Stereotype Shifts in Multilingual Language Models

Gerard Planella **Veljko Kovac** **Adam Valin**

Abstract

Warning: This paper contains content that may be offensive or upsetting.

In this paper, we explore the problem of stereotypes and biases in Language Models, focusing on how different social groups are represented in different languages and how fine-tuning these models for specific languages can affect these representations. This paper focuses on stereotypes in Multilingual Language Models (MLM) which are pre-trained on extensive multilingual text collections. Specifically, we investigate the emotion profiles of certain social groups across different languages within these MLMs and analyze the impact of language-specific fine-tuning on biased data. By comparing the emotion profiles before and after fine-tuning, we analyse whether this process triggers a shift in stereotypes across languages. In this study, we aim to gain a better understanding of the intricacies between language-specific fine-tuning, inherent biases, and their combined influence on MLMs.

1 Introduction

The advent of sophisticated natural language processing techniques, specifically Language Models (LM), has recently brought a lot of attention. These models have demonstrated remarkable capabilities in understanding, generating, and translating text, thereby driving advancements in fields ranging from artificial intelligence to computational linguistics and beyond. Nevertheless, the potential of these models is hampered by a critical issue, namely bias. The inherent biases within these models, often mirroring those in the data they are trained on, can inadvertently perpetuate and even amplify societal prejudices and stereotypes. One increasingly important field of research is about quantifying those biases on multiple natural language processing applications. To this effect, one of the initial studies on the topic (Bolukbasi et al., 2016) highlighted the gender biases captured in contextual word embeddings. The research revealed that embeddings for gendered male words tend to be closer to higher-status jobs such as *doctor programmer* while gendered female words were associated with *homemaker* and *nurse*.

While bias is an unintended preference or prejudice that can be harmful when it interferes with the ability to be impartial, stereotypes can be defined as a preconceived idea that (incorrectly) attributes general characteristics to all members of a group. Previous work investigated which stereotypical associations are encoded in Language Models for different target groups (Choenni et al., 2021). Moreover, an emotion profile was built using the NRC Emotion Lexicon of Mohammad and Turney (2013b) for each target group based on the words with which the Language Models would complete certain template queries for a particular group. The lexicon quantifies the emotions within each word and allows a quantification of the emotions associated with specific social groups by analysing the predicted words of the Language Model.

In this project, we study stereotypes that emerge within pre-trained Multilingual Language Models (MLMs). These models are typically trained on large-scale multilingual text corpora, learning to represent the different languages in a shared embedding space. This shared representation allows the model to generalize knowledge learned in one language to other languages, a phenomenon known as cross-lingual transfer (Conneau et al., 2019). Our objective is to build upon the work of Choenni et al. (2021) by initially contrasting emotion profiles of identical social groups across diverse languages within these MLMs. Considering that pre-training strategies for these large-scale MLMs often depend on human-generated datasets such as CommonCrawl, we hypothesize that the behavioural patterns of a model will exhibit language-specific differences. This variability may

provide insights into the intrinsic biases of the models and the text corpora they were trained on. To further our understanding, we plan to conduct a quantitative investigation to evaluate the impact of language-specific fine-tuning on biased data. This data is drawn from selected news articles in several languages as well as from other specific stereotype datasets. By comparing the emotion profiles before and after the fine-tuning process, we aim to discern to what extent language-specific fine-tuning instigates a shift in stereotypes across different languages. The overarching goal of our research is to provide a better understanding of the relationship between language-specific fine-tuning, inherent biases, and their collective impact on MLMs.

Our results support the hypothesis that structurally similar languages share a greater number of stereotypes compared to others. Moreover, we analysed the way in which emotion profiles evolve during the fine-tuning process, finding how for similar languages (based on data from the World Altas of Language Structures) emotion profiles seem to evolve in correlation with each other. However, some similar languages do not experience the same kinds of parallel shifts due to extrinsic reasons such as cultural or geographical differences.

2 Related Work

2.1 Bias and Stereotypes in Language Models

The effect of model biases on social groups has been studied in several works such as Blodgett et al. (2020). To address the issue of stereotypes, recent works have evaluated which stereotypical associations are encoded in LMs for different target groups (Nadeem et al., 2021; Nangia et al., 2020; de Vassimon Manela et al., 2021). They examined whether models show a preference for stereotypical sentences over anti-stereotypical ones by employing datasets consisting of pairs of sentences. Recent work (Choenni et al., 2021) further investigated which stereotypes emerge within pre-trained LMs for different target groups, studying the LM's top k salient attributes, where salient attributes refer to the properties or characteristics related to the top-k predictions of a LM for a specific prompt. They built emotion profiles for each target group based on the words with which the LMs would complete certain template queries for the groups, basing their work on that of Kurita et al. (2019) and steering away from the more restrictive *she/he* sentence completion approach shown in their work.

2.2 Debiasing in LM Training

Several works have addressed the issue of bias in LMs by debiasing the training data or modifying the learning objective. Examples of such works include WinoBias (Zhao et al., 2018) and CrowS-Pairs (Nangia et al., 2020). Recently, Gira et al. (2022) proposed an efficient debiasing method, which outperforms previous state-of-the-art debiasing techniques on multiple benchmarks for word embeddings and LMs.

2.3 Quantifying Bias in Languages

Different social groups evoke varying emotional profiles, which may be influenced by stereotypes. Some works have explored stereotypes in intergroup contexts (Cottrell and Neuberg, 2005; Tapias et al., 2007; Mackie et al., 2000; Choenni et al., 2021). Such research can inform the development of more nuanced metrics to quantify biases and stereotypes in LMs. Following this initial focus on emotions in certain languages, researchers turned their attention to a more specific aspect of linguistic bias, namely gender. The first efforts to quantify bias in gendered languages were conducted by Zhou et al. (2019) for Spanish and French. More recently, Malik et al. (2022) carried out a similar study for Hindi, although gender is structured differently in this language (Hall, 2002). These studies aimed to assess the degree to which LMs associate masculine and feminine gender with specific professions.

3 Methods

Language Similarities When comparing languages' similarities, the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) can be used to detect their structural properties. The most commonly used categories in language comparison are Phonology, Morphology, Syntax, Grammar and Lexicality. In this paper, although we focus on the stereotypes resulting from different languages rather than the actual differences among languages, it is still important to analyse these linguistic differences. Statistical analysis was performed for the different chosen languages (Croatian, English, French, Greek, Spanish), using features obtained from WALS and ranking them by Euclidean distance. The goal of this analysis

was to support the idea that the emotion profiles of similar languages will also vary in a similar way after fine-tuning a LM.

Emotion Profiles Qualificatives predicted by the models to characterize a specific group can be associated with an emotion. Quantifying emotion profiles allows the evaluation of stereotypes for every social group. To obtain an emotion profile for each target group in different languages we first translated the social groups and prompts introduced by Choenni et al. (2021) in all chosen languages (English, Greek, French, and Spanish, Croatian). Once all of the prompts had been translated, we used XLM-R, to generate the top predictions for each translated prompt. Secondly, we performed a re-ranking by typicality, measuring the association between the words by computing the chance of completing the template, which corrects for the high-probability words with no stereotypical value. Moreover, we used the stereotype elicitation method introduced by Choenni et al. (2021). As explained in the paper, this method quantifies typicality by computing the log probability of the model probability for the predicted completion, corrected by the prior probability of the completion e.g:

$$P_{\text{post}}(y = \text{strict}|\text{Why are parents so } y?) \quad (1)$$

$$P_{\text{prior}}(y = \text{strict}|\text{Why are [MASK] so } y?) \quad (2)$$

$$p = \log\left(\frac{P_{\text{post}}}{P_{\text{prior}}}\right) \quad (3)$$

However, with multilingual vocabularies, words in languages different to that of the prompt are assigned a low prior probability and are thus ranked highly. To adapt the stereotype elicitation method for multilingual analysis, words belonging to a different language than the one of the provided prompt were removed and the priors were recalculated based on the remaining words, e.g :

$$P_{\text{prior}}(y = \text{strict}|y \in Vocab_{Eng}, \text{Why are [MASK] so } y?) \quad (4)$$

This adjustment aimed to reduce the likelihood of detecting words from languages different to the prompt's language as stereotypes. By employing this adapted method, we were able to effectively analyze the emotion profiles of various target groups across different languages.

Emotion Profile Analysis Once we obtained the predictions for the prompts, we matched these to the NRC Emotion Lexicon (Mohammad and Turney, 2013a). We constructed a matrix for each language with social groups as rows and Ekman's eight basic emotions (fear, joy, anticipation, trust, surprise, sadness, anger, and disgust)(Ekman, 1992) as columns, with two additional columns representing the emotions of positivity and negativity. Using the Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) method, we compared the similarities between the stereotypes that emerge from different languages within the model. RSA was chosen as the comparison method as it takes into account the relative relations between languages and groups, which is crucial for understanding how stereotypes differ among different languages. Our choice of comparison method is essential, as the number of words present in the emotion lexicon varies significantly between languages, ranging from approximately 20% to 75% which could lead to unfair comparisons if absolute numbers would be compared. For each language, a Representational Similarity Matrix (RSM) is constructed, where each element a_{ij} represents the cosine similarity between the i-th and j-th group in a list of social groups. After obtaining the RSM for each language, the Spearman correlation is computed for each RSM (language) pair. This generates a correlation score for each social group in a language pair. Finally, the mean over all social groups is used as a measure for correlation between the emotion profile of each language pair.

Beyond the interest in RSA coefficients, our study also aims to explore the relative shift of emotion profiles for each language pair after each fine-tuning. To achieve this, we subtract the emotion profiles of the pre-trained model from those of the fine-tuned models for each language under consideration. This procedure results in 10-dimensional vectors per social group, corresponding to each column in the emotion profile matrix (each column corresponding to an emotion). These vectors can then be averaged

to find a global average vector per languages across social groups or per social groups across languages. Lastly, they can be compared using cosine similarity, a measure that can capture the degree of similarity between two vectors.

Fine-Tuning Having the results for the RSA correlations between the emotion profiles of several languages in pre-trained models, our goal is to examine the shifts of these language's emotion profiles relative to each other following language-based fine-tuning processes. Specifically, we focus on the RSA correlations between the emotion profiles for all selected languages. This is done by comparing the emotion profiles obtained from different fine-tunings on datasets of only these four languages: English, Greek, French, and Spanish. Regrettably, we could not find a specifically biased Croatian dataset for our study and thus no fine-tuning was performed for this language. Given the inherent bias towards stereotypes in our datasets, we draw upon specific *Sensitive Minorities* as subjects for creating and comparing the emotion profiles before and after fine-tuning. This group encompasses 35 manually selected social groups (Appendix C) more susceptible to discriminative and racial biases, such as "Black People" and "Homosexuals". Our decision to concentrate on a sub-selection of data was driven by our hypothesis that it would provide a more significant shift of emotion profiles after fine-tuning. The motivation was the understanding that the examination of a large array of social groups could result in the neutralization of certain meaningful emotion shifts, due to the added noise related to potentially less sensible minorities.

4 Experiments

The first experiment involved conducting a language analysis on a selected set of languages. As outlined in Section 3, we conducted comparisons using language features obtained from the World Atlas of Language Structures (WALS). The analysis was performed on the selected languages, namely: Spanish, English, French, Croatian, Greek. Subsequently, any features that were identical across all languages or had empty values were removed and the following features were dropped: wals_code, iso_code, glottocode, countrycodes, latitude, longitude, and macroarea. This resulted in a 26-dimensional feature vector for each selected language. One can refer to Appendix A for a list of the selected features. We then employed Principal Component Analysis (PCA) (F.R.S., 1901) for dimensionality reduction, yielding a two-dimensional feature vector for each language. Finally, we calculated the Euclidean distance between each pair of languages, providing a similarity measure for all languages. This measure was used as a direct comparison between the linguistic properties of each language.

The LM used in this study is XLM-R (Conneau et al., 2019), a multilingual adaptation of the RoBERTa model (Liu et al., 2019) based on the Transformer architecture (Vaswani et al., 2017), consisting of stacked self-attention and feed-forward layers. This model leverages the efficiency of unsupervised pre-training and fine-tuning on downstream tasks. XLM-R incorporates cross-lingual training, where it learns from parallel data in different languages, enabling it to transfer knowledge across languages. It utilizes a shared sub-word vocabulary and introduces language-specific embeddings to capture language nuances effectively. This architecture enables XLM-R to achieve impressive performance on various cross-lingual tasks, including machine translation, document classification, and sentiment analysis.

Having examined the differences between the chosen languages, the next step consisted in analysing the similarity of the emotion profiles for the pre-trained XLM-R model using the prompts and social groups for the languages stated in Section 3. After computing the RSA for each social group for different pairs of languages, we obtained a measure that allowed us to quantify the similarity of emotional profiles across these languages. Averaging for each social group, we obtained a value for each language pair indicating the proximity between the emotional profiles for different languages.

Once we obtained the emotion profile similarity for each language pair, we proceeded to fine-tune the model on different datasets. The distinct datasets used for the various fine-tuning processes include Stereoset (Nadeem et al., 2021), French CrowS-Pairs (Nangia et al., 2020), a collection of Spanish news articles web-scraped from the right-winged *La Razón* journal, and a compilation of Greek offensive tweets (which can both be found in Appendix B).

The fine-tuning process was standardized across

all models by running for one epoch, with an attempt to train each model on a similar amount of data. This was done to mitigate disparities in sample size between different datasets. For instance, the Spanish dataset, derived from news articles, was considerably larger than the English Stereoset dataset, which primarily consisted of sentences. Consequently, we adjusted the training split size to ensure comparable dataset lengths across all language models. The selected batch size was 8 and the learning rate was 1e-5. The used optimizer was Adam and the training objective was Mask Language Modelling. Following the fine-tuning process, we examined the RSA correlation coefficients between the emotion profiles of languages from the base model and the fine-tuned model. If the correlation coefficient for a particular language is close to 1, it signifies that the fine-tuning has not significantly modified the model's top-k salient attributes for the same prompts (k = 300 in this study). Consequently, this implies that the stereotypes associated with the language have remained relatively unchanged. If, however, the correlation coefficient is low or approaching zero, it denotes a substantial influence of the fine-tuning process on the model's outputs. In this scenario, the model's predictions have changed noticeably after fine-tuning. Moreover, a significant negative correlation could suggest that the emotion profiles shift in opposite directions as a result of the fine-tuning process. Finally, we compare shifts in emotion profiles between language pairs before and after fine-tuning using cosine similarity as previously explained in Section 3.

5 Results and Analysis

The first set of results allows us to investigate the similarities between the linguistic properties of the chosen languages. Table 1 shows the distances between these. It can be seen how Spanish and French are the closest in terms of language features, followed by English and French. Furthermore, the most distant languages are French and Croatian, English and Croatian, and Greek and Croatian. We therefore expect that the emotion profiles for the closest languages vary in a similar way to each other whereas with more distant languages variation becomes negatively correlated.

After computing the distances between the language features obtained from WALS, the correlation between the emotion profiles for different languages across all social groups was computed using the pre-trained XLM-R model. The results can be observed in Table 2, where a larger value represents a larger correlation between the language's emotion profiles. It can be observed that language pairs like English and French, or English and Spanish have higher RSA correlation values than English and Greek or French and Croatian. These results support the idea that similar languages should share similar stereotypes, which is reflected in a higher RSA correlation between their emotion profiles. Nevertheless, this does not hold for all language pairs. For example, in Table 2, English and Greek are categorized much closer than French and Croatian. However, the mean RSA values or the former pair are the lowest in the table. This was to be expected, as even though the linguistic properties of the languages may be similar, the latent stereotypes present in the training data for each language may differ largely due to cultural or geographic factors not considered in Table 1.

	English	French	Greek	Croatian	Spanish
English	0.00	–	–	–	–
French	1.64	0.00	–	–	–
Greek	1.99	3.43	0.00	–	–
Croatian	7.91	8.90	6.07	0.00	–
Spanish	1.29	0.58	3.23	8.95	0.00

Table 1: Euclidean distance between language vectors using the WALS data. Lower half of the symmetric matrix including diagonal is shown.

Language 1	Language 2	Mean RSA
English	Croatian	0.317
English	French	**0.445**
English	Greek	**0.172**
French	Croatian	0.291
French	Greek	0.209
French	Spanish	0.322
Greek	Croatian	0.264
Spanish	Croatian	0.310
Spanish	English	0.346
Spanish	Greek	0.234

Table 2: This table shows the mean value for the RSA correlation across all social groups between languages when using the pre-trained XLM-R model.

Once information regarding the initial RSA correlations between stereotypes of different

languages was acquired, we analysed the variation of the emotion profiles for each language before and after fine-tuning. As seen in Table 3, all languages experience significant shifts in their emotion profiles after fine-tuning on different language-specfic datasets. The most notable correlations are observed between English and French in the English fine-tuning scenario, with specific values of 0.604 and 0.594 respectively. Moreover, we can see that Greek and French fine-tuning produces highly decorrelated emotion profiles for every language. One possible explanation could be the amount of data on which the model has been pre-trained on regarding those two languages, possibly being trained on much more French data than Greek data. Nonetheless, drawing conclusions about the shift of emotion profiles relative to the languages used for fine-tuning proves challenging due to the different nature of our datasets.

Nevertheless, trends in the evolution of emotion profiles in relation to each other can still be discerned. It appears that structurally similar languages according to Table 1 tend to follow the same patterns. For example, French, English and Spanish appear to exhibit parallel shifts for every fine-tuning, maintaining a consistent degree of correlation to their respective baseline emotion profiles. Moreover, Table 4 and Table 5 present high cosine similarity values for shifts in emotion profiles for English and French during English and Spanish fine-tuning, respectively measured as 0.638 and 0.881. This suggests that these two languages not only share stereotypes but also evolve in closer alignment than other pairs, such as English and Croatian. An intriguing observation is the low RSA coefficient of the Spanish language during English fine-tuning, at 0.228, in contrast to French and English languages (0.604 and 0.594). This implies a substantial shift in the emotion profile of the Spanish language from its initial state during English fine-tuning. Nevertheless, according to Table 4 and Table 5, the shifts in Spanish align with those of French and English. This indicates that while the emotion profiles of these languages evolve similarly, their initial states vary, explaining why the Spanish emotion profile is highly decorrelated from its initial pre-trained state despite undergoing shifts akin to those of the French and English languages. In reference to Greek and Croatian, these languages do not appear to exhibit a significant level of similarity in the emotion profile shift with the other languages in our group. Table 1 could provide an explanation for this observation, as it shows that Croatian and Greek stand out with the highest global Euclidean distance from the rest of the languages.

	English FT	French FT	Greek FT	Spanish FT
English	0.604	0.082	0.119	0.218
French	0.594	0.097	0.069	0.212
Spanish	0.228	0.140	0.094	0.315
Croatian	0.212	0.183	0.229	0.182
Greek	0.134	0.203	0.203	0.153

Table 3: RSA comparison of emotion profiles for *Sensitive Minorities* across different languages after fine-tuning. The top row represents the languages used for fine-tuning, while the first column represents the language used to calculate the correlations.

	English	French	Spanish	Greek	Croatian
English	1.000	–	–	–	–
French	0.638	1.000	–	–	–
Spanish	0.660	0.861	1.000	–	–
Greek	-0.206	-0.466	-0.308	1.000	–
Croatian	-0.309	-0.124	0.145	-0.067	1.000

Table 4: Cosine similarity values of emotion profiles shifts for each language when fine-tuned on English for *Sensitive Minorities*.

	English	French	Spanish	Greek	Croatian
English	1.000	–	–	–	–
French	0.881	1.000	–	–	–
Spanish	0.858	0.958	1.000	–	–
Greek	-0.297	-0.234	-0.146	1.000	–
Croatian	0.253	0.342	0.281	0.409	1.000

Table 5: Cosine similarity values of emotion profiles shifts for each language when fine-tuned on Spanish for *Sensitive Minorities*.

6 Conclusion

In this study, we analyzed the structural properties and emotion profiles of selected languages using the XLM-R language model. Our findings shed light on the similarities and variations in stereotypes associated with different languages.

Firstly, we compared the linguistic properties of the chosen languages using WALS features. The results indicate that these properties can vary significantly, and the similarity between languages does not always align with their geographic or

cultural proximity. Secondly, we explored the emotion profiles of the pre-trained XLM-R model for the selected languages and social groups. The RSA correlation coefficients revealed that structurally similar languages tended to have higher correlation values in their emotional profiles. However, there were exceptions to this pattern, suggesting the influence of cultural and geographic factors on the associated stereotypes. Through fine-tuning experiments, we observed significant shifts in the emotion profiles of all languages. Similar languages tend to display comparable patterns of development in their emotion profiles during fine-tuning, while languages with distinct linguistic properties showed distinct shifts. These findings underscore the complex relationship between language, culture, and stereotypes, emphasizing the need to consider multiple factors when analyzing language model outputs. Our findings highlight the importance of considering linguistic properties, cultural contexts, and training data biases when examining stereotype shifts in LMs.

Future Research To further enhance our understanding of this complex relationship between LMs, stereotypes, and linguistic properties, several avenues for future research emerge. First, validating and generalising our claims by testing our approach on different LMs would be valuable. By exploring how our findings hold across diverse architectures and training methodologies, we can better understand the generalizability of our conclusions. Additionally, we could conduct fine-tuning experiments on datasets of similar nature, such as exclusively using news datasets in different languages. This focused analysis would allow us to investigate how fine-tuning impacts the stereotypes associated with specific domains or media sources, providing a better understanding of the relationship between fine-tuning and contextual biases. Expanding our analysis to include more dissimilar languages, such as Hindi or Mandarin, would offer a broader perspective on the impact of linguistic and cultural variations on stereotype shifts. Exploring how stereotypes evolve in languages that differ significantly from the ones examined in this study could reveal novel insights into cross-lingual emotion processing and highlight the influence of cultural factors. Furthermore, varying the number of epochs during fine-tuning could provide insights into the dynamics of stereotype shifts over time. By tracking how the model's predictions change as it overfits the training data, we can gain a better understanding of the stability and robustness of the learned stereotypes. Finally, exploring different training objectives beyond Masked Language Modelling, such as Natural Language Inference, could offer alternative perspectives on how fine-tuning impacts stereotype representations. Comparing the results of varying training objectives would provide a more comprehensive understanding of the relationship between training objectives, stereotype shifts, and LM's outputs.

By addressing these future research directions, we can deepen our understanding of LM's behaviour and progress towards developing more inclusive and unbiased AI systems.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? *arXiv preprint arXiv:2109.10052*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Catherine A Cottrell and Steven L Neuberg. 2005. Different emotional reactions to different groups: A sociofunctional threat-based approach to "prejudice". *Journal of Personality and Social Psychology*, 88(5):770–789.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European*

Chapter of the Association for Computational Linguistics: Main Volume, pages 2232–2242, Online. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Paul Ekman. 1992. An argument for basic emotions. In *Emotions: Essays on emotion theory*, pages 169–200. Erlbaum, Hillsdale, NJ.

Karl Pearson F.R.S. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69.

Kira Hall. 2002. Unnatural" gender in hindi. *Gender across languages: The linguistic representation of women and men. Amsterdam: John Benjamins*, pages 133–162.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Diane M Mackie, Thierry Devos, and Eliot R Smith. 2000. Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal of Personality and Social Psychology*, 79(4):602–616.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013a. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M Mohammad and Peter D Turney. 2013b. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Maria P Tapias, Jack Glaser, Dacher Keltner, Karina Vasquez, and Thomas Wickens. 2007. Emotion and prejudice: Specific emotions toward outgroups. *Group Processes & Intergroup Relations*, 10(1):27–39.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. *arXiv preprint arXiv:1909.02224*.

A List of used WALS features

As mentioned in Section 4, we compared the linguistic properties of each language using the WALS features. The remaining features used after cleaning and processing were the following:

- **Genus**: Classification of languages into groups based on shared characteristics.
- **Family**: Grouping of languages into families based on their historical relationships.
- **53A Ordinal Numerals**: The presence or absence of ordinal numerals in the language.
- **70A The Morphological Imperative**: The presence or absence of a morphological imperative in the language.

- **71A The Prohibitive**: The presence or absence of a prohibitive construction in the language.
- **72A Imperative-Hortative Systems**: The presence or absence of imperative-hortative systems in the language.
- **74A Situational Possibility**: The presence or absence of situational possibility constructions in the language.
- **75A Epistemic Possibility**: The presence or absence of epistemic possibility constructions in the language.
- **76A Overlap between Situational and Epistemic Modal Marking**: The presence or absence of overlap between situational and epistemic modal marking in the language.
- **81A Order of Subject, Object and Verb**: The order of subject, object, and verb in basic declarative clauses in the language.
- **82A Order of Subject and Verb**: The order of subject and verb in basic declarative clauses in the language.
- **83A Order of Object and Verb**: The order of object and verb in basic declarative clauses in the language.
- **85A Order of Adposition and Noun Phrase**: The order of adposition and noun phrase in the language.
- **86A Order of Genitive and Noun**: The order of genitive and noun in the language.
- **87A Order of Adjective and Noun**: The order of adjective and noun in the language.
- **92A Position of Polar Question Particles**: The position of polar question particles in the language.
- **95A Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase**: The relationship between the order of object and verb and the order of adposition and noun phrase in the language.
- **97A Relationship between the Order of Object and Verb and the Order of Adjective and Noun**: The relationship between the order of object and verb and the order of adjective and noun in the language.
- **112A Negative Morphemes**: The presence or absence of negative morphemes in the language.
- **115A Negative Indefinite Pronouns and Predicate Negation**: The presence or absence of negative indefinite pronouns and predicate negation in the language.
- **116A Polar Questions**: The presence or absence of polar questions in the language.
- **138A Tea**: The presence or absence of a specific word for "tea" in the language.
- **143F Postverbal Negative Morphemes**: The position of negative morphemes relative to the verb in the language.
- **144A Position of Negative Word With Respect to Subject, Object, and Verb**: The position of negative words with respect to subject, object, and verb in the language.
- **143E Preverbal Negative Morphemes**: The position of negative morphemes relative to the verb in the language.
- **143A Order of Negative Morpheme and Verb**: The order of negative morpheme and verb in the language.

B Spanish and Greek Datasets

Used Spanish dataset containing articles from La Razón: Spanish News Article Dataset.
Greek Dataset used containing offensive Tweets: Offensive Greek Tweet Dataset.

C Sensitive Minorities

The social groups that are part of the "Sensitive Minorities are the following: "Black people", "Mexicans","Blacks","Arabs", "Black men", "Asian women", "Asian men", "Black women", "Indian men", "Indian women", "Chinese men", "Chinese women", "Latinas", "Latinos", "Black kids", "African Americans", "Afghans", "Iraqis", "Iranian people", "Iranians", "Syrians", "Pakistanis", "Romanians", "Turkish people", "Muslims", "Mormons", "immigrants", "gay people", "lesbians", "queer people", "homosexuals", "faggots", "fags".

D Emotion profile RSA Correlations and Cosine Similarities for different models

This section shows the emotion profiles obtained for all selected language pairs for the pre-trained model and after fine-tuning on specific languages. The final column of each table computes the mean RSA values for each language pair accross all social groups.

Language 1	Language 2	Age	Countries	Gender	Lifestyle	Political	Profession	Race	Religion	Sexuality	Mean
English	French	0.407	0.340	0.268	0.570	0.927	0.300	0.333	0.424	0.439	0.445
English	Croatian	0.267	0.2744	0.260	0.231	0.515	0.297	0.198	0.496	0.312	0.317
English	Greek	0.145	0.056	0.128	0.283	0.219	0.104	0.095	0.306	0.215	0.172
English	Spanish	0.316	0.502	0.233	0.452	0.315	0.257	0.282	0.446	0.313	0.346
French	Croatian	0.295	0.293	0.178	0.155	0.559	0.265	0.201	0.425	0.245	0.294
French	Greek	0.243	0.026	0.247	0.198	0.265	0.109	0.129	0.448	0.215	0.214
French	Spanish	0.251	0.227	0.306	0.463	0.334	0.293	0.273	0.399	0.353	0.311
Spanish	Croatian	0.454	0.211	0.117	0.208	0.449	0.286	0.123	0.252	0.694	0.316
Spanish	Greek	0.248	0.046	0.231	0.207	0.410	0.156	0.147	0.238	0.420	0.233
Greek	Croatian	0.338	0.094	0.152	0.274	0.329	0.140	0.142	0.544	0.367	0.262

Table 6: This table highlights the RSA correlations between the emotion profiles for different languages using the base pre-trained XLM-R model. Highlighted in blue we can observe the smallest correlation values per social group and in red the largest correlation values. The last column provides the mean value for the correlations for each language pair.

Language 1	Language 2	Age	Countries	Gender	Lifestyle	Political	Profession	Race	Religion	Sexuality	Mean
English	French	0.480	0.262	0.275	0.449	0.601	0.328	0.391	0.144	0.415	0.389
English	Croatian	0.365	0.338	0.282	0.308	0.340	0.217	0.229	0.535	0.181	0.326
English	Greek	0.232	0.077	0.134	0.211	0.258	0.058	0.064	0.206	-0.002	0.151
English	Spanish	0.397	0.119	0.323	0.105	0.312	0.100	0.155	0.378	0.230	0.248
French	Croatian	0.229	0.207	0.126	0.281	0.164	0.172	0.206	0.276	0.283	0.216
French	Greek	0.480	0.113	0.223	0.167	0.146	0.142	0.105	0.210	0.158	0.199
French	Spanish	0.449	0.091	0.281	0.171	0.454	0.089	0.136	0.351	0.341	0.269
Spanish	Croatian	0.458	0.109	0.240	0.069	0.169	0.107	0.208	0.292	0.426	0.255
Spanish	Greek	0.047	0.051	0.258	0.155	0.341	0.092	0.108	0.271	0.226	0.196
Greek	Croatian	0.338	0.065	0.394	0.241	0.243	0.054	0.102	0.278	0.368	0.241

Table 7: This table highlights the RSA correlations between the emotion profiles for different languages using the fine-tuned English model. Highlighted in blue we can observe the smallest correlation values per social group and in red the largest correlation values. The last column provides the mean value for the correlations for each language pair.

Language 1	Language 2	Age	Countries	Gender	Lifestyle	Political	Profession	Race	Religion	Sexuality	Mean
English	Croatian	0.312	0.037	0.182	0.097	0.344	0.061	0.055	0.255	0.185	0.170
English	French	0.293	0.047	0.115	0.207	0.457	0.107	0.095	0.251	0.218	0.199
English	Greek	0.092	0.059	0.215	0.180	0.178	0.032	0.007	0.136	0.192	0.121
French	Croatian	0.474	0.115	0.166	0.037	0.183	0.100	0.041	0.504	0.393	0.224
French	Greek	0.303	0.065	0.057	0.196	0.281	0.026	0.063	0.278	0.277	0.172
French	Spanish	0.295	0.112	0.384	0.139	0.025	0.028	0.154	0.216	0.420	0.197
Greek	Croatian	0.602	-0.004	0.281	0.176	0.375	0.073	0.094	0.447	0.198	0.249
Spanish	Croatian	0.329	0.116	0.177	0.427	0.461	0.009	0.053	0.320	0.603	0.277
Spanish	English	0.115	0.009	0.220	0.036	0.189	0.082	0.051	0.394	0.425	0.169
Spanish	Greek	0.342	0.044	0.057	0.130	0.368	0.103	0.102	0.175	0.151	0.163

Table 8: This table highlights the RSA correlations between the emotion profiles for different languages using the fine-tuned French model. Highlighted in blue we can observe the smallest correlation values per social group and in red the largest correlation values. The last column provides the mean value for the correlations for each language pair.

Language 1	Language 2	Age	Countries	Gender	Lifestyle	Political	Profession	Race	Religion	Sexuality	Mean
English	Croatian	0.312	0.037	0.182	0.097	0.344	0.061	0.055	0.255	0.185	0.141
English	French	0.293	0.047	0.115	0.207	0.457	0.107	0.095	0.251	0.218	0.292
English	Greek	0.092	0.059	0.215	0.180	0.178	0.032	0.007	0.136	0.192	0.223
French	Croatian	0.474	0.115	0.166	0.037	0.183	0.100	0.041	0.504	0.393	0.207
French	Greek	0.303	0.065	0.057	0.196	0.281	0.026	0.063	0.278	0.277	0.212
French	Spanish	0.295	0.112	0.384	0.139	0.025	0.028	0.154	0.216	0.420	0.171
Greek	Croatian	0.602	-0.004	0.281	0.176	0.375	0.073	0.094	0.447	0.198	0.212
Spanish	Croatian	0.329	0.116	0.177	0.427	0.461	0.009	0.053	0.320	0.603	0.150
Spanish	English	0.115	0.009	0.220	0.036	0.189	0.082	0.051	0.394	0.425	0.209
Spanish	Greek	0.342	0.044	0.057	0.130	0.368	0.103	0.102	0.175	0.151	0.267

Table 9: This table highlights the RSA correlations between the emotion profiles for different languages using the fine-tuned Spanish model. Highlighted in blue we can observe the smallest correlation values per social group and in red the largest correlation values. The last column provides the mean value for the correlations for each language pair.

Language 1	Language 2	Age	Countries	Gender	Lifestyle	Political	Profession	Race	Religion	Sexuality	Mean
English	Croatian	0.233	0.026	0.130	0.223	0.270	0.093	0.084	0.330	0.504	0.210
English	French	0.325	0.040	0.140	0.171	0.526	0.099	0.072	0.115	0.546	0.226
English	Greek	0.337	0.011	0.080	0.081	0.253	0.061	0.062	0.155	0.466	0.167
French	Croatian	0.281	0.053	0.134	0.236	0.253	0.053	0.061	0.257	0.503	0.203
French	Greek	0.274	0.042	0.235	0.106	0.189	0.067	0.055	0.373	0.355	0.189
French	Spanish	0.566	0.129	0.203	0.237	0.478	0.044	0.078	0.510	0.161	0.267
Greek	Croatian	0.309	0.074	0.164	0.154	0.184	0.022	0.007	0.382	0.494	0.199
Spanish	Croatian	0.310	0.007	0.104	0.236	0.241	0.091	0.102	0.495	0.320	0.212
Spanish	English	0.338	0.076	0.013	0.175	0.247	0.180	0.024	0.394	0.458	0.212
Spanish	Greek	0.308	0.081	0.105	0.223	0.305	0.050	0.048	0.400	0.370	0.210

Table 10: This table highlights the RSA correlations between the emotion profiles for different languages using the fine-tuned Greek model. Highlighted in blue we can observe the smallest correlation values per social group and in red the largest correlation values. The last column provides the mean value for the correlations for each language pair.

	English	French	Spanish	Greek	Croatian
English	1.000	–	–	–	–
French	0.848	1.000	–	–	–
Spanish	0.676	0.908	1.000	–	–
Greek	-0.578	-0.337	-0.253	1.000	–
Croatian	0.174	-0.113	-0.148	0.086	1.000

Table 11: Cosine similarity values of emotion profiles shifts for each language when fine-tuned on French for Sensitive Minorities

	English	French	Spanish	Greek	Croatian
English	1.000	–	–	–	–
French	0.838	1.000	–	–	–
Spanish	0.717	0.935	1.000	–	–
Greek	-0.511	-0.491	-0.438	1.000	–
Croatian	0.098	0.168	0.084	-0.070	1.000

Table 12: Cosine similarity values of emotion profiles shifts for each language when fine-tuned on Greek for *Sensitive Minorities*.