Analyzing Gender Representation in Multilingual Models

Anonymous ACL submission

Abstract

Multilingual language models were shown to allow for nontrivial transfer across scripts and languages. In this work, we study the structure of the internal representations that enable this transfer. We focus on the representations of gender distinctions as a practical case study, and examine the extent to which the gender concept is encoded in *shared subspaces* across different languages. Our analysis shows that gender representations consist of several prominent components that are shared across languages, alongside language-specific components. The existence of language-independent and language-specific components provides an explanation for an intriguing empirical observation we make: while gender classification transfers well across languages, bias mitigation interventions trained on a single language do not transfer easily to others.

1 Introduction

Pretrained models of contextualized representations (Peters et al., 2018; Devlin et al., 2018; Liu et al., 2020) are known in their ability to capture both explicit and implicit information during training. A special case of these models are multilingual models (Devlin et al., 2018; Conneau et al., 2020), which are pretrianed with texts in multiple languages. These models were shown to induce the emergence of similar representations in different languages, a phenomenon that was put to use for transfer between languages in end-tasks (Pires et al., 2019; Muller et al., 2020; Gonen et al., 2020). However, the underlying mechanism is still not clear, and we do not know yet the full extent to which the representations of these models share information across languages.

The rise of pretrained models has been accompanied with growing concern regarding sensitive information they might encode, e.g. gender or ethnic distinctions. Pre-trained language models were shown to be sensitive to gender information, both when it is explicitly stated in texts, as well as when it can be inferred from implicit information (Zhao et al., 2019; May et al., 2019). We still lack a complete understanding of what the model captures, and the ways to control and change the information in this context as well.

In this work,[1] we aim to shed light on the way human-interpretable concepts, such as gender, are represented by multilingual models, and whether they are encoded in a language-dependant way. In a series of experiments, we uncover a seemingly surprising finding: gender-identification ability is highly transferable across languages (section 4.1) but neutralizing gender identification is not (section 4.2). While these two findings may seem contradictory at first glance, this is explained by several levels of gender marking: both cross-lingual and language-specific (section 5).

We start our analysis by training gender classifiers and examining their ability to transfer across languages. We then proceed to identifying "gender subspaces" — subspaces that encode gender — in each language, with the goal of understanding which information is language-specific, and which is shared across languages. Following recent work on linear interventions (Ravfogel et al., 2020; Elazar et al., 2021; Ravfogel et al., 2021), we take an "amnesic" approach: we study the extent to which **neutralizing** the gender subspace in one language interferes with gender prediction in another language. Finally, we analyze the similarity in the gender-encoding components across languages.

We find that while linear probes for gender transfer well between languages — that is, a gender classifier that is trained on one language predicts gender well in another language, the linear bias mitigation procedure we employ fails to transfer. A deeper analysis reveals a fine-grained organization of the gender-encoding subspaces across languages:

[1] We will make the code available upon publication.

they are spanned by a few main directions, which are largely similar across languages; but in addition to these directions, there are other directions that are language-specific. The existence of several similar directions explains the high degree of transferability of linear gender classifiers across languages, while the existence of a large amount of language-specific information explains the inability to efficiently mitigate bias in one language based on another language's representation.

2 Related Work

Multilingual Representation Analysis Pires et al. (2019) begin a line of work that studies mBERT's representations and capabilities. In their work, they inspect the model's zero-shot transfer abilities using different probing experiments, and propose a way to map sentence representations in different languages, with some success. Karthikeyan et al. (2020) further analyze the properties that affect zero shot transfer of bilingual BERTs. Wu and Dredze (2019) perform transfer learning from English to 38 languages, on 5 different downstream tasks and report good results. Wang et al. (2019) learn alignment between contextualized representations, and use it for zero shot transfer. Dufter and Schütze (2020) make an attempt to control different aspects of mBERT and identify those that contribute the most to its transfer ability.

Beyond focusing on zero-shot transfer abilities, an additional line of work studies the representations of mBERT and the information it stores. Using hierarchical clustering based on the CCA similarity scores between languages, Singh et al. (2019) are able to construct a tree structure that faithfully describes relations between languages. Chi et al. (2020) learn a linear syntax-subspace in mBERT, and point out to syntactic regulartieis in the representations that transfer across languages. In Cao et al. (2020), the authors define the notion of *contextual* word alignment. They design a fine-tuning loss for improving alignments and show that they are able to improve zero-shot transfer after this alignment-based fine-tuning. In Libovický et al. (2019), the authors assume that mBERT's representations have a language-neutral component, and a language-specific component and provide an experimental setting to partially support this assumption. Finally, in Gonen et al. (2020), the authors propose an explicit *decomposition* of the representations to language-encoding and language-neutral components, and also demonstrate that implicit word-level translations can be easily distilled from the model when exposed to the proper stimuli.

Unlike previous works, we pay attention specifically to how gender is manifested in the representations, as a case study for the analysis of a concrete societal property. We do that by focusing on the information included in the representations themselves, rather than on downstream tasks.

Gender Representation in Multilingual Models To the best of our knowledge, no previous work focuses on the way gender is represented in multilingual models and the extent to which such representations are shared across languages.

Some work has been done on identifying and mitigating gender bias in languages other than English. Gonen et al. (2019) identify and debias a new type of gender bias, unique to gender-marking languages. Williams et al. (2021) look at the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. In Zmigrod et al. (2019), the authors suggest a method for converting between masculine-inflected and feminine-inflected sentences in morphologically rich languages, and use them for counterfactual data augmentation in order to reduce gender stereotyping.

Zhao et al. (2020) analyze gender bias in multilingual word embeddings, and evaluate it intrinsically and extrinsically. They point to several factors that influence the gender bias in multilingual embeddings, among which are the pretrained monolingual word embeddings, and the alignment method used. Additionally, Liang et al. (2020) focus on contextualized embeddings, analyze the gender representation in BERT, and also put efforts into English-Chinese cross lingual debiasing. Finally, Bansal et al. (2021) focus on Indian languages when debiasing multilingual embeddings.

3 Datasets and Multilingual Representations

For our experiments we use the BiosBias Dataset (De-Arteaga et al., 2019), the Multilingual BiosBias Dataset (Zhao et al., 2020) and the multilingual BERT model (mBERT, (Devlin et al., 2018)) as detailed below.

Multilingual Gender Data. De-Arteaga et al. (2019) collected the English BiosBias dataset, a

set of short-biographies written in third person, and annotated by perceived gender. They have demonstrated that profession classifiers trained on this dataset condition on the gender concept, resulting in fairness issues. Zhao et al. (2020) evaluate the bias in cross-lingual transfer settings, for which they created the Multilingual BiosBias (MLBs) Dataset which contains a similar set of biographies in three additional languages: French, Spanish and German.

For our experiments we use both datasets to have English, Spanish and French data. These are not available online, so we used the scripts the authors provide for crawling the dataset ourselves.[2] To avoid noisy results, we filter out examples of professions with less than 500 occurrences. Table 1 describes the statistics of the dataset in all languages.

	examples	female	male	majority	# prof
EN	255682	118344	137338	53.71	28
FR	42773	12196	30577	71.49	19
ES	46931	12867	34064	72.58	27

Table 1: Statistics of the MLBs dataset.

Multilingual Representations. To study the representation of the gender concept in a multilingual setting, we use multilingual BERT (mBERT) (Devlin et al., 2018). For each example in the dataset, we extract its representation from mBERT by averaging the representations in context of all the tokens in the paragraph.

4 Gender Representation across Language

4.1 Transfer of Gender Probes

As a first step in understanding gender representation in multilingual models, we start with a basic experiment that aims to evaluate the extent to which gender is represented similarly across languages.

We train a linear classifier for gender classification in a SOURCE language, and use it as is to predict gender in a TARGET language. The training is done over the mBERT representations of the training examples.

The results, presented in Table 2, indicate that gender classifiers transfer very well across languages, with only a slight degradation in performance when applied in a different language. For example, the accuracy of the English gender classifier is 99.27%, but when the French or Spanish classifiers are used to predict gender in English data, the accuracy is 98.10% and 97.29%, respectively. The same trend is observed for the French and Spanish datasets. These results suggest that gender information is linearly accessible in mBERT representation and is shared between languages.

	EN classifier	FR classifier	ES classifier
EN data	**99.27**	98.10	97.29
FR data	95.97	**97.50**	94.61
ES data	84.04	84.10	**85.97**

Table 2: Accuracy of gender classification across languages with linear classifiers trained on average representations. Rows represent the language of the prediction data, columns represent the language in which the classifier was trained.

4.2 Cross-lingual Linear Bias Mitigation

The experiment described above suggests some gender components are shared between languages. As bias mitigation techniques focus on the *removal* of bias information, a natural question that arises is whether mitigation efforts focused on one language would transfer to another. This question is important for two reasons. First, if possible, this has a potential practical utility – e.g., enabling bias mitigation in low-resource languages, for which training data is scarce. Second, the degree of success in transfer of bias mitigation efforts is a complementary way to assess whether the representation of gender is indeed multilingual.

Previous experiments on removing the gender concept from neural representations show encouraging results in-language for English. These are done using INLP (Ravfogel et al., 2020), an existing approach for the identification and neutralization of "concept subsapces", e.g. the gender concept. In these experiments, Ravfogel et al. (2020) show they manage to neutralize the ability of linear probes to recover gender information from the representations. In light of the above results that show high quality transfer of gender classifiers **across** languages, we leverage the INLP method, and attempt to *remove* gender information from the representations **across** languages.

Iterative Null-space Projection (INLP) INLP (Ravfogel et al., 2020) aims to remove linearly-

[2]The German portion we were able to extract was too small, so we decided to aviod experimenting with it.

decodable information from vector representations. Given a dataset of representations X (in our case, mBERT representations) and annotations Z for the information to be removed (gender) the method renders Z linearly unpredictable from X. It does so by iteratively training linear predictors $w_1, \ldots, w_n$ of Z, calculating the projection matrix onto their nullspace $P_N := P_N(w_1), \ldots, P_N(w_n)$, and transforming $X \leftarrow P_N X$. By the nullsapce definition, this guarantees $w_i P_N X = 0, \forall w_i$, i.e., the features that w_i uses for gender prediction are neutralized. Note that the guarantee is only with respect to linear separation.

While the nullsapce $N(w_1, \ldots, w_n)$ is a subspace in which Z is not linearly predictable, the complement rowspace $R(w_1, \ldots, w_n)$ is a subspace of the representation space X that corresponds to the property Z. In our case, this subspace is the *gender subspace*. As part of the analysis in this work, we utilize INLP in two complementary ways: (1) we use the *null-space* projection matrix P_N to zero out the gender subspace, in order to render the representations gender-neutral[3], this projection is onto the **gender-neutral subspace**; and (2) we use the *rowspace* projection matrix $P_R = I - P_N$ to project mBERT representations onto the **gender subspace**, keeping only the parts that are useful for gender prediction.

Method We start by training INLP in one language (En, Fr, Es) and identifying the complementing subspaces: the gender-neutral subspace – *nullspace*, and the gender subspace – *rowspace* (for later use, see Section 5). We then neutralize that subspace in *another* language. Finally, we examine the influence of this intervention and asses the effect of gender information reduction.

We run INLP with the objective of identifying the gender, with SGD classifiers (using SKlearn[4]) for 100 iterations. We use the average representations (averaging over the representations in context of all tokens) of the training paragraphs.

Results Tables 3 and 4 depict the results of gender and profession prediction in each language (rows) before and after applying INLP (each column stands for a different language for training INLP). We get that in-language, the accuracy of gender prediction drops to majority after applying INLP, while profession classification is only slightly hurt. For example, for English we get gender prediction accuracy of 53.7 compared to 99.3 before using INLP, and profession prediction accuracy of 78.1 compared to 79.9 before INLP. However, across languages, there is virtually no effect, both for gender prediction and profession prediction. For example, English gender and profession predictions drop from 99.3 to 98.1 and from 79.9 to 79.5, respectively, after applying Spanish INLP.

	before	EN INLP	FR INLP	ES INLP
EN	99.3	53.7	97.6	98.1
FR	97.8	95.1	71.4	94.9
ES	85.7	82.8	82.6	72.5

Table 3: Gender prediction before and after applying INLP. Rows are the language in which we predict, columns are the languages in which we train INLP. Using 100 iterations of INLP in each language.

	before	EN INLP	FR INLP	ES INLP
EN	79.9	78.1	79.2	79.5
FR	73.0	72.4	68.2	72.4
ES	57.8	57.1	57.3	51.8

Table 4: Profession prediction before and after applying INLP. Rows are the language in which we predict, columns are the languages in which we train INLP. Using 100 iterations of INLP in each language.

Interestingly, the largest drops in performance of profession classification due to application of INLP are in-language. This can be explained by the inherent correlations between gender and profession signals – removing gender information hurts the ability to predict the profession in the same language. This is not the case across-language since, as seen by the gender prediction results, gender information is not removed from the representations when applying INLP across languages.

5 Analyzing the Cross-linguality of Gender Representation

At first glance, the two results presented in Section 4 look contradicting: linear gender classification transfers well across languages while gender removal using INLP does not. In this section we provide a detailed analysis that accounts for this discrepancy: under this more fine-grained view, gender representation is neither shared between

[3] to the extent that gender is indeed encoded in a linear subspace, and that INLP finds this subspace.

[4] https://scikit-learn.org/stable/

languages nor unique per language, but is actually only partially shared between languages. This allows for some transferability, but prevents debiasing across languages.

To define the term "partial sharing" formally, we represent gender in each language as a collection of linear directions that together span the gender subspace of that language. This collection of directions can be identified using INLP – when training INLP in a specific language, we get a sequence of orthogonal linear classifiers that are able to predict gender with a decreasing level of accuracy, with the first classifier being the most accurate one. Together, these directions define the gender subspace of the language. This formulation allows us to more easily analyze the extent to which gender is similarly encoded across languages.

We hypothesize that the two aforementioned results are compatible because **some of these gender directions are shared between languages, while others are language-specific**. The shared directions allow high quality transfer of gender classification across languages, while the language-specific directions allow gender prediction even after applying INLP cross-lingually since they are not identified in the source language. In what follows, we devise two experiments to quantify this phenomenon.

5.1 Shared Gender Directions across Languages

High Level and Intuition In the following experiment we leverage the formulation of gender representation as a collection of many different directions in the space as well as the ability to project representations on the gender and gender-neutral subspaces, to analyze the relation between gender representations in the different languages. We are looking to answer the following question: are gender directions fully shared across languages, fully disjoint, or split – some are shared between languages and some are disjoint?

Intuitively, when projecting representations on the gender subspace, we expect all the information relevant to gender prediction to be kept in the projected representations. Similarly, when projecting representations on the gender-neutral subspace, we expect the opposite – that the projected representations will not include any gender-related information.

With this intuition we seek to determine the extent to which gender information is shared between languages by comparing their gender subspaces (with similar gender subspaces indicating high amount of shared information). To quantify the shared information, we carefully compare the different directions, taking their significance into account. This process is explained in detail below.

Method In this experiment, we make use of our projection mechanism as a way to control the information included or excluded from the representations. We compare the original mBERT representations of the training data before and after projecting them on the learned gender and gender-neutral subspaces of the different languages (see "Compared Representations" below). For each set of compared representations we perform PCA and look at the explained variance of each PCA direction (i.e., the ratio between its variance and the total variance of the data), from large to small – this tells us the variance in the representations. When comparing the explained variance before and after a projection, we are able to quantify the information that was lost by that projection.

We take English and French as our running example. We perform two projections subsequently and compare the representations before, in between and after the projections: the first projection is on the English gender subspace – this preserves the gender directions in English; The second projection is on the French gender-neutral subspace – this eliminates the gender directions in French. In case there are **no shared gender directions** between the two languages: the representations after the first projection encode gender information in English, and **no information is lost** when further eliminating French gender directions – we expect the plots after the first and the second projections to be identical. Conversely, we expect that **full sharing of gender directions** between the languages will result in **zero variance** after the two projections – the first projection keeps only English gender directions, and these are eliminated when eliminating the (same) French gender directions (by projecting on the French gender-neutral subspace).

Compared Representations We start by training INLP and obtaining a collection of 100 gender directions in each language (EN FR, ES), from the most prominent to the least prominent one. We use 100 dimensions regardless of what was needed for INLP to converge, so as to be consistent across

languages and avoid artifacts due to the number of dimensions. We compare different sets of representations as detailed below, for English vs. French, English vs. Spanish and French vs. Spanish (the explanation below is assuming English vs. French):

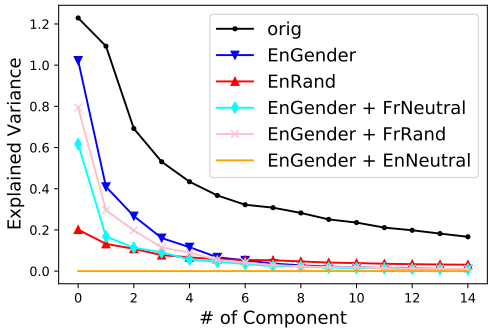
- ORIG: Original representations (in English).
- ENGENDER: ORIG projected on the English gender subspace (rowspace).
- ENRAND: ORIG projected with a random matrix with the same dimensions as the EnGender matrix (for comparison).
- ENGENDER+FRNEUTRAL: ENGENDER projected on the French gender-neutral subspace (nullspace).
- ENGENDER+FRRAND: ENGENDER projected on a random matrix with the same dimensions as the French gender-neutral matrix (for comparison).
- ENGENDER+ENNEUTRAL: ENGENDER projected on English gender-neutral subspace (nullspace).

Result Analysis The results for English vs. French, English vs. Spanish and French vs. Spanish are shown in Figure 1.

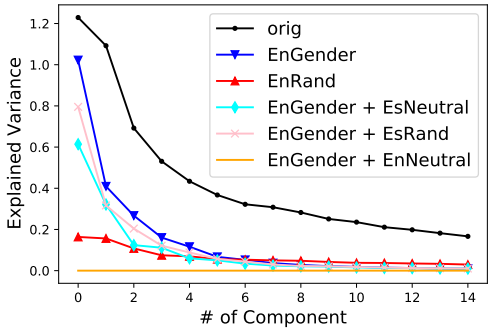
The plots support our initial hypothesis: indeed, we get that gender directions are shared between languages but only partially. Focusing on English vs. French, we can see that as expected, the curve of ENGENDER+FRNEUTRAL (cyan) is lower than that of ENGENDER (blue), implying that there are shared gender directions between English and French. Recall that projecting the representations on the English gender subspace (ENGENDER) keeps mainly English gender directions, and then projecting on French gender-neutral subspace (ENGENDER+FRNEUTRAL) removes French gender directions. If no directions are shared, this should result with similar values for both ENGENDER and ENGENDER+FRNEUTRAL. However, the sharing is only partial: if all directions are shared, we expect ENGENDER+FRNEUTRAL to be zero (similar to ENGENDER+ENNEUTRAL), which is not the case.

Controls The ENGENDER+FRRAND projections are intended as reference for ENGENDER+FRNEUTRAL. If there are shared gender directions between English and French, we expect the curve of ENGENDER+FRNEUTRAL to be lower than that of ENGENDER+FRRAND, since by projecting on the French gender-neutral subspace we are expected to lose more information than with a random projection with the same dimensions. In Figure 1a we can see that indeed the curve of ENGENDER+FRNEUTRAL (cyan) is lower than that of ENGENDER+FRRAND (pink), indicating that the loss of information is not due to random shared directions.

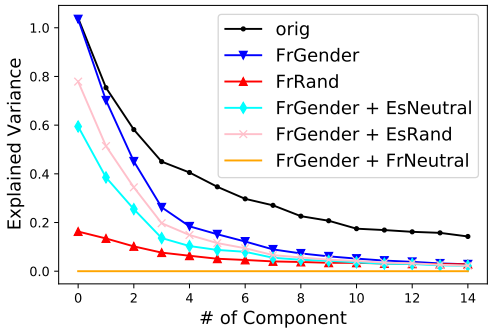
Note also that the curve of ENGENDER (blue)



(a) Explained variance of PCA of different representations, focusing on English and French.



(b) Explained variance of PCA of different representations, focusing on English and Spanish.



(c) Explained variance of PCA of different representations, focusing on French and Spanish.

Figure 1: Explained variance of PCA of different representations, for all three language pairs.

is significantly higher than that of EnRand (red). We hypothesize that this is due to the fact that gender is usually dominant in natural texts, especially in a dataset that includes information about individuals, as this one. Thus, keeping only gender information by projecting on the English gender subspace keeps a large portion of the information, compared to projecting on arbitrary directions of the same dimension.

Another sanity check is obtained by projecting EnGender on the English gender-neutral subspace (EnGender+EnNeutral), this should, by definition, result in a 0 line, which is indeed the case (orange).

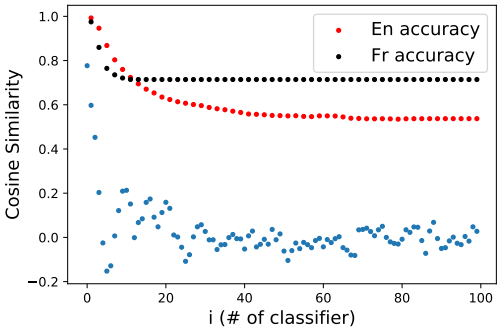
5.2 Similarities of Dominant Directions

In the previous section we established the hypothesis that some gender directions are shared between languages while others are language-specific. Now, we turn to perform a more fine-grained analysis where we look at the specific directions in the different languages.

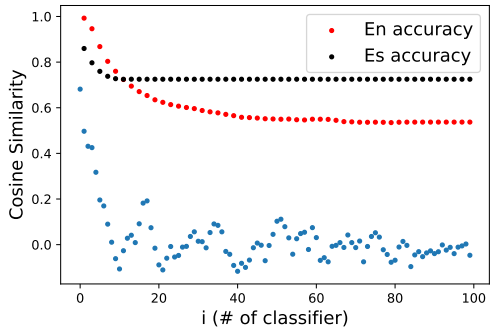
We look at the first 100 classifiers (trained during INLP) in two languages, and compute all pairwise cosine similarities between them (across language). This leads us to a surprising result – only the **first** classifiers in both languages are similar to each other, while the rest are not: we get that the 3 highest similarities are between the first En classifier and the first Fr classifier, between the second En classifier and the second Fr classifier, and between the third En classifier and the third Fr classifier, with values of 0.777, 0.597 and 0.453, respectively. The average absolute cosine similarity among all pairwise similarities of the first 100 classifiers in English and French is 0.037. Interestingly, the more dominant directions are those that are shared cross lingually, while the less predictive directions are those that are language specific.

Figure 2 depicts the similarities of the i^{th} classifiers for the two languages (English-French, English-Spanish and French-Spanish). We also plot the gender classification accuracy in-language for reference.

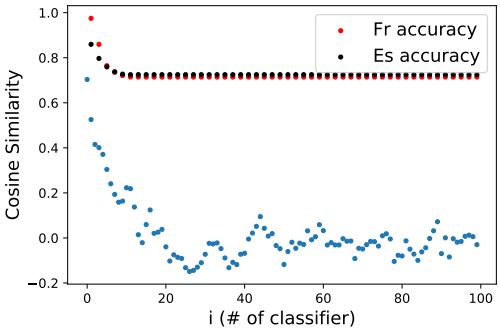
This result completes the picture and serves as an explanation for the extremely high quality transfer of gender classification across languages – the most dominant directions that represent gender in each languages are cross-lingual, which enables high accuracy in zero-shot transfer of linear gender classifiers across languages. However, less dominant gender directions are language specific, but are predictive enough so as to prevent gender neutralization across languages using INLP.



(a) Similarity between the i^{th} classifiers in English and French.



(b) Similarity between the i^{th} classifiers in English and Spanish.



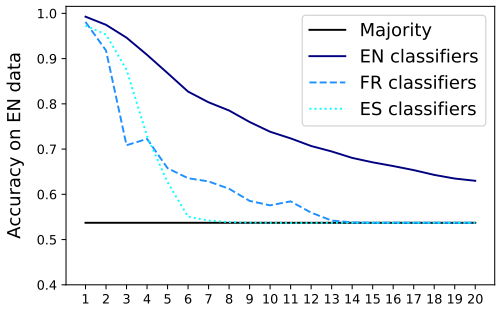
(c) Similarity between the i^{th} classifiers in French and Spanish.

Figure 2: Similarity between the i^{th} classifiers in all three language pairs. The gender classification accuracy in-language is added for reference.

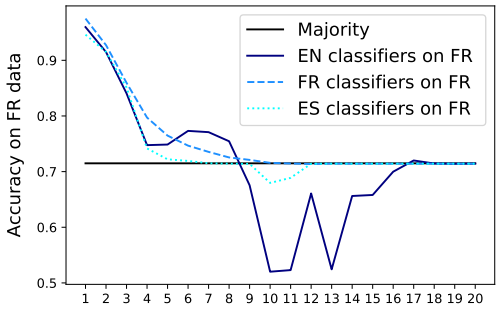
5.3 Accuracy across Language

Finally, we also look at the performance of each classifier (trained during INLP) across language. In Figure 3, we depict the gender prediction accuracy in-language and across-language. We consistently get that the performance of the first 2-3 classifiers

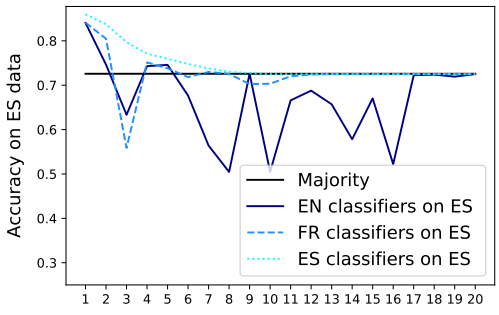
trained in-language and also across-language is relatively similar, with a significant divergence between in-language and cross-language trainings for the subsequent classifiers. This matches the results of the previous experiment which shows high similarity only between the first classifiers in the different languages.



(a) Gender prediction accuracy in English with the different classifiers in- and across-language.



(b) Gender prediction accuracy in French with the different classifiers in- and across-language.



(c) Gender prediction accuracy in Spanish with the different classifiers in- and across-language.

Figure 3: Gender prediction accuracy with the different classifiers in- and across-language.

6 Conclusion

As part of the efforts to better understand the underlying mechnism of multilingual modeling, we focus in this work on the way gender is represented across languages. We analyze and quantify the extent to which gender information is shared across English, French and Spanish.

We find that on the one hand, gender prediction transfers very well across languages: training a linear classifier on English data yields a high quality classifier for French and Spanish as well (true for all three languages in both directions). On the other hand, our attempt to neutralize gender information across languages using INLP, which was shown to work in English, was unsuccessful.

We show that these two results are compatible, and together they shed light on the structure of the representation space: we provide experimental evidence that the most salient directions are shared between languages (which enables good transfer of the classifiers), while others are unique per language (which interferes with gender removal across languages). The key observation is that a *single* "good" direction of the gender subspace in one language is enough for cross-lingual gender prediction transfer, while transfer of gender neutralization requires *all* directions to be shared, otherwise, the remaining ones can be used to recover gender information after the removal of the shared ones.

7 Ethical Considerations

Gender bias mitigation has attracted a lot of attention as a practical and socially important field of study. This paper contributes to this effort by studying the internal organization of gender representations. We note that gender and bias are complicated and multi-faceted constructs. When studying gender bias in neural models, we unavoidably rely on a narrow notion of gender, as reflected in several annotated datasets. As such, we see this study as a preliminary attempt that is based on a relatively narrow concept of gender bias, that does not reflect the subtle ways by which social gender is manifested. We advise for caution when applying the conclusions of this study to other notions of gender or other definitions of bias.

We acknowledge that gender is not a binary property. Due to lack of existing resources, we use binary gender as a rough approximation of reality. We hope to account for this in future work.

References

Srijan Bansal, Vishal Garimella, Ayush Suhane, and Animesh Mukherjee. 2021. Debiasing multilingual

word embeddings: A case study of three indian languages. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 27–34.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *arXiv:2002.03518*.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. *CoRR*, abs/2005.04511.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805v1*.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing word-level translations from multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Online.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv:1911.03310*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach. In *ICLR*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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.

Benjamin Muller, Beno^ıt Sagot, and Djame Seddah. 2020. Can multilingual language models transfer to an unseen dialect? a case study on north african arabizi. *arXiv:2005.00318*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *ArXiv*, abs/2004.07667.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2021. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*, 9:139–159.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.