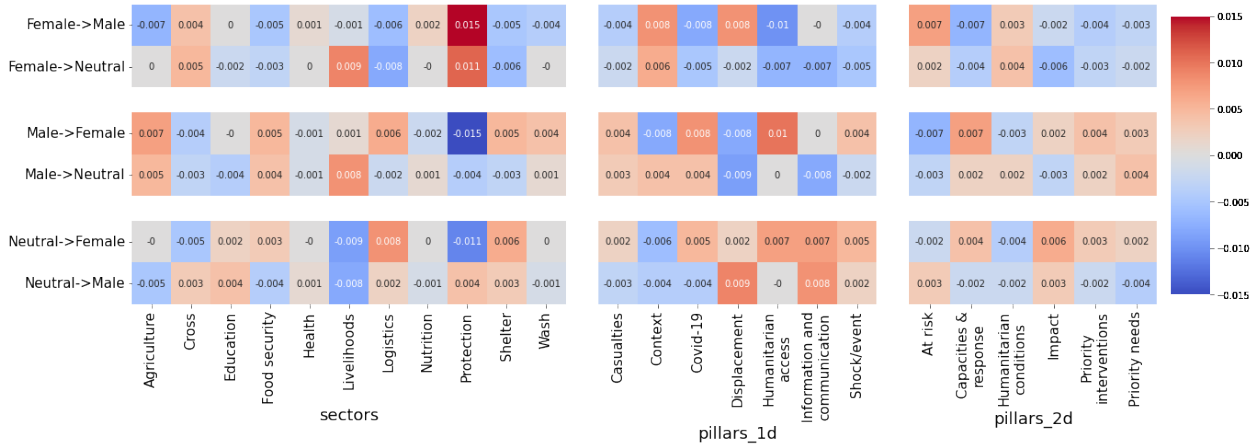
Figure 11: BASE Training Setup Country Tag-wise results



(a) Average Tags Flips count

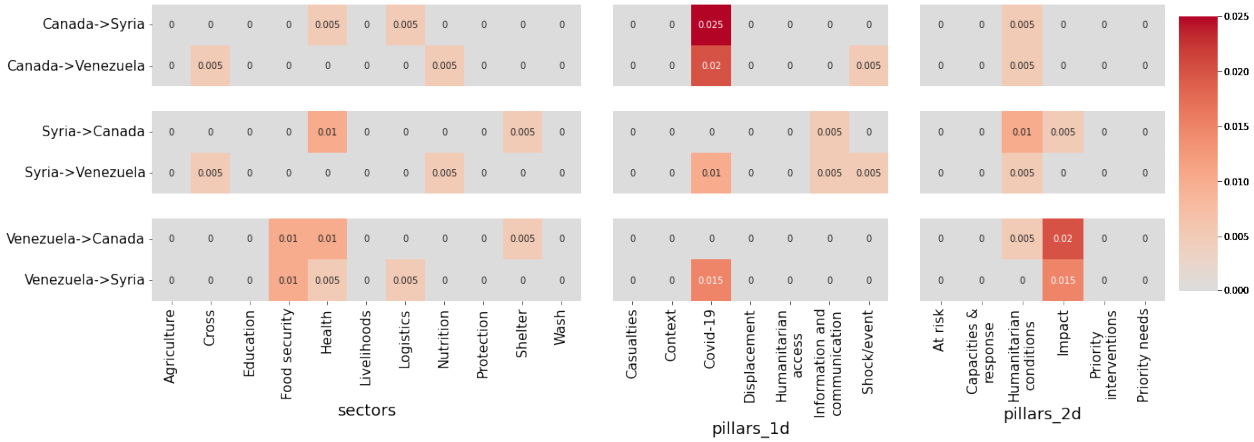


(b) Probabilities Discrepancy

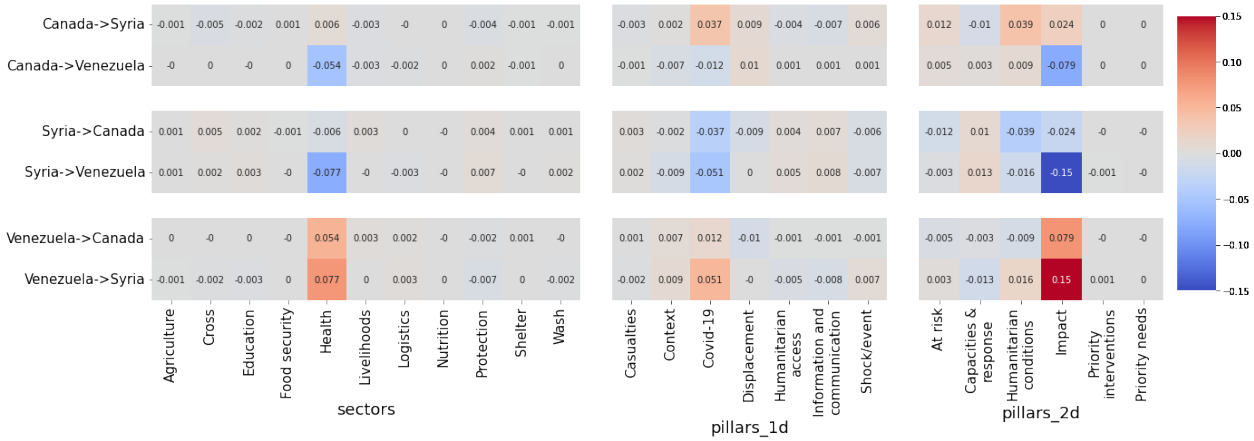


(c) Salience Discrepancy

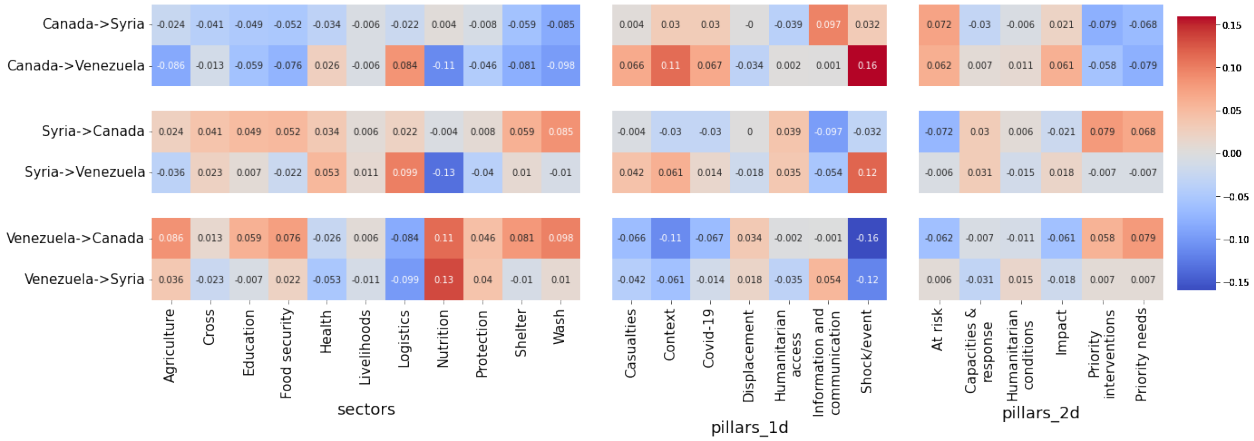
Figure 12: CDA Training Setup Gender Tag-wise results



(a) Average Tags Flips count

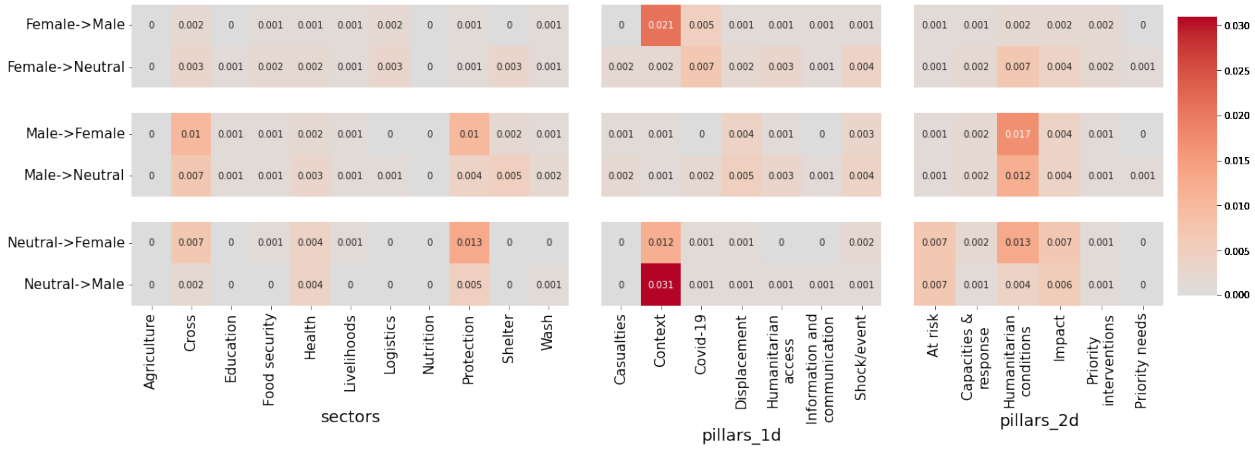


(b) Probabilities Discrepancy

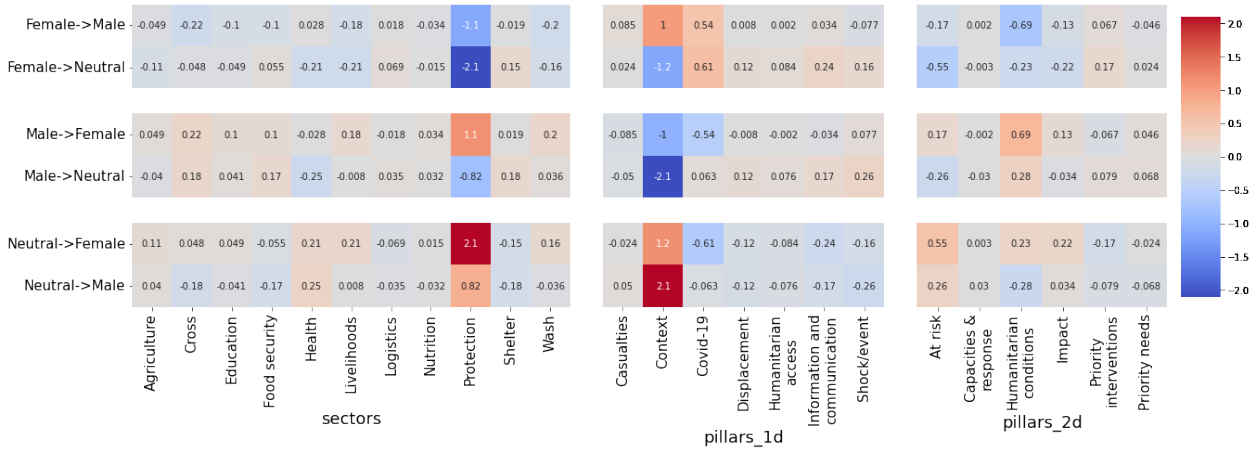


(c) Salience Discrepancy

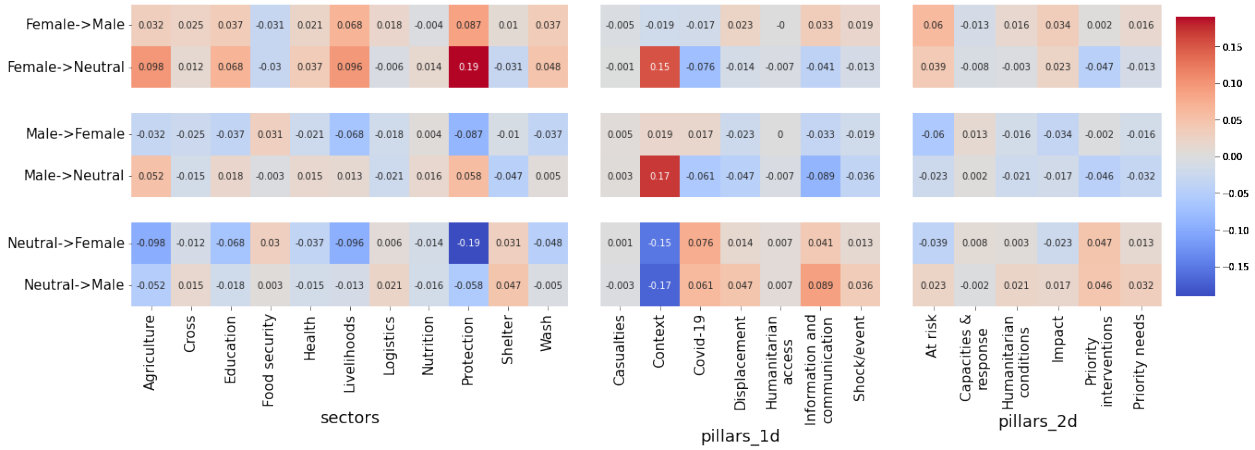
Figure 13: CDA Training Setup Country Tag-wise results



(a) Average Tags Flips count

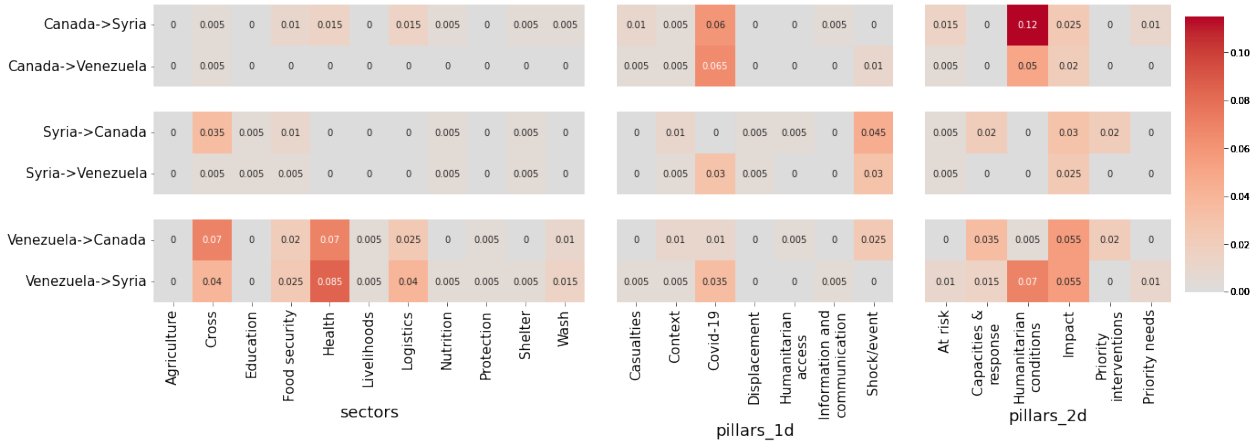


(b) Probabilities Discrepancy

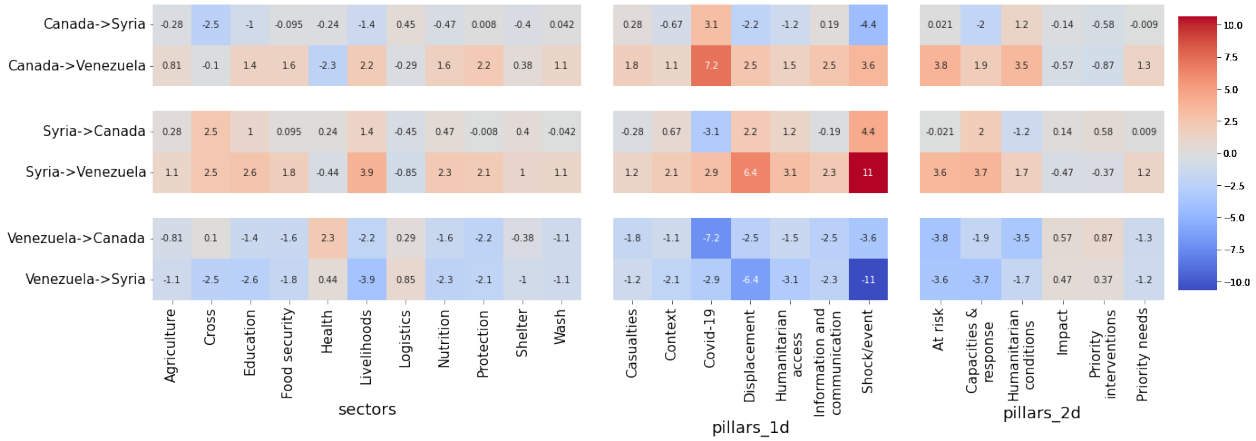


(c) Salience Discrepancy

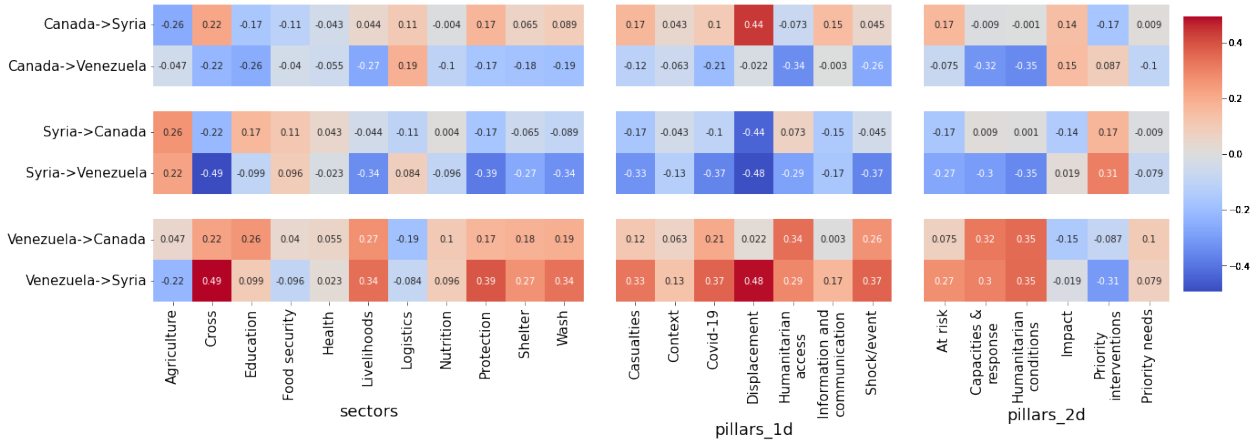
Figure 14: FL Training Setup Gender Tag-wise results



(a) Average Tags Flips count



(b) Probabilities Discrepancy



(c) Salience Discrepancy

Figure 15: FL Training Setup Country Tag-wise results