

Without Demographics, Without Labels: A Diagnostic Framework for Fairness in Embedders

CentraleSupélec Internship Report

Erwan DAVID*
ed.erwandardavid@gmail.com

December 18, 2025



This report details the work conducted during my internship at the International Laboratory on Learning Systems (ILLS), under the co-supervision of Ulrich Aivodji and Pablo Piantanida. It is submitted to CentraleSupélec as my final internship report.

Abstract

Foundation embedders have become central to modern AI systems, powering a wide variety of tasks across domains such as natural language processing, vision, and multimodal reasoning. Their role is to encode raw inputs into vector representations that can be reused in many downstream applications. However, evaluating their fairness remains an open challenge. Unlike task-specific models, embedders are trained to be general-purpose and are rarely associated with labeled data, which makes classical fairness metrics inapplicable. Furthermore, retraining or fine-tuning these large models to enforce fairness constraints is often computationally prohibitive and practically infeasible.

In this work, I study fairness evaluation for embedders with a fully label-free representation-level diagnostic validated against supervised fairness criteria. Building on the Information Sufficiency (IS) framework, I first analyze a natural group-conditional extension (FairIS) and show that, although theoretically motivated, it does not provide a discriminative proxy for downstream group-wise fairness. Motivated by this limitation, I propose *FairEntropy*, a simple and efficient representation-level diagnostic based on worst-group conditional entropy, designed to probe structural worst-case imbalances in how embeddings represent different subpopulations.

Empirical results on multiple ACS benchmarks show that FairEntropy exhibits strong monotonic correlations with Worst-Group Accuracy (WGA) when demographic attributes are available (up to average $\rho_p \approx 0.83$), particularly on tasks where group disparities are primarily driven by representational effects. Building on this foundation, I extend FairEntropy to an unsupervised setting, which is especially relevant in realistic scenarios where demographic attributes cannot be accessed due to privacy constraints or are unavailable. In this case, proxy latent groups are induced via clustering, and FairEntropy continues to exhibit strong monotonic correlations with Worst-Group Accuracy (up to average $\rho_p \approx 0.84$).

These results suggest that FairEntropy is a practical and easy- and fast-to-compute diagnostic for fairness-aware model selection under severe informational constraints.

Throughout the paper, we use the terms *supervised* and *unsupervised* interchangeably with *with demographics* and *without demographics*, respectively.

*CentraleSupélec, ILLS (Mila - ETS - McGill - CentraleSupélec - CNRS - Université Paris-Saclay)

Contents

1	Introduction	3
1.1	Context	3
1.2	Development of the Research Topic	3
1.3	Reformulation of the Problem	4
1.4	Overview of the Document	4
2	Literature Review	4
2.1	Fairness with Demographics	4
2.2	Fairness without Demographics	6
2.3	Information Sufficiency Score	7
3	Methodology	9
3.1	From FairIS to FairEntropy: Creation of the Score	9
3.1.1	Base score with demographics	9
3.1.2	Extension without demographics.	11
3.2	Evaluation Strategy	12
3.3	Datasets	12
3.4	Embedding models	13
4	Experiments and Results	14
4.1	Fairness with demographics	14
4.2	Fairness without demographics	15
4.3	Applicability and limitations	16
5	Future Work	16
6	Technical Aspects	17
7	Main Learnings from the Internship	17
8	Ethics	18
9	Acknowledgements	18
10	Conclusion	18
11	Appendix	18
11.1	Worst-Group Bayes Risk and Groupwise Deficiency: A Simple Bound	18
11.2	Embedding models	19
11.3	Impact of the value of K – Unsupervised settings	19
11.4	Final ranking of embedders based on the mean worst FairEntropy score across the 3 datasets.	20
11.5	FairIS: Group-Conditional Information Sufficiency	21

1 Introduction

1.1 Context

This internship takes place within the International Laboratory on Learning Systems (ILLS), a joint research initiative launched in 2022 by McGill University, École de technologie supérieure (ÉTS), Mila – Quebec AI Institute, the Centre National de la Recherche Scientifique (CNRS), Université Paris-Saclay, and CentraleSupélec. The laboratory’s mission is to advance both the theoretical foundations and practical applications of artificial intelligence, with ongoing projects spanning natural language processing, computer vision, robust systems, and trustworthy machine learning.

Embedders occupy a central place in this landscape. By mapping raw data such as text, images, or speech into structured vector representations, they provide the backbone of most modern AI systems. Their widespread adoption is particularly visible in large language models (LLMs), where pre-trained embedders are reused across a variety of downstream tasks ranging from classification to retrieval and reasoning. This universality, while powerful, also creates new challenges for evaluation: unlike task-specific models, foundation embedders are not directly tied to a single dataset or objective, making evaluation more complex.

Fairness is a key concern because these models are often deployed in sensitive domains. Yet, assessing fairness in generic embedders is not straightforward. Since they are task-agnostic by design, traditional fairness metrics that rely on labels or specific downstream performance are difficult to apply. Moreover, the challenge could be even greater in realistic scenarios where access to sensitive demographic attributes is restricted due to privacy regulations such as GDPR. This has motivated recent work on fairness without demographics, which aims to provide principled alternatives when protected attributes are unavailable (Kenfack et al., 2024). Importantly, recent studies show that even the latest generation of large models continues to exhibit biases and can produce disparate harms across groups (Chehbouni et al., 2024). This highlights the persistent tension between the technical success of foundation models and the ethical requirements of fairness.

In this context, the present work first investigates whether it is possible to evaluate and compare embedders in terms of fairness without relying on labeled data, and further explores the more challenging setting where demographic attributes are also unavailable.

1.2 Development of the Research Topic

The starting point of this internship is the suggestion of my supervisors, Pablo Piantanida and Ulrich Aïvodji, to investigate fairness evaluation for text embedders. Conventional fairness metrics usually rely on labels and demographic attributes, but these requirements are not always met in practice. The challenge is thus to design a framework that compares embedders in terms of fairness without access to such information.

The initial intuition of this internship was to build on the *Information Sufficiency Score* (IS), originally introduced as a task-agnostic measure of embedding informativeness (DARRIN et al., 2024), and to investigate whether it could be extended to serve as a proxy for fairness. This led to a critical examination of group-conditional variants of IS, including a direct extension to group-sensitive settings, which we show empirically to be insufficient for producing a reliable fairness ranking in this context. Motivated by this limitation, I then propose *FairEntropy*, a simpler and significantly more efficient entropy-based metric that operates purely at the representation level and avoids pairwise, relative comparisons. Beyond the theoretical analysis, I designed and implemented an experimental pipeline to validate this approach empirically across multiple datasets. Concretely, I (i) review the literature on fairness in machine learning, both with and without demographics, to identify suitable definitions and evaluation criteria; (ii) analyze specific challenges arising in the fairness assessment of embedders and large language models, and (iii) evaluate the proposed FairEntropy through controlled experiments. See the internship timeline for a detailed plan in Figure 1.

This approach refines an initial intuition into a concrete research contribution, resulting in a theoretically grounded, task-agnostic, and computationally efficient framework for assessing representational fairness in embedding models.

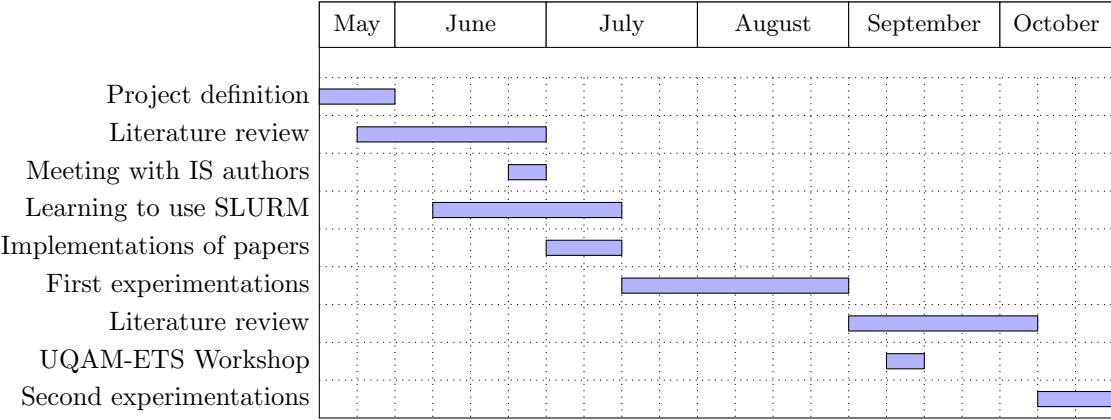


Figure 1: Gantt chart of the internship

1.3 Reformulation of the Problem

The research question addressed in this internship is how to rank different embedders with respect to fairness in a relative manner. The ultimate goal is to provide a diagnostic framework that allows a decision maker, given a set of pre-trained models, to identify which one is the fairest according to a well-motivated definition of fairness. This perspective is particularly relevant because retraining or fine-tuning large models is computationally expensive and often impractical.

The core contribution of this research is to investigate fairness evaluation when demographic attributes are available. In this setting, our framework is validated by testing whether it correlates with established supervised metrics such as Worst-Group Accuracy (WGA). This forms the main part of the work, as it provides a theoretically grounded and empirically tested proxy for fairness in embedders. Building on this foundation, we then explore the more challenging but realistic case where demographics are not accessible, reporting preliminary results that suggest promising directions for future research.

It is worth noting that the framework assumes that the decision maker has some knowledge of the general domain in which the embedders will be applied (e.g., medicine, finance, customer interaction, or multi-domain contexts). Finally, the fairness ranking established by our framework should not be interpreted as a global performance ranking: fairness and global accuracy are distinct dimensions, and our metric is explicitly designed to address fairness.

1.4 Overview of the Document

This report first reviews the literature within the broader landscape of fairness in machine learning, distinguishing between approaches that rely on demographic attributes and those that operate without them, as well as considering unsupervised metrics that do not depend on labels. Building on this foundation, the methodology is then introduced, detailing the design of a new fairness diagnostic, the evaluation strategy, and the datasets and models used. The subsequent section presents the experimental results, contrasting the demographic and non-demographic settings and assessing the originality and applicability of the approach. The report concludes with a discussion of future research directions, the main learnings gained during the internship, and the ethical considerations surrounding this work.

2 Literature Review

2.1 Fairness with Demographics

Sources of bias. Biases in machine learning systems can emerge at multiple stages of the pipeline. Mehrabi et al. (2021) distinguishes three main categories: data bias, algorithmic bias, and evaluation bias. Data-related issues include historical bias, sampling bias, and representation bias, all of which reflect inequalities in the underlying distribution of examples. Algorithmic bias arises when optimization objectives favor majority groups or when models are poorly specified for minority populations. Finally, evaluation and deployment can amplify existing disparities if benchmarks are not representative or if user interaction reinforces biased feedback loops. While this taxonomy provides a high-level understanding, the crucial point is that these biases are often

entangled and difficult to isolate, which motivates the development of fairness-aware definitions and methods.

Fairness definitions. A central contribution of the literature has been to clarify what it means for a model to be “fair.” Verma and Rubin (2018) provide a systematic taxonomy of fairness definitions, broadly divided into group fairness, individual fairness, and causal reasoning. Group fairness requires that individuals sharing a sensitive attribute (e.g., gender) receive statistically comparable outcomes. Notable criteria include statistical parity, predictive parity, equal opportunity, and equalized odds. Individual fairness instead focuses on treating similar individuals similarly, usually relying on a task-specific similarity metric. Causal reasoning approaches such as counterfactual fairness propose that a model’s prediction should remain unchanged in a counterfactual world where only the sensitive attribute is altered. These definitions are not equivalent and may even be mutually incompatible, illustrating the normative nature of fairness choices. Figure 2 illustrates a taxonomy of the most popular fairness definitions.

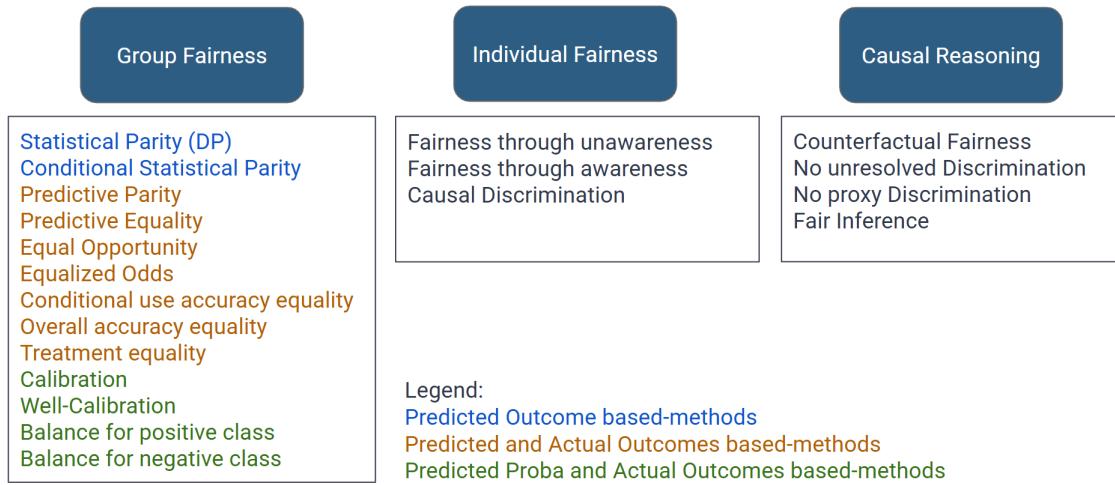


Figure 2: Taxonomy of fairness *definitions* with demographics, adapted from Verma and Rubin (2018). Evaluation metrics such as WGA and optimization methods such as DRO are discussed in the text.

Mitigation strategies. Once fairness is defined, the challenge is to design algorithms that mitigate unfairness. Caton and Haas (2024) categorizes these approaches into three main families. Pre-processing methods aim to modify the data before training, for example by reweighting, resampling, or learning fair representations. The line of work on fair representation learning (Zemel et al., 2013; Edwards & Storkey, 2015; Madras et al., 2018; Balunovic et al., 2021) explicitly seeks encoders that obfuscate sensitive information while preserving task utility. In-processing methods introduce fairness constraints directly into the training objective, using techniques such as adversarial learning, constrained optimization, or regularization. Post-processing methods adjust predictions after training, e.g., by thresholding, calibration, or equalized odds adjustments. The suitability of each of these categories often depends on the availability of sensitive attributes and on the label during training or deployment.

Evaluation metrics. Fairness evaluation relies on quantitative measures that capture disparities between groups. Classical metrics include statistical parity difference, disparate impact ratio, equalized odds difference, and calibration measures. While these definitions highlight different aspects of fairness, they often focus on average-case behavior across groups. In contrast, more recent work emphasizes worst-case performance as a robust criterion. The *Worst-Group Accuracy* (WGA) (Sagawa* et al., 2020) captures the accuracy of the least advantaged group:

$$\text{WGA}(f) = \min_{g \in \mathcal{G}} \Pr [f(X) = Y \mid g],$$

where f is the predictor, Y the true label, and \mathcal{G} the set of sensitive groups. This metric embodies the principle that “no group should be left behind,” and has become a standard baseline in the fairness literature. Closely related, distributionally robust optimization (DRO) is not an evaluation metric but a training method that explicitly minimizes worst-case risk across groups (Hashimoto

et al., 2018). It complements WGA by targeting the same fairness principle during model learning rather than evaluation.

Relevance to this work. In the context of this research, WGA provides a well-motivated supervised reference, against which we can validate whether our proposed label-free proxy (described later in this paper) is meaningful. Once this link is established with demographics, the framework can then be extended to the more challenging case where no labels or sensitive attributes are available, which is precisely the scenario embedders are most often deployed in.

2.2 Fairness without Demographics

Sources of bias. The absence of demographic attributes does not eliminate bias. Models may still amplify structural inequalities present in the data, even when sensitive features are removed. Indeed, proxies for demographic attributes often remain in the feature space, creating indirect pathways for discrimination. This motivates definitions and methods explicitly designed for fairness without demographics.

Fairness definitions. Kenfack et al. (2024) provide a recent survey identifying four main approaches: individual fairness, fairness through unawareness, proxy fairness, and Rawlsian min–max fairness. Individual fairness requires similar individuals to receive similar predictions. Fairness through unawareness excludes sensitive features from decision-making, though correlated attributes may still introduce bias. Proxy fairness replaces true demographic groups with approximated groups (e.g., via clustering), on which fairness metrics can be applied. Finally, Rawlsian min–max fairness requires maximizing the worst-case group performance, even when the groups are not pre-defined. (Hashimoto et al., 2018). Figure 3 illustrates a recent taxonomy of the fairness definitions when sensitive attributes are not accessible.

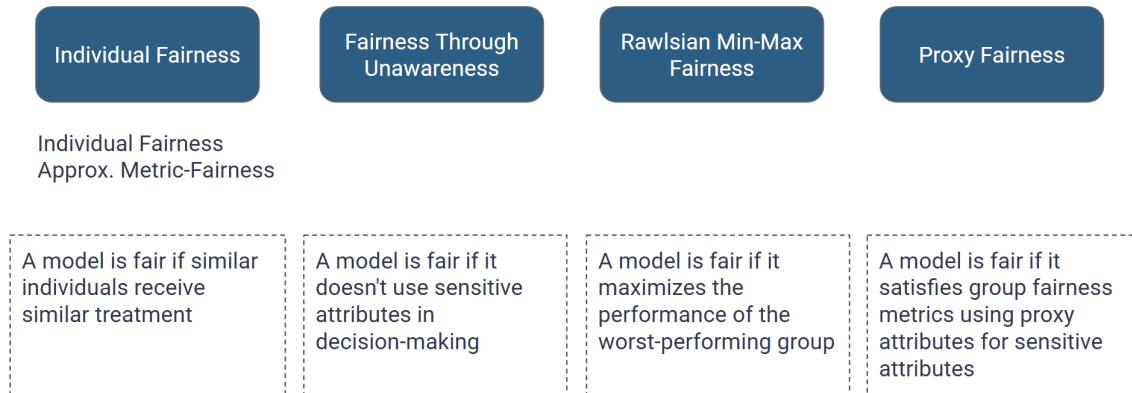


Figure 3: Taxonomy of fairness *without demographics*, adapted from Kenfack et al. (2024). In this work, proxy fairness is instantiated through k-means clustering and worst-case aggregation.

Mitigation strategies. Several strategies have been proposed for fairness without demographics. Representation learning methods aim to learn embeddings where fairness emerges from latent structure (Lahoti et al., 2020). Clustering-based approaches create proxy groups and then apply group fairness criteria to them. Distributionally robust optimization (DRO) extends naturally to this setting by treating latent groups as potential worst-case subpopulations and minimizing their maximum risk during training (Hashimoto et al., 2018). Other methods approximate sensitive attributes through proxies, for example by leveraging correlated features or adversarially reweighted learning, in order to enforce fairness constraints without direct access to demographics. Finally, heuristic methods rely on weak supervision or external knowledge to approximate demographic partitions. Some work on fair class balancing demonstrates that fairness can also be enhanced directly in training by artificially augmenting the data (Yan et al., 2020).

Evaluation metrics. Evaluating fairness without demographics remains difficult, as most methods still assume access to task labels to compute group-wise performance. For instance, proxy-based WGA requires sensitive attributes to define the groups to estimate accuracy per cluster. In this sense, existing approaches remain fundamentally supervised, even when sensitive attributes are not

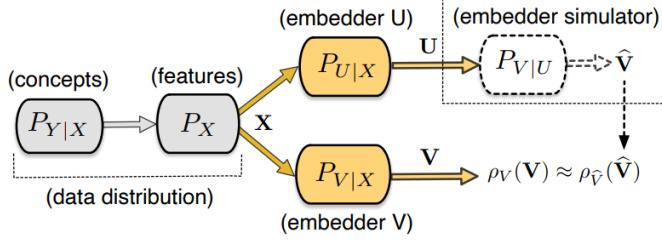


Figure 4: Channel simulation view of the Information Sufficiency score taken from DARRIN et al. (2024). Embedders U and V induce conditional distributions $P_{U|X}$, $P_{V|X}$. The deficiency quantifies how well V can be simulated from U .

available. **This limitation is crucial for foundation models and embedders**, where both labels and demographics are absent.

Relevance to this work. Fairness without demographics is directly relevant to this research. Large-scale text embedders are trained to be task-agnostic and to give an abstract representation of an input, often without demographic information. Existing approaches in the literature generally assume labels are available, which limits their applicability in this context.

In this work, proxy fairness is operationalized through unsupervised clustering of the latent space, and a fairness-aware extension of the Information Sufficiency (IS) score (discussed later in this paper) is computed. The method will be discussed in detail in the rest of the paper.

2.3 Information Sufficiency Score

As we discussed, the main concern is the lack of knowledge around the downstream task for embedders. In other words, we cannot count on the label to rank the embedders provided.

A recent line of work introduces the *Information Sufficiency* (IS) score as a task-agnostic way to evaluate and compare embedders (DARRIN et al., 2024). The foundation of this approach lies in the notion of *deficiency* (Cam, 1964), which quantifies the discrepancy between two statistical experiments.

Let us define one of the most important terms of what is following. Given two embedding models U and V , represented as channels $P_{U|X}$ and $P_{V|X}$ from the input space X , the deficiency from U to V is

$$\delta(P_{U|X} \rightarrow P_{V|X}) = \inf_{M \in \mathcal{K}(V|U)} \mathbf{E}[\|M \circ P_{U|X} - P_{V|X}\|_{\text{TV}}],$$

where $\mathcal{K}(V|U)$ denotes the set of Markov kernels from U to V . Intuitively, this measures how well the distribution of V can be simulated from U .

Le Cam showed that the difference in Bayes risks between U and V is controlled by this deficiency:

$$R_U - R_V \leq \delta(P_{U|X} \rightarrow P_{V|X}),$$

which provides a theoretical bridge between channel simulation and downstream predictive risk. In practice, if U can simulate V with low deficiency, then U cannot perform substantially worse than V across tasks.

Formal definition of IS. Building on this, DARRIN et al. (2024) define the IS between U and V as

$$IS(U \rightarrow V) = \inf_{f \in \mathcal{F}_{\Theta}(V)} \mathbb{E}[-\log f(V)] - \mathbb{E} \left[\inf_{M \in \mathcal{K}_{\Theta}(V|U)} \mathbb{E}[-\log M(V|U) | U] \right],$$

where the first term measures the inherent uncertainty of V , and the second term the residual uncertainty when simulating V from U . A higher $IS(U \rightarrow V)$ means that U reduces more uncertainty about V , i.e. it is more informative.

Aggregating across embedders. Given a set of embedders $\mathcal{E} = \{E_1, \dots, E_m\}$, pairwise IS values are computed for all ordered pairs (E_i, E_j) . The overall IS score of an embedder E_i is then defined as the *median* of its outgoing comparisons:

$$\text{IS}(E_i) = \text{median}\left\{ \text{IS}(E_i \rightarrow E_j) \mid j \neq i \right\}.$$

The median is used instead of the mean to ensure robustness against outliers and asymmetries between particular pairs of embedders.

Algorithm 1 Computation of IS Score

```

1: Input: Dataset  $X = \{x_n\}_{n=1}^N$ , embedders  $\mathcal{E} = \{E_1, \dots, E_m\}$ 
2: for each pair  $(E_i, E_j)$  do
3:   Compute embeddings  $U = E_i(X)$  and  $V = E_j(X)$ 
4:   Learn simulator  $M \in \mathcal{K}_\Theta(V|U)$  (e.g. Gaussian mixtures or Neural Network)
5:   Estimate  $\text{IS}(E_i \rightarrow E_j)$  using the definition above
6: end for
7: for each embedder  $E_i$  do
8:    $\text{IS}(E_i) \leftarrow \text{median}\left(\left\{ \frac{\text{IS}(E_i \rightarrow E_j)}{d_{E_i}} : j \neq i \right\}\right)$ , where  $d_{E_i}$  denotes the embedding dimension of  $E_i$ .
9: end for
10: Output: IS scores  $\{\text{IS}(E_i)\}_{i=1}^m$  used to rank the embedders

```

Empirical validation. The IS score was validated on 56 NLP datasets across tasks including retrieval, classification, clustering, semantic textual similarity (STS), and reranking. Results show strikingly high correlations with downstream task performance: average Kendall $\tau = 0.73$, Spearman $\rho_s = 0.90$, Pearson $\rho_p = 0.94$ across tasks. The following table summarizes their findings:

Task	ρ_p	ρ_s	τ
Retrieval (15 datasets)	0.89	0.89	0.69
Classification (12 datasets)	0.92	0.88	0.73
Clustering (11 datasets)	0.86	0.85	0.66
STS (10 datasets)	0.92	0.83	0.63
Reranking (4 datasets)	0.84	0.78	0.64
Average (56 datasets)	0.94	0.90	0.73
Additional Classif (8 datasets)	0.89	0.84	0.66

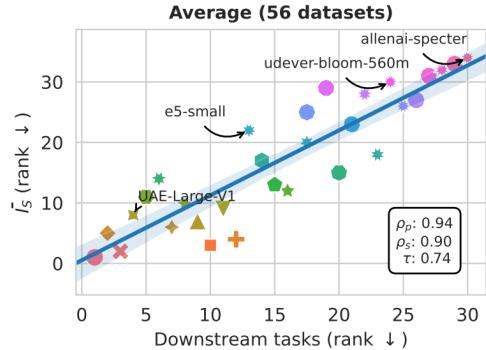


Figure 5: Empirical validation of the IS score on NLP benchmarks taken from DARRIN et al. (2024). The left panel reports the correlation coefficients (Pearson ρ_p , Spearman ρ_s , and Kendall τ) between IS-based rankings and downstream task performance across multiple tasks. The right panel illustrates the empirical relationship between IS scores and the average downstream performance, showing the strong monotonic trend across embedders.

Strengths and limitations. The IS score offers several advantages: it is fully task-agnostic (no need of evaluating each embedder on every downstream tasks) and requires no labels, it correlates strongly with downstream performance across a wide range of NLP benchmarks, and it allows the comparison of embedders of heterogeneous nature. However, it also has limitations: the estimation is quadratic in the number of embedders, it depends on the choice of simulator family (often Gaussian mixtures), and requires some prior assumptions about the domain of the task the embedders will be used on (e.g., medicine, finance, customer interaction, or multi-domain contexts).

These results suggest that the IS score provides a robust, task-agnostic proxy for ranking embedders in NLP, without requiring task-specific labels. While originally designed for general informativeness, this score forms the theoretical and algorithmic foundation of our fairness-oriented adaptation.

3 Methodology

3.1 From FairIS to FairEntropy: Creation of the Score

3.1.1 Base score with demographics

As discussed above, a key difficulty in evaluating fairness without access to labels is that standard definitions, such as worst-group accuracy (WGA), rely on groupwise downstream performance. In the absence of task labels, we must rely on representation-level criteria to assess whether an embedder treats sensitive groups in a balanced manner.

The theoretical framework based on deficiency provides a principled connection between representation quality and worst-group Bayes risk. In particular, extending the deficiency framework of DARRIN et al. (2024) and Cam (1964) to the groupwise setting yields the following bound.

Worst Group Accuracy in the context. For an embedding model E and a group $z \in \mathcal{Z}$, let $R_E(z)$ denote the Bayes risk, i.e. the minimal classification error achievable from $E(X)$ on the conditional distribution $P_{X,Y|Z=z}$. The worst-group Bayes risk is then

$$\Delta R(E) = \max_{z \in \mathcal{Z}} R_E(z),$$

and the corresponding worst-group accuracy (WGA) is

$$\text{WGA}(E) = 1 - \Delta R(E).$$

A model is considered more fair if it achieves a higher WGA, i.e. if the performance of its worst-performing group is higher.

Extending the deficiency framework of DARRIN et al. (2024) and Cam (1964) to the groupwise case leads to the following bound:

Proposition 1. *For any two embedding models U, V ,*

$$\Delta R(U) - \Delta R(V) \leq \max_{z \in \mathcal{Z}} \delta_z(U \rightarrow V),$$

where $\delta_z(U \rightarrow V)$ is the deficiency restricted to group z .

The proof of this proposition is available in the Appendix.

This result suggests that, in principle, groupwise efficiencies control the degradation of worst-group Bayes risk between representations. A natural idea, following DARRIN et al. (2024), is therefore to approximate these deficiencies using group-conditional Information Sufficiency (IS) scores.

We implemented this approach by computing IS scores restricted to each sensitive group and aggregating them in a worst-group fashion, mirroring Worst-Group Accuracy. However, empirical results reveal that under group conditioning, reduced distributional diversity causes IS scores to saturate, and worst-group aggregation further amplifies this effect, leading to coarse rankings. So groupwise IS primarily clusters embedders according to their mutual simulability. In terms of fairness, it's less appropriate when compared to absolute structural metrics. (see figure 9)

This behavior is consistent with the nature of deficiency and IS: they induce a partial order based on pairwise simulability, rather than an absolute scalar ranking, and become particularly noisy when estimated on group-conditioned distributions.

Group-Conditional Entropy as a Fairness Criterion. Motivated by these observations, we propose a simpler, fully representation-level fairness diagnostic based on group-conditional entropy, termed FairEntropy, which departs from IS-style relative comparisons by providing an absolute diagnostic of group-wise representational dispersion.

Rather than approximating deficiencies or Bayes-risk bounds, this score quantifies whether an embedder allocates comparable representational capacity across sensitive groups.

For an embedding model E and a sensitive group $z \in \mathcal{Z}$, we define the groupwise entropy as

$$H_z(E) = H(E(X) | Z = z),$$

where $H(\cdot)$ denotes a *differential entropy* estimate computed over the continuous embedding distribution, estimated from samples.

The overall fairness diagnostic score of an embedder is then given by the least informative group:

$$\text{FairEntropy}(E) = \min_{z \in \mathcal{Z}} \frac{H_z(E)}{d_E}, \quad \text{where } d_E \text{ denotes the embedding dimension of } E.$$

The underlying hypothesis is that low group-conditional entropy may reflect representational collapse or over-compression, which can limit the diversity of features available to downstream classifiers for that group across tasks. From this perspective, entropy can be viewed as a task-agnostic indicator of the expressive capacity available to the worst-off group, rather than as a direct measure of downstream performance.

The construction of FairEntropy directly mirrors the worst-group perspective underlying WGA: an embedder is considered favorable under our diagnostics only if its least favorable group retains sufficient representational variability. Low entropy for a given group indicates representation collapse or over-compression, which is likely to limit downstream separability for that group across tasks. To ensure comparability across embedders with different dimensionalities, the entropy score is normalized by the embedding dimension.

FairEntropy is intended as a relative ranking criterion within a given pool of embedders, rather than as an absolute, scale-invariant fairness score. While differential entropy is theoretically sensitive to scale, we empirically observe that enforcing ℓ_2 -normalization of embedding vectors removes variations that are predictive of worst-group performance across standard benchmarks. We therefore compute FairEntropy on raw embedding representations, treating representational dispersion as a meaningful component of embedding structure rather than as a nuisance factor.

FairEntropy does not aim to capture all forms of unfairness, but focuses on representational imbalance as an important contributing condition for downstream group disparities. Nonetheless, it remains purely analytical, task-agnostic, and fast to compute: it requires no labels, no downstream optimization, and can be applied post-hoc to any pretrained embedder.

Algorithm 2 describes the proposed procedure.

Algorithm 2 Computation of FairEntropy (with demographics)

```

1: Input: Dataset  $X = \{(x_n, z_n)\}_{n=1}^N$  with sensitive groups  $\mathcal{Z}$ , embedders  $\mathcal{E} = \{E_1, \dots, E_m\}$ 
2: for each embedder  $E_i \in \mathcal{E}$  do
3:   for each group  $z \in \mathcal{Z}$  do
4:      $X_z \leftarrow \{x_n : z_n = z\}$                                  $\triangleright$  fixed (true) demographic group
5:      $U \leftarrow E_i(X_z)$ 
6:     Estimate  $H_z(U)$  (entropy estimation)
7:   end for
8:    $\text{FairEntropy}(E_i) \leftarrow \min_{z \in \mathcal{Z}} \frac{H_z(U)}{d_{E_i}}$ ,    where  $d_{E_i}$  denotes the embedding dimension of  $E_i$ .
9: end for
10: Output:  $\{\text{FairEntropy}(E_i)\}_{i=1}^m$ 
```

Interpretation. This construction ensures that an embedder receives a high score only if its worst-case sensitive group retains sufficient representational diversity. By taking the minimum across groups, the method enforces a worst-case perspective consistent with fairness principles such as Worst-Group Accuracy (WGA). Finally, FairEntropy operates purely at the representation level and does not require access to task labels or downstream objectives. We emphasize that FairEntropy is inherently defined with respect to a specific group partition Z ; as a consequence, the resulting fairness assessment is conditional on the chosen sensitive attributes and their granularity, and different group definitions may induce different fairness rankings even for the same embedding model.

We do not claim that FairEntropy constitutes a universal or normative definition of fairness. Rather, it should be understood as a fast-to-compute representation-level diagnostic that strongly correlates with worst-group performance.

Entropy estimation. We estimate group-wise marginal entropies from continuous embeddings using the KNIFE estimator from the Information Sufficiency codebase (DARRIN et al., 2024), which models densities with Gaussian mixture kernels. In our setting, embeddings are treated as continuous variables and we report only the marginal entropy term $H(Y)$, computed from the learned marginal density and normalized by the embedding dimension. The estimator relies on a mixture-of-Gaussians parameterization with diagonal (optionally extended) covariance structure

and is trained with fixed hyperparameters and early stopping based on entropy stabilization. In all experiments we use `margin_modes=8` with `batch_size=256`.

3.1.2 Extension without demographics.

In addition to the supervised setting, we also consider the more realistic case where demographic attributes are not available, for instance due to privacy regulations (e.g., GDPR) or because they are not collected at scale. In such situations, we extend FairEntropy by constructing proxy groups in an unsupervised manner, approximating latent subgroup structure directly from the embedding space. This is not the main focus of the present work, but rather a natural and practical extension of the framework toward deployment scenarios where sensitive attributes cannot be observed.

Clustering-based grouping. Given an embedder E , we apply k -means clustering to its representation space $E(X)$, yielding a partition of the dataset into clusters $\{C_1, \dots, C_k\}$. These clusters act as surrogate groups, replacing explicit demographics.

Unsupervised FairEntropy score. Let $C \in \{1, \dots, k\}$ denote the random variable corresponding to cluster assignment, obtained by applying k -means to the representations $E(X)$.

For a fixed embedder E , we define the family of cluster-conditional entropies $\{H_c(E)\}_{c=1}^k$ by

$$\forall c \in \{1, \dots, k\}, \quad H_c(E) = H(E(X) | C = c).$$

The fairness score of E is then defined analogously to the demographic case, by taking the worst-performing cluster:

$$\text{FairEntropy}_{\text{unsup}}(E) = \min_{c \in \{1, \dots, k\}} H_c(E).$$

This procedure preserves the worst-case perspective of WGA while avoiding reliance on demographic attributes. The algorithm used is described below.

The underlying idea is that sensitive groups, or at least subpopulations exhibiting different behaviors, often manifest as distinct regions of the embedding space. By applying k -means clustering, we partition the embeddings into proxy groups that approximate latent structures. Although these clusters do not correspond to true demographic categories, they capture heterogeneity in the data distribution. Evaluating fairness on such proxy groups reflects the hypothesis that any systematic disparity across clusters may signal hidden biases, in the same way disparities across demographic groups reveal unfairness.

Algorithm 3 Computation of Unsupervised FairEntropy (via k -means)

```

1: Input: Dataset  $X = \{x_n\}_{n=1}^N$ , embedders  $\mathcal{E} = \{E_1, \dots, E_m\}$ , number of clusters  $k$ 
2: for each embedder  $E_i \in \mathcal{E}$  do
3:    $U_i \leftarrow E_i(X)$ 
4:   Run  $k$ -means on  $U_i$  to obtain cluster assignments  $C_i(x_n) \in \{1, \dots, k\}$ 
5:   for each cluster  $c \in \{1, \dots, k\}$  do
6:      $X_c \leftarrow \{x_n : C_i(x_n) = c\}$ 
7:      $U \leftarrow E_i(X_c)$ 
8:     Estimate  $H_c(U)$  (entropy estimation)
9:   end for
10:   $\text{FairEntropy}_{\text{unsup}}(E_i) \leftarrow \min_{c \in \{1, \dots, k\}} \frac{H_c(E_i)}{d_{E_i}}$ , where  $d_{E_i}$  denotes the embedding dimension of  $E_i$ .
11: end for
12: Output:  $\{\text{FairEntropy}_{\text{unsup}}(E_i)\}_{i=1}^m$ 

```

Interpretation. The proposed unsupervised variant of FairEntropy should be understood as a partition-dependent diagnostic: the clustering procedure induces a set of proxy groups that reflect a particular geometric hypothesis about latent subpopulations, and FairEntropy quantifies worst-case representational dispersion relative to this induced partition rather than capturing an intrinsic or universal notion of fairness. Although the resulting clusters do not correspond to explicit sensitive attributes, any systematic disparity in entropy across clusters may reveal hidden structural biases in the embedding.

3.2 Evaluation Strategy

The purpose of the evaluation is to validate whether the proposed FairEntropy score is a reliable proxy for Worst-Group Accuracy (WGA). Since WGA is label-dependent, we construct a controlled binary classification task on benchmark fairness datasets. These datasets are widely used in the literature for fairness analysis and contain sensitive demographic attributes. It will be presented below.

The evaluation proceeds in three steps. First, we train a simple logistic regression model on the full dataset, using one’s embeddings as input features. Second, we evaluate the trained model on each subgroup—either defined by demographics (supervised case) or by clusters obtained via k -means in embedding space (unsupervised case). Finally, we compute WGA for every embedders as the minimum subgroup accuracy and measure its correlation with the FairEntropy score across embedders. We report three correlation coefficients: Pearson’s ρ_p , Spearman’s ρ_s , and Kendall’s τ , providing a robust assessment of monotonic and linear associations.

Algorithm 4 Evaluation Strategy

```
1: Input: Dataset  $X = \{(x_n, y_n, z_n)\}_{n=1}^N$ , embedders  $\mathcal{E}$ , groups  $\mathcal{Z}$ 
2: for each embedder  $E \in \mathcal{E}$  do
3:   Compute embeddings  $U = E(X)$ 
4:   Train logistic regression on  $(U, y)$ 
5:   for each group  $z \in \mathcal{Z}$  do
6:     Evaluate accuracy  $Acc_E(z)$ 
7:   end for
8:   Compute WGA:  $WGA(E) = \min_{z \in \mathcal{Z}} Acc_E(z)$ 
9:   Retrieve score FairEntropy( $E$ )
10: end for
11: Compute correlation  $\text{corr}(\{\text{FairEntropy}(E)\}, \{WGA(E)\})$ 
```

This protocol ensures that correlations are assessed on a downstream task familiar in the fairness literature, but with the sole purpose of validating our metric. The logistic regression is deliberately simple, so as not to confound the evaluation with model-specific complexities.

For the unsupervised setting without demographics, the algorithm is the same but the algorithm is evaluated on fixed groups from the Cartesian product between *Sex* and *Race*. In the unsupervised setting, sensitive attributes are used exclusively for evaluation purposes to compute WGA, and are never used in the computation of FairEntropy itself.

3.3 Datasets

We conduct experiments on tasks drawn from the Folktale suite (Ding et al., 2022), which are based on American Community Survey (ACS) Public Use Microdata Sample (PUMS) data. In order to remain focused and manageable, we select three of the canonical ACS tasks: **Income**, **Employment**, and **Public Coverage**.

- **ACSIIncome.** Predict whether an individual’s annual income exceeds \$50,000. The dataset is filtered to include only individuals older than 16 who reported working at least one hour per week and whose reported earnings are at least \$100.
- **ACSEmployment.** Predict whether an individual is employed. We restrict the sample to individuals between 16 and 90 years old.
- **ACSPublicCoverage.** Predict whether an individual has government-provided health insurance. We filter to include individuals younger than 65 and those whose income is less than \$30,000, focusing on low-income populations not covered by Medicare.

These three tasks span different socioeconomic phenomena while presenting real-world fairness challenges. For each task, we randomly sample 100,000 individuals from the 2018 ACS PUMS, pooling data across all U.S. states.

Preprocessing for embeddings. To integrate these tabular datasets into embedding-based pipelines, we transform each individual’s record into a textual description (a sentence or a short phrase) such that each attribute-value pair is represented explicitly in natural language. This

transformation allows a text embedder (e.g. sentence transformer) to process the data in the same format as standard textual input. Techniques for converting table rows or structured data into text have been explored in the literature—for instance, the TabLLM framework proposes to convert each line in simple statements for classification (Hegselmann et al., 2023). We apply a straightforward template-based approach (e.g. *“She is 37 years old. She identifies as female. She is white alone. She is never married or under 15 years old.”*) to ensure that each attribute is articulated as a short clause.

The textualization procedure may itself influence representational geometry; as such, the reported fairness diagnostics should be interpreted in light of this design choice, which reflects a realistic deployment scenario for text embedders applied to structured data.

3.4 Embedding models

To evaluate fairness, we consider a diverse set of embedding models spanning different generations and training paradigms. The selection follows a taxonomy based on fine tuning methods and intended use cases. This taxonomy ensures that our analysis does not rely on a single family of models, but rather covers baselines, general-purpose models, and specialized architectures.

Baseline models. Early word embedding approaches such as Word2Vec and GloVe provide static vector representations of words without any attention mechanism. Although largely surpassed by transformer-based encoders, they serve as useful baselines for highlighting the evolution of fairness properties across generations.

General self-supervised models. Modern encoders trained on large corpora with self-supervised objectives form the backbone of current NLP applications. We include lightweight sentence-transformer variants such as `all-MPNet-base-v2`, `all-MiniLM-L12-v2`, and `all-DistilRoBERTa-v1`, as well as more recent second-generation models like `GTE-large`, `Ember-v1`, and `Stella-base-en-v2`. These models typically combine masked language modeling (MLM) with contrastive learning or distillation, yielding general-purpose representations.

Paraphrase-supervised models. A different line of work directly optimizes embeddings for semantic similarity between paraphrases. Representative models include `MPNet-paraphrase-v2`, `MiniLM-L12-paraphrase-v2`, and its multilingual extension covering over 50 languages, as well as `DistilRoBERTa-paraphrase-v1`. These models emphasize sentence-level meaning preservation.

NLI and STS supervised models. Natural Language Inference (NLI) and Semantic Textual Similarity (STS) benchmarks have motivated another family of supervised embedders. Examples include `BERT-base-NLI`, `DistilBERT-base-NLI-STBS`, `sup-simcse-bert-base`, and `UAE-Large-V1`. These models are trained to capture logical relations between premises and hypotheses or graded semantic similarity, which often improves reasoning sensitivity.

Retrieval-focused models. Finally, several embedders are explicitly optimized for retrieval tasks. These include `E5-large-v2`, `gtr-t5-large`, `sentence-t5-large`, and `LaBSE`. They leverage either encoder-decoder backbones (e.g., T5 variants) or multilingual setups (e.g., LaBSE), with the objective of matching documents and queries across languages or domains.

This taxonomy reflects the diversity of the embedding ecosystem. By spanning static word vectors, general-purpose transformers, paraphrase- and NLI-focused encoders, and retrieval-specialized architectures, our evaluation covers both classic baselines and state-of-the-art systems. This variety is crucial to assess whether fairness properties depend more on the training paradigm or on the intended use case of the embedder. DARRIN et al. (2024) shows that the IS score stabilizes only when computed across a sufficiently large pool of models (≈ 15 embedders in NLP). A detailed summary of the models used, with their dimensions and maximum token lengths, is provided in subsection 11.2.

4 Experiments and Results

4.1 Fairness with demographics

Table 3 reports the correlation between the proposed FairEntropy score and Worst-Group Accuracy (WGA) across ACS datasets when true demographic attributes are available. Across the evaluated tasks, FairEntropy shows a clear monotonic relationship with WGA on several datasets (up to average $\rho_p \approx 0.83$), while the strength of this relationship can vary on some specific known datasets.

Dataset	Pearson ρ_p	Spearman ρ_s	Kendall τ
ACSIIncome	$0.82^* \pm 0.000$	$0.78^* \pm 0.047$	$0.61^* \pm 0.048$
ACSPublicCoverage	$0.88^* \pm 0.000$	$0.80^* \pm 0.025$	$0.64^* \pm 0.033$
ACSEmployment	$0.80^* \pm 0.000$	$0.60^* \pm 0.051$	$0.48^* \pm 0.053$
Average	0.83	0.72	0.58

Table 1: Correlation between FairEntropy scores and WGA across ACS datasets with demographics. 4 groups based on the cartesian product between “Sex” and “Race”. * indicates that all of the 3 seeds used have a statistically significant correlations ($p < 0.05$)

Figure 6 visualizes these correlations at the level of individual embedders, highlighting alignment between FairEntropy and worst-group accuracy.

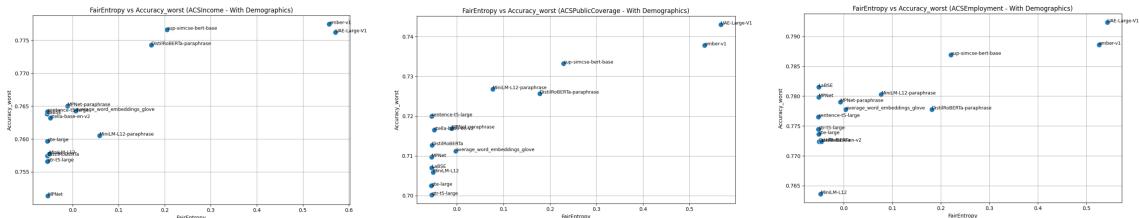


Figure 6: Correlation between FairEntropy score and WGA across ACS datasets with demographics: (left) ACSIIncome, (center) ACSPublicCoverage, and (right) ACSEmployment. 4 groups based on the cartesian product between “Sex” and “Race”.

Across datasets, a clear monotonic relationship is generally observed: embedders achieving higher FairEntropy scores tend to exhibit higher worst-group accuracy. This indicates that FairEntropy captures representation-level geometrical properties that are strongly associated with downstream fairness.

While representation quality is a major driver of worst-group performance, it is interesting to note that WGA could also be influenced by task-specific decision boundaries and label base rates, as also noted in prior analyses of ACS benchmarks (Mehrabi et al., 2021; Ding et al., 2022). Figure 6 illustrates these task-dependent deviations at the level of individual embedders.

Beyond dataset-specific effects, the relationship between FairEntropy and WGA also depends on how sensitive groups are defined. To analyze this effect under challenging conditions, we focus on ACSEMPLOYMENT, which consistently exhibits the weakest correlations among the ACS benchmarks. Table 2 reports correlation scores for multiple group partitions on this dataset, allowing us to assess the sensitivity of FairEntropy to group granularity in a setting where downstream fairness is known to be harder to capture.

Groups	Pearson ρ_p	Spearman ρ_s	Kendall τ
Sex	$0.63^* \pm 0.000$	$0.51^* \pm 0.022$	$0.36^* \pm 0.039$
Sex + Age	$0.57^* \pm 0.001$	0.43 ± 0.053	0.27 ± 0.058
Sex + Race	$0.80^* \pm 0.000$	$0.60^* \pm 0.036$	$0.48^* \pm 0.022$
Sex + Race + Age	0.31 ± 0.000	0.22 ± 0.042	0.14 ± 0.048

Table 2: Correlation between FairEntropy scores and WGA on ACSEmployment across different sensitive attribute groupings. * indicates that all of the 3 seeds used have a statistically significant correlations ($p < 0.05$)

Consistent with prior work on subgroup fairness, we observe that the definition and granularity of sensitive groups has a critical impact on the measured alignment with worst-group accuracy Kearns et al., 2018. As shown in Table 2 on ACSEMPLOYMENT—the dataset where correlations are overall the weakest—increasing the number of sensitive attributes does not systematically improve correlation with WGA and can instead introduce instability across random seeds. In particular, the Cartesian product of *Sex* and *Race* (four subgroups (*white men, non-white men, white women, non-white women*) yields the most stable and consistently significant correlations, whereas finer partitions involving age lead to noisier estimates and reduced statistical significance. These results highlight that selecting relevant sensitive attributes is a non-trivial but crucial design choice, and motivate our focus on *Sex* × *Race* groupings for the remainder of this study.

Overall, when sensitive attributes are available, FairEntropy exhibits a strong and statistically robust monotonic association with worst-group accuracy. These results confirm that representation-level uncertainty, when evaluated over carefully chosen sensitive groupings, provides a meaningful proxy for downstream fairness disparities. This sensitivity to group definitions further motivates the need for fairness diagnostics that remain applicable when sensitive attributes are unavailable.

4.2 Fairness without demographics

We now consider a more realistic setting in which sensitive demographic attributes are unavailable and proxy groups are induced through unsupervised clustering. In this context, FairEntropy is computed over clusters obtained via k -means, and its ability to explain downstream fairness is evaluated through correlation with Worst-Group Accuracy (WGA) measured on the true sensitive groups (*Sex* × *Race*).

Table 3 reports the correlations obtained in the unsupervised setting. Across all ACS datasets, FairEntropy remains strongly and significantly correlated with WGA, with an average correlation up to $\rho_p = 0.84$. We emphasize that this alignment critically depends on the clustering granularity, as discussed in Appendix 11.3, and degrades for both overly coarse and overly fine partitions.

While no demographic information is used to compute FairEntropy, these results indicate that unsupervised FairEntropy captures structural properties of the embedding space that are highly predictive of downstream worst-group disparities on the ACS benchmarks.

Dataset	Pearson ρ_p	Spearman ρ_s	Kendall τ
ACSIIncome	0.82* _{±0.003}	0.81* _{±0.052}	0.64* _{±0.038}
ACSPublicCoverage	0.89* _{±0.003}	0.89* _{±0.069}	0.76* _{±0.072}
ACSEmployment	0.80* _{±0.005}	0.74* _{±0.066}	0.60* _{±0.067}
Average	0.84	0.81	0.67

Table 3: Correlation between FairEntropy scores and WGA across ACS datasets without demographics. 4 groups based on k -means clustering. * indicates that all of the 3 seeds used have a statistically significant correlations ($p < 0.05$)

Figure 7 further illustrates this behavior across datasets, showing that embedders with higher FairEntropy systematically exhibit higher WGA, despite the absence of demographic supervision.

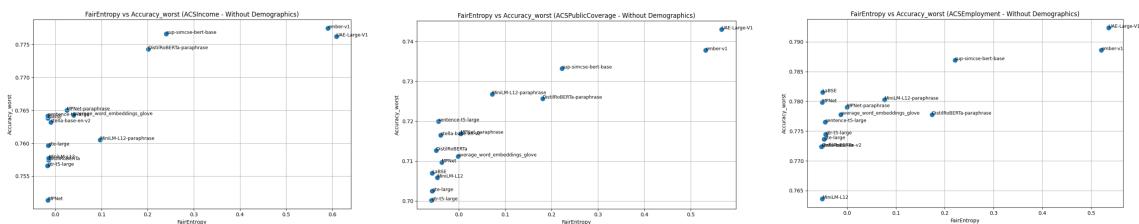


Figure 7: Correlation between FairEntropy score and WGA across ACS datasets without demographics: (left) ACSIIncome, (center) ACSPublicCoverage, and (right) ACSEmployment using K-means with 2 clusters.

Following the analysis in Appendix 11.3, we fix $K = 4$, as this value yields the most stable correlations on ACS benchmarks; notably, this granularity happens to align with the four sensitive

subgroups used for evaluation (*white men, non-white men, white women, non-white women*), but no demographic information is used in the computation of FairEntropy. Also, Worst-Group Accuracy (WGA) is consistently evaluated on the true sensitive groups defined by the Cartesian product of *Sex* and *Race*.

Interestingly, the correlations obtained in this unsupervised setting are on average comparable to, and in some cases slightly higher than, those observed when sensitive attributes are explicitly available; while this may appear counterintuitive, we attribute this behavior to the fact that, on the ACS benchmarks, intermediate clustering granularities induce proxy partitions that capture a level of heterogeneity at which worst-group performance is largely driven by representation collapse rather than by highly fine-grained subgroup structure, allowing unsupervised FairEntropy to recover a signal closely aligned with WGA; this result should not be interpreted as unsupervised evaluation being strictly superior, but rather as evidence that, on these datasets, fairness-relevant disparities manifest at a granularity that can be effectively approximated by appropriately chosen clustering structures.

We further analyze the effect of the number of clusters K in Appendix 11.3, which report silhouette scores and FairEntropy–WGA correlations for a wide range of clustering granularities.

Overall, while supervised FairEntropy remains conceptually preferable when sensitive attributes are available, our results indicate that, on the ACS benchmarks, unsupervised FairEntropy provides a surprisingly strong proxy for downstream worst-group performance, highlighting both the strengths and the limitations of representation-level fairness analysis.

4.3 Applicability and limitations

Applicability. FairEntropy provides a simple and practical tool to rank and compare embedding models in terms of group-wise fairness, without relying on task labels and potentially without demographics. Its main purpose is to help practitioners select, among several candidate embedders, those that are less likely to disadvantage certain demographic groups due to representational limitations. Because FairEntropy is task-agnostic and fast to compute, it can be applied early in the model selection process, before downstream tasks or labels are known. Importantly, FairEntropy serves as a complementary fairness-oriented criterion, guiding model choice alongside other evaluation metrics. We propose a ranking of the embedders used in this study in the Appendix ??.

Limitations. FairEntropy is designed to assess representation-level properties and should be interpreted accordingly. It does not guarantee high downstream accuracy, nor does it capture all sources of unfairness in a predictive system. While the downstream task may remain unknown, one still needs prior knowledge of the general application domain in order to preprocess the data into a suitable textual form and compute FairEntropy scores (DARRIN et al., 2024). A fundamental limitation of FairEntropy is its dependence on an explicit or implicit population partition: whether defined via sensitive attributes or induced through clustering, the fairness signal it provides is conditional on the chosen grouping structure, and inappropriate granularity may lead to under- or over-estimation of representational disparities, particularly in complex or highly heterogeneous datasets. Like most fairness criteria in the literature, FairEntropy provides a partition-aware diagnostic whose effectiveness depends on selecting a grouping granularity that meaningfully captures population heterogeneity.

All empirical results are obtained on ACS benchmarks, and further validation is required to assess whether similar behaviors arise on datasets with different forms of heterogeneity or weaker global structure.

Overall, FairEntropy is best interpreted as a practical and flexible proxy: it does not provide absolute guarantees, but it offers valuable easy-to-compute guidance for selecting embedders with stronger fairness properties without relying on task labels and could also be used without access on the demographics.

5 Future Work

This work opens several directions for further investigation. First, extending the experimental evaluation to a broader set of ACS tasks would help assess the robustness of the observed correlations beyond the three benchmarks considered here.

A natural extension is to apply FairEntropy to other data modalities. In particular, evaluating it on purely textual datasets would be more directly aligned with the intended use of text embedders,

while extending the approach to vision datasets could help assess whether similar representation-level fairness signals emerge in image-based embedding spaces.

From a methodological perspective, future work could explore refinements of the unsupervised setting. Although the present results suggest that FairEntropy is largely robust to the choice of clustering granularity, alternative grouping strategies or regularization schemes could be investigated to better control the partitioning of the embedding space. For instance, one could augment the score with a term reflecting inter-cluster geometric relationships, such as cluster separation or overlap, in order to explicitly account for global structure beyond worst-group uncertainty.

Finally, further work should study how FairEntropy relates to a wider range of fairness metrics beyond Worst-Group Accuracy, in order to clarify its scope, limitations, and role as a complementary diagnostic tool within the broader fairness evaluation landscape.

These directions represent natural continuations of this work, but could not be fully explored within the limited timeframe of the internship.

6 Technical Aspects

This work also involved significant technical challenges, both in terms of implementation and infrastructure. I summarize here the main aspects of my setup and contributions.

Codebase and Adaptations. I started from the official code of the *Information Sufficiency (IS) score* GitHub and adapted it for fairness evaluation and for FairEntropy. I modified parts of the pipeline to obtain the outputs I needed and to enable parallel execution. In the end, I wrote around 1,000 lines of Python code for the final pipeline, complemented by several hundred lines of Bash scripts to automate runs on SLURM.

Computational Infrastructure. I used Compute Canada clusters and relied on SLURM job scheduling (`sbatch`) to run my experiments. Most jobs were executed on the Nibi cluster, but I had to move between Narval, Béluga, Tamia, and Nibi due to availability and technical issues. I parallelized the code at the job level: for instance, each calculation of the form $\text{FairEntropy}(U)$ was submitted as an independent job.

Data and Models. In total, I worked with about 64 GB of pre-trained embedding models. For each dataset, the embeddings represented roughly 15 GB of data. The original datasets contained several million rows, but I subsampled them to 100,000 random examples to keep the experiments tractable.

Runtime and Scalability. Running the FairEntropy pipeline on one dataset with 17 embedding models typically required around 30 minutes on Compute Canada’s GPU on SLURM, reflecting the efficiency of absolute, representation-level fairness diagnostics. In contrast, the FairIS pipeline, which relies on pairwise and purely relative comparisons between embedders, required around 10 hours under the same experimental conditions, highlighting the substantially higher computational cost of such approaches.

7 Main Learnings from the Internship

Over the course of this internship, I have gained both technical and personal skills that I believe will be valuable for my future. Working on fairness in AI allowed me to explore a fascinating and necessary domain to better understand the societal stakes of AI. Along the way, I also experienced firsthand that research often leads to setbacks and dead ends, and I learned how important it is to pick myself up after failure and keep moving forward.

On a practical level, I became more proficient with computational tools such as SLURM, which enabled me to organize and run large-scale experiments efficiently. I also strengthened my use of L^AT_EX and improved at conducting structured literature reviews, which helped me connect diverse ideas into a coherent research direction. Finally, this internship gave me the chance to refine my communication skills, particularly when presenting in front of professors and researchers from different backgrounds—for example during the joint UQAM–ÉTS workshop at the end of the project where I presented my work. These experiences gave my the opportunity to discuss with researchers from diverse fields and backgrounds internationally.

8 Ethics

This research is centered on fairness, which is itself a core aspect of ethics. Working on such a subject highlighted for me how crucial it is to develop ethical dimensions of AI in parallel with technical advances. Fairness research does not only improve the trustworthiness of models, but also provides a foundation for responsible deployment in society. By addressing these questions during my internship, I gained awareness of the responsibility that comes with designing and evaluating AI systems.

9 Acknowledgements

I would like to express my deepest gratitude to Ulrich Aïvodji and Pablo Piantanida for their guidance, advice, and for giving me the opportunity to carry out this great internship. Their support has been essential to my progress and learning throughout this project. I am also grateful to the PhD students at the lab, in particular Patrik Kenfack, for his insightful advice, and to all others who made every day at the lab enjoyable with their presence. Finally, I warmly thank my family and friends for their constant encouragement and support during this internship.

10 Conclusion

Throughout this internship, I had the opportunity to explore in depth one of the central challenges of modern AI: evaluating fairness in foundation embedders under severe informational constraints, including settings where labels or demographic attributes may be unavailable. By building on the Information Sufficiency (IS) framework, I first explored a group-conditional extension, *FairIS*, and showed empirically that it does not provide a reliable fairness ranking in our setting. This negative result motivated the design of *FairEntropy*, a simpler and easy-to-compute entropy-based metric that operates purely at the representation level. *FairEntropy* correlates strongly with Worst-Group Accuracy (WGA) on all of the three ACS tasks I tested with demographics and exhibits substantial and consistent alignment with WGA in the unsupervised clustering setting when an appropriate partitioning granularity is used. These results show that fairness-relevant information can be extracted from embeddings even under severe informational constraints, and I am proud to have contributed a clearer understanding of both the limitations of IS-based approaches and the potential of entropy-based structural metrics for fairness assessment.

On a personal level, this internship has been extremely formative. I learned to navigate the full research cycle, from a thorough literature review to theoretical reformulation, metric design, and large-scale experimentation. Implementing the *FairIS* and *FairEntropy* frameworks required me to develop new technical skills, particularly in adapting complex codebases, managing large-scale computational pipelines on SLURM, and handling significant amounts of data and models. Beyond technical expertise, I also strengthened my ability to connect abstract theoretical concepts with empirical validation, which I see as a crucial step in becoming a well-rounded professional in AI.

This experience also highlighted the importance of resilience and collaboration in research. Many experiments did not work on the first attempt, and I learned to view these setbacks as part of the scientific process rather than failures. Discussions with my supervisors and other researchers at the lab were invaluable for refining my ideas and keeping the project on track. I greatly appreciated the stimulating environment of the International Laboratory on Learning Systems in Montréal, where I could exchange with people from diverse backgrounds, countries and research areas.

11 Appendix

11.1 Worst-Group Bayes Risk and Groupwise Deficiency: A Simple Bound

We denote by $R_E(z)$ the Bayes risk of an embedding model E restricted to group z , and by $\Delta R(E) = \max_{z \in \mathcal{Z}} R_E(z)$ its worst-group Bayes risk. All notions of Bayes risk and groupwise deficiency $\delta_z(U \rightarrow V)$ are defined above in the paper.

Proposition 2. *For any two embedding models U, V ,*

$$\Delta R(U) - \Delta R(V) \leq \max_{z \in \mathcal{Z}} \delta_z(U \rightarrow V).$$

Proof. For each $z \in \mathcal{Z}$, we have the following inequality (DARRIN et al., 2024)

$$R_U(z) - R_V(z) \leq \delta_z(U \rightarrow V).$$

Therefore,

$$\begin{aligned} \Delta R(U) - \Delta R(V) &= \max_{z \in \mathcal{Z}} R_U(z) - \max_{z \in \mathcal{Z}} R_V(z) \\ &\leq \max_{z \in \mathcal{Z}} (R_U(z) - R_V(z)) \\ &\leq \max_{z \in \mathcal{Z}} \delta_z(U \rightarrow V). \end{aligned}$$

□

11.2 Embedding models

Model	Dimension	Max tokens
Baseline Models		
average_word_embeddings_glove.6B	300	N/A
General Self-Supervised Models		
all-MiniLM-L12-v2	384	512
stella-base-en-v2	768	512
all-distilroberta-v1	768	512
all-mpnet-base-v2	768	512
ember-v1	1024	512
gte-large	1024	512
Paraphrase Supervised Models		
MiniLM-L12-paraphrase-v2	384	512
MiniLM-L12-paraphrase-multilingual-v2	384	512
DistilRoBERTa-paraphrase-v1	768	512
MPNet-paraphrase-v2	768	512
NLI & STS Supervised Models		
bert-base-nli	768	128
distilbert-base-nli-stsb	768	128
sup-simcse-bert-base	768	512
UAE-Large-V1	1024	512
Retrieval Models		
LaBSE	768	512
gtr-t5-large	768	512
e5-large-v2	1024	512
sentence-t5-large	1024	512

Table 4: Summary of embedding models with dimension size and maximum input length.

11.3 Impact of the value of K – Unsupervised settings

Figure 8a shows that the global silhouette score consistently decreases as the number of clusters K increases, indicating progressively weaker geometric separation of the unsupervised clusters. Figure 8b further reveals that the correlation between worst-group FairEntropy and WGA is sensitive to the choice of K : very small values ($K = 1$ or $K = 2$) yield weaker and less stable rank correlations, while correlations peak for intermediate values of K , typically when K matches or slightly exceeds the number of true sensitive groups. For larger clusterings, the correlation gradually degrades and becomes more variable across seeds, suggesting that excessive partitioning introduces noise by fragmenting the representation space into overly small and unstable clusters. As illustrated in Figure 8c, increasing K also induces a systematic rightward shift of FairEntropy values, reflecting higher worst-cluster entropy due to finer partitions, while only partially preserving the relative ranking of embedders. The results are similar on the 2 other ACS datasets.

Overall, these results indicate that FairEntropy is most informative when the clustering granularity is commensurate with the underlying sensitive group structure, suggesting that selecting K on the order of the approximated number of latent subgroups provides a practical and robust operating regime, while avoiding both under-partitioning and excessive fragmentation that may reduce stability, especially on datasets with more complex group structures.

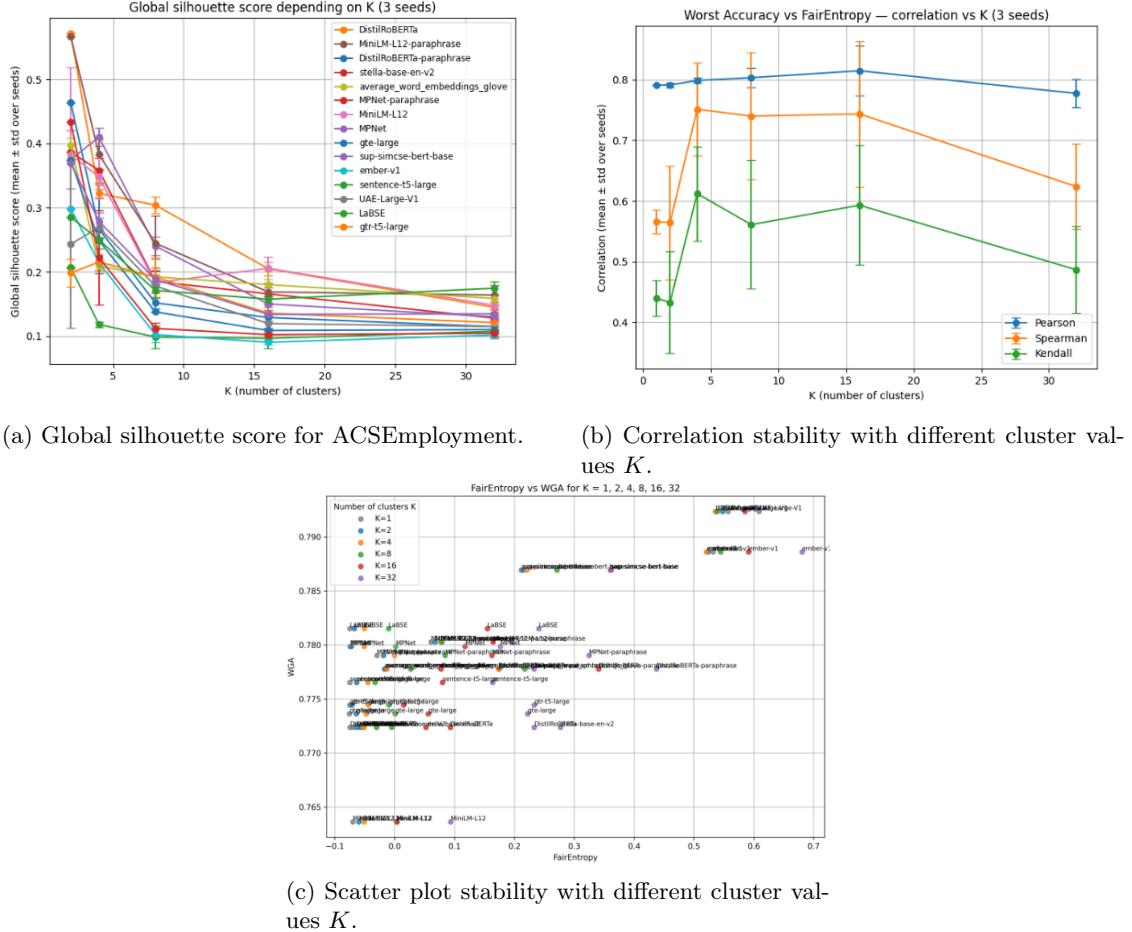


Figure 8: Global silhouette score, scatter plot and correlation for ACSEmployment as a function of K . Cluster quality decreases with larger K , but correlation with WGA remains stable

11.4 Final ranking of embedders based on the mean worst FairEntropy score across the 3 datasets.

Table 5 reports the final ranking of embedding models according to their average worst FairEntropy score across the 3 ACS datasets. Higher values indicate embeddings for which the minimum estimated entropy across sensitive groups is larger, suggesting that no group concentrates representations into a low-entropy region of the latent space, reducing worst-case disparities in representational uncertainty.

Note that FairEntropy values can be negative, as entropy is estimated from a continuous density modeled with a Gaussian Mixture Model, which does not impose non-negativity constraints on the resulting differential entropy estimates.

Table 5: Final ranking of embedders based on the mean worst IS score across datasets.

Rank	Embedder	Mean FairEntropy
1	UAE-Large-V1	0.560596
2	ember-v1	0.538469
3	sup-simcse-bert-base	0.217774
4	DistilRoBERTa-paraphrase	0.176930
5	MiniLM-L12-paraphrase	0.070956
6	average_word_embeddings_glove	0.002346
7	MPNet-paraphrase	-0.009850
8	stella-base-en-v2	-0.047800
9	MiniLM-L12	-0.049827
10	LaBSE	-0.053059
11	DistilRoBERTa	-0.053342
12	MPNet	-0.053430
13	gtr-t5-large	-0.053454
14	sentence-t5-large	-0.053638
15	gte-large	-0.054079

11.5 FairIS: Group-Conditional Information Sufficiency

In this section, let’s explore how we built FairIS and the associated results.

Following Proposition 1, which provides a group-conditional upper bound on excess risk via deficiency, a natural extension is to define a group-conditional Information Sufficiency score by restricting IS estimates to sensitive groups. While theoretically well bounded, we show in this appendix that this direct extension—referred to as *FairIS*—does not provide a sufficiently discriminative fairness proxy in our setting.

Fairness IS score definition. Since estimating the true deficiency from samples is intractable, we rely on the IS score as a practical proxy. To adapt it to fairness, we extend the definition to the group-wise setting. For each embedder E and group z , we compute

$$IS_z(E) = \text{median}\{IS_z(E \rightarrow E') \mid E'\},$$

where $IS_z(E \rightarrow E')$ denotes the information sufficiency estimated by restricting both source and target embeddings to the sensitive group z . The fairness score of embedder E is then defined by the worst-performing group:

$$\text{FairIS}(E) = \min_{z \in \mathcal{Z}} IS_z(E).$$

This construction mirrors the worst-group perspective underlying WGA: an embedder is considered more fair if even its least informative group retains sufficient information relative to others. The full procedure is summarized in Algorithm 5.

Algorithm 5 Computation of FairIS (group-conditional IS)

```

1: Input: Dataset  $X = \{(x_n, z_n)\}_{n=1}^N$  with sensitive groups  $\mathcal{Z}$ , embedders  $\mathcal{E} = \{E_1, \dots, E_m\}$ 
2: for each embedder  $E_i \in \mathcal{E}$  do
3:   for each group  $z \in \mathcal{Z}$  do
4:      $X_z \leftarrow \{x_n : z_n = z\}$ 
5:     for each  $E_j \in \mathcal{E}$ ,  $j \neq i$  do
6:        $U \leftarrow E_i(X_z)$ ,  $V \leftarrow E_j(X_z)$ 
7:       Estimate  $IS_z(E_i \rightarrow E_j)$ 
8:     end for
9:      $IS_z(E_i) \leftarrow \text{median}(\{IS_z(E_i \rightarrow E_j)\})$ 
10:   end for
11:    $\text{FairIS}(E_i) \leftarrow \min_{z \in \mathcal{Z}} \frac{IS_z(E_i)}{d_{E_i}}$ , where  $d_{E_i}$  denotes the embedding dimension of  $E_i$ .
12: end for
13: Output:  $\{\text{FairIS}(E_i)\}_{i=1}^m$ 

```

Empirical behavior. Figure 9 reports the relationship between FairIS and worst-group accuracy on ACS datasets. FairIS exhibits little to no alignment with WGA: embedders with similar FairIS values can display substantially different worst-group performance, and the resulting ordering fails to reflect downstream group-wise fairness. The results are similar on the 2 other ACS datasets.

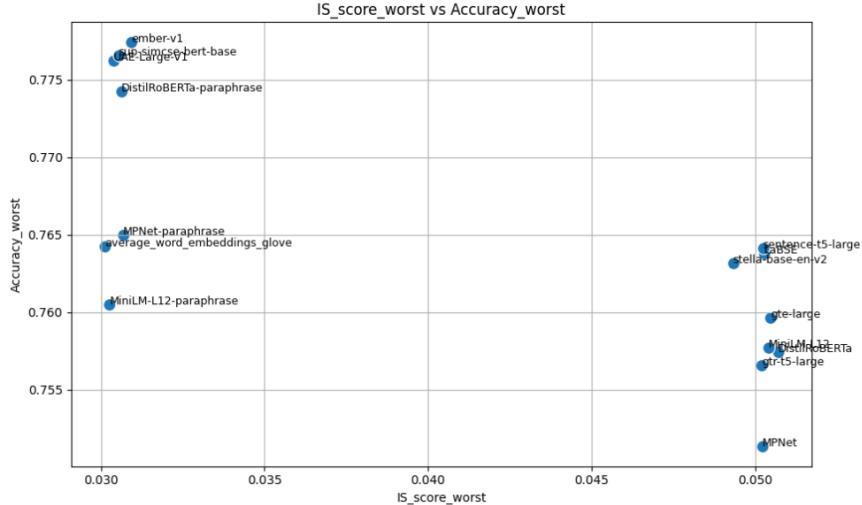


Figure 9: FairIS versus worst-group accuracy on ACSEmployment. The score exhibits weak or no correlation with WGA, indicating that group-conditional IS does not yield a meaningful fairness ranking.

Why FairIS fails in this setting. This behavior stems from two complementary issues inherent to Information Sufficiency under group conditioning. First, IS is intrinsically *relative*: it quantifies how informative one embedding is with respect to another over the same data distribution. When restricted to a single sensitive group, IS primarily reflects inter-embedder similarity within that group, rather than the absolute quality, diversity, or separability of the corresponding representations.

Second, conditioning on groups substantially reduces distributional diversity, causing IS scores to saturate across embedders. As many models exhibit similar levels of mutual informativeness within narrowly defined groups, FairIS lacks discriminative power. The subsequent worst-group aggregation further amplifies this non-smooth behavior, collapsing embedders into broad equivalence classes and preventing the emergence of a meaningful ordering.

As a result, FairIS is largely insensitive to representation collapse: an embedder may receive a high score even when group-specific representations are highly concentrated or degenerate, provided that other embedders behave similarly. In contrast, FairEntropy directly measures group-wise representational variability, making it sensitive to collapse and over-compression effects while remaining absolute rather than relative. This distinction explains the strong empirical alignment between FairEntropy and worst-group accuracy observed in the main experiments, and further motivates our focus on entropy-based structural metrics for fairness assessment. Moreover, FairEntropy is significantly faster to compute due to its non-pairwise formulation.

References

- Kenfack, P. J., Kahou, S. E., & Aïvodji, U. (2024). A survey on fairness without demographics. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=3HE4vPNIfX>
- Chehbouni, K., Roshan, M., Ma, E., Wei, F., Taik, A., Cheung, J., & Farnadi, G. From representational harms to quality-of-service harms: A case study on llama 2 safety safeguards (L.-W. Ku, A. Martins, & V. Srikumar, Eds.). In: *Findings of the association for computational linguistics: Acl 2024* (L.-W. Ku, A. Martins, & V. Srikumar, Eds.). Ed. by Ku, L.-W., Martins, A., & Srikumar, V. Bangkok, Thailand: Association for Computational Linguistics, 2024, August, 15694–15710. <https://doi.org/10.18653/v1/2024.findings-acl.927>
- DARRIN, M., Formont, P., Ayed, I. B., Cheung, J. C., & Piantanida, P. When is an embedding model more promising than another? In: In *The thirty-eighth annual conference on neural information processing systems*. 2024. <https://openreview.net/forum?id=VqFz7iTGcl>

- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6). <https://doi.org/10.1145/3457607>
- Verma, S., & Rubin, J. Fairness definitions explained. In: In *Proceedings of the international workshop on software fairness*. FairWare '18. Gothenburg, Sweden: Association for Computing Machinery, 2018, 1–7. ISBN: 9781450357463. <https://doi.org/10.1145/3194770.3194776>
- Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7). <https://doi.org/10.1145/3616865>
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. Learning fair representations. In: In *Proceedings of the 30th international conference on international conference on machine learning - volume 28*. ICML'13. Atlanta, GA, USA: JMLR.org, 2013, III–325–III–333.
- Edwards, H., & Storkey, A. J. (2015). Censoring representations with an adversary. *CoRR*, *abs/1511.05897*. <https://api.semanticscholar.org/CorpusID:4986726>
- Madras, D., Creager, E., Pitassi, T., & Zemel, R. S. (2018). Learning adversarially fair and transferable representations. *ArXiv*, *abs/1802.06309*. <https://api.semanticscholar.org/CorpusID:3419504>
- Balunovic, M., Ruoss, A., & Vechev, M. T. (2021). Fair normalizing flows. *ArXiv*, *abs/2106.05937*. <https://api.semanticscholar.org/CorpusID:235390444>
- Sagawa*, S., Koh*, P. W., Hashimoto, T. B., & Liang, P. Distributionally robust neural networks. In: In *International conference on learning representations*. 2020. <https://openreview.net/forum?id=ryxGuJrFvS>
- Hashimoto, T. B., Srivastava, M., Namkoong, H., & Liang, P. (2018). Fairness without demographics in repeated loss minimization. *ArXiv*, *abs/1806.08010*. <https://api.semanticscholar.org/CorpusID:49343170>
- Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., & Chi, E. H. Fairness without demographics through adversarially reweighted learning. In: In *Proceedings of the 34th international conference on neural information processing systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.
- Yan, S., Kao, H.-t., & Ferrara, E. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In: In *Proceedings of the 29th acm international conference on information & knowledge management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, 2020, 1715–1724. ISBN: 9781450368599. <https://doi.org/10.1145/3340531.3411980>
- Cam, L. L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.*, 35(4), 1419–1455. <http://dml.mathdoc.fr/item/1177700372>
- Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2022). Retiring adult: New datasets for fair machine learning. <https://arxiv.org/abs/2108.04884>
- Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., & Sontag, D. (2023). Tabllm: Few-shot classification of tabular data with large language models. <https://arxiv.org/abs/2210.10723>
- Kearns, M., Neel, S., Roth, A., & Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness (J. Dy & A. Krause, Eds.). In: *Proceedings of the 35th international conference on machine learning* (J. Dy & A. Krause, Eds.). Ed. by Dy, J., & Krause, A. 80. Proceedings of Machine Learning Research. PMLR, 2018, 2564–2572. <https://proceedings.mlr.press/v80/kearns18a.html>