Probing Pre-Trained Language Models for Cross-Cultural Differences in Values

Arora, Arnav; Kaffee, Lucie-Aimée Frimelle; Augenstein, Isabelle

DOI:
10.48550/arXiv.2203.13722

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Abstract

Language embeds information about social, cultural, and political values people hold. Prior work has explored potentially harmful social biases encoded in Pre-Trained Language Models (PTLMs). However, there has been no systematic study investigating how values embedded in these models vary across cultures. In this paper, we introduce probes to study which cross-cultural values are embedded in these models, and whether they align with existing theories and cross-cultural value surveys. We find that PTLMs capture differences in values across cultures, but those only weakly align with established value surveys. We discuss implications of using mis-aligned models in cross-cultural settings, as well as ways of aligning PTLMs with value surveys.

1 Introduction

A person’s identity, values and stances are often reflected in the linguistic choices one makes (Jaffe, 2009; Norton, 1997). This is why, when language models are trained on large text corpora, they not only learn to understand language, but also pick up on a variety of societal and cultural biases (Stanczak et al., 2021). While biases picked up by the PTLMs have a potential to cause harm when used in a downstream application, they may also serve as tools which provide insights into understanding cultural phenomena. Further, while studying ways of surfacing and mitigating potentially harmful biases is an active area of research, cultural biases and values picked up by PTLMs remain understudied. Here, we investigate cultural values and differences among them picked up by PTLMs through means of their pre-training on Web text.

In a wide range of social science research fields, values are a crucial tool for understanding cross-cultural differences. Values are the “core conceptions of the desirable within every individual and society” (Rokeach, 2008), i.e., the foundation for the beliefs guiding a person’s actions and on a society level the base for the guiding principles. In this paper, we base our understanding of values across cultures on two studies: Hofstede (2005), which defines 6 dimensions to describe cross-cultural differences in values, and the World Value Survey (WVS) (Haerpfer et al., 2022). Both surveys are widely used as a foundation to understand cross-cultural differences in values.

The surveys work on the assumption that they can capture cultural values through text, i.e., the answers in the surveys. PTLMs are trained on text from the Web, in several languages, giving them the potential to capture different information across different languages. In Wikipedia, cross-cultural differences have been established (Miquel-Ribé and Laniado, 2019), and analysed by Hara et al. (2010) based on Hofstede’s cultural dimensions theory. As Wikipedia is one of the primary sources of training data for multilingual LMs, there is a potential that these cultural differences can also be found in the LMs trained on its data. In this paper, we explore the novel research question of whether PTLMs capture cultural differences in terms of values across different language models. We probe PTLMs using the value surveys of both Hofstede’s cultural dimensions theory and the World Value Survey. We reformulate the survey questions to probe PTLMs and extract the answers to evaluate whether language models can capture cultural differences based on their training data. We focus on 13 languages, each of which is mostly geographically restricted to one country, to compare the results of the language models to the value surveys.

Our work makes the following contributions:

- We present the first study measuring cultural values embedded in large pre-trained language models
- We propose a methodology for probing for values by converting survey questions to cloze
style questions

• We conduct experiments across 13 languages with three language models (mBERT, XLM, and XLM-R), showing value alignment correlations with two large scale value surveys
• We present discussion around potential implications of deploying these models in a multicultural context

Finally, we release the code and data used for our experiments.¹

2 Related Work

Values and Norms CH-Wang and Jurgens (2021) study change in attitudes towards sexuality and gender by analysing change in use of lexical variables on texts from Twitter and Reddit across 10 years. Sap et al. (2017) introduce connotation frames for power and agency to study gender bias in modern films. Forbes et al. (2020) and Emelin et al. (2021) introduce large-scale annotated datasets for studying existing social norms useful for commonsense reasoning. Roy et al. (2021) present work on extracting moral sentiment using the Moral Foundation Theory. Sap et al. (2020) introduce the Social Bias Inference dataset to study inferred dynamics around social reasoning and bias in toxicity detection. Liu et al. (2021) introduce a multimodal multilingual dataset to diversify visually grounded reasoning across cultures.

Probing Probing has been used as tool to study a variety of knowledge and biases picked up by PTLMs. Wallace et al. (2019) study reasoning around numeracy captured by PTLMs by using a Question Answering model. Mosbach et al. (2020) study linguistic knowledge in 3 PTLMs by making use of fine-tuning and sentence-level probes. Stanczak et al. (2021) study gender bias towards politician names in cross-lingual language models. Ousidhoum et al. (2021) further study toxicity bias towards social groups.

To the best of our knowledge, there is no existing work on probing for values embedded in PTLMs in a comparative cultural context.

3 Value Probing

In this paper, we explore how PTLMs capture differences in values across cultures, and whether those differences reflect the ones found in values across cultures at large. To compare the PTLMs’ encodings of values, we compare them with established surveys capturing cross-cultural differences in values, namely Hofstede’s cultural dimensions theory and the World Value Survey (WVS) (Section 4). We transform the survey questions introduced in those surveys for compatibility with PTLMs by reformulating them semi-automatically to convert them into probes (Section 5). We then assess the variance in cross-cultural values in PTLMs and compare the probing results to the established value surveys in Section 7.

We investigate the following research questions as a first step to exploring this novel area of probing cross-cultural differences in value:

RQ1 Do PTLMs capture diversity across cultures for the established values?
RQ2 Are there similarities in the embedded values across different PTLMs?
RQ3 Do values embedded in PTLMs align with existing value surveys?

4 Value Surveys

We base our work on previous studies on how values differ across cultures. When discussing values, we refer to the core beliefs that motivate peoples’ actions. As these are central to a number of research fields, there are a large number of studies. Among the most common ones are Hofstede’s cultural dimensions theory and the World Value Survey. These studies build on the body of work in different fields: Hofstede’s theory is derived from management studies (Hofstede, 1984), while the WVS was developed in the field of political science (Inglehart, 2006). Both studies have since been widely used across fields.

4.1 Hofstede’s cultural dimensions theory

Hofstede started his surveys of cross-cultural differences in values in 1980. This first survey (Hofstede, 1984) included 116,000 participants from 40 countries (extended to 111 countries and regions in the 2015 version) working with IBM, and created 4 cultural dimensions, which were subsequently extended to 6 cultural dimensions that are also used in this paper. These 6 dimensions are: Power Distance (pdi), Individualism (idv), Uncertainty Avoidance (ua), Masculinity (mas), Long-term Orientation (lto), Indulgence (ivr). The full survey contains 24

¹https://github.com/copenlu/value-probing
questions. Each dimension is calculated using a formula defined by Hofstede using 4 of the questions in the survey, see Appendix C. Hofstede shows the influence that culture has on values by defining distinctly different numerical values in those 6 dimensions for the cultures observed. While critics of Hofstede’s cultural dimensions theory point out, among others, the simplicity of the approach of mapping cultures to countries and question the timeliness of the approach (Nasif et al., 1991), this model of representing values is now a foundation for large body of work on cross-cultural differences in values (Jones, 2007).

4.2 World Value Survey (WVS)

The World Value Survey (WVS, Haerpfer et al. (2022)) collects data on peoples’ values across cultures in a more detailed way than Hofstede’s cultural dimensions theory. The survey started in 1981 and is conducted by a nonprofit organisation, which includes a network of international researchers. The survey is conducted in waves, to collect data on how values change over time. The latest wave, wave 7, ran from 2017 to 2020. Compared to the European Value Study\(^2\), WVS targets all countries and regions, and includes 57 countries. While Hofstede’s cultural dimensions theory aggregates the findings of their survey into the 6 cultural dimensions, WVS publishes the results of their survey per question. Those are organised in 11 categories: (1) Corruption, (2) Ethical Values and Norms, (3) Happiness and Well-being, (4) Migration, (5) Political Culture and Regimes, (6) Political Interest and Political Participation, (7) Religious Values, (8) Science and Technology, (9) Security, (10) Social Capital, Trust and Organisational Membership, (11) Social Values, Attitudes and Stereotypes.

Inglehart (2006), who established WVS, further defines the Inglehart–Welzel cultural map, which processes the surveys and defines two dimensions in relation to each other: traditional versus secular-rational values and survival versus self-expression values, and summarise values for countries on a scatter plot describing these dimensions. In the following, we only use the originally defined 11 categories and leave an analysis based on the Inglehart–Welzel cultural map for future work.

5 Probe Generation

In order to make the surveys compatible with language models, we reformulate the survey questions to cloze-style question probes (Taylor, 1953; Hermann et al., 2015) that we can then perform masked language modelling inference on. Since this is the task PTLMs were trained on, we argue it is a suitable methodology to measure embedded cultural biases in these models.

Hofstede’s Cultural Dimensions Based on the English survey questions, the questions in the survey are manually reformulated to question probes (QPs). This is done analogously to iterative categorisation, in which a set of possible labels \(y_i^+, y_i^-\) corresponding to either end of the response options available in the survey are defined, which are the words the language models are probed for. The sentences are then reformulated to probes, and the labels masked. Those labels are based on the answers of the original survey, for instance, the original question like have sufficient time for personal or home life with answer options consisting of different degrees of importance, the probe is reformulated to Having sufficient time for personal or home life is [MASK]., where [MASK] should be replaced by important or unimportant. So \(QPs = W_i, y_i^+, y_i^-\) where \(W_i\) is the masked probe and \(y_i^+\) and \(y_i^-\) are the set of labels. There are a total of 24 questions with repeating labels.

World Value Survey Analogous to the probes created from the Hofstede survey, we create probes from the English questionnaire of the WVS. As there are more questions than for Hofstede (238 in total), there are also a larger number of labels to replace and a higher variety of question types.

Multilingual Probes To probe across several languages, we follow a semi-automatic methodology for translating the created probes in English to the target language. We use a translation API\(^3\) that covers all target languages. We translate each QP from English into the target language with the [MASK] token replaced by the label words \([y_i^+, y_i^-]\) in order to maintain grammatical structure and aid the translation API. One challenge of cross-cultural research is information loss when translating survey questions (Nasif et al., 1991; Hofstede, 1984). Therefore we opted for this approach rather than reformulating the translated survey questions by

\(^2\)https://europeanvaluesstudy.eu/

\(^3\)https://cloud.google.com/translate
<table>
<thead>
<tr>
<th>Country</th>
<th>Language</th>
<th>Wikipedia size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romania</td>
<td>Romanian (ro)</td>
<td>428,330</td>
</tr>
<tr>
<td>Greece</td>
<td>Greek (el)</td>
<td>207,647</td>
</tr>
<tr>
<td>Pakistan</td>
<td>Urdu (ur)</td>
<td>168,587</td>
</tr>
<tr>
<td>Iran</td>
<td>Farsi (fa)</td>
<td>872,240</td>
</tr>
<tr>
<td>Philippines</td>
<td>Tagalog (tl)</td>
<td>43,145</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Indonesian (id)</td>
<td>618,395</td>
</tr>
<tr>
<td>Germany</td>
<td>German (de)</td>
<td>2,675,084</td>
</tr>
<tr>
<td>Malaysia</td>
<td>Malay (ms)</td>
<td>356,937</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>Bengali (bn)</td>
<td>119,619</td>
</tr>
<tr>
<td>Serbia</td>
<td>Serbian (sr)</td>
<td>656,627</td>
</tr>
<tr>
<td>Turkey</td>
<td>Turkish (tr)</td>
<td>475,984</td>
</tr>
<tr>
<td>Vietnam</td>
<td>Vietnamese (vi)</td>
<td>1,270,712</td>
</tr>
<tr>
<td>Korea South</td>
<td>Korean (ko)</td>
<td>582,977</td>
</tr>
</tbody>
</table>

Table 1: Mapping of countries (cultures) to languages used throughout this paper, including number of articles per Wikipedia language as of March 2022.

Hofstede. The target labels \( y^+, y^- \) for each QP are then translated individually as single words (e.g. *important* is translated from English to the German *wichtig*), followed by lowercased string matching to check if the translated label can be found and replaced in the translated probe. If the target label cannot be found directly in the translated probe due to differences in word choice, we use a cross-lingual word aligner Dou and Neubig (2021) to align the English probe and its translated version. With this approach, we identify the label word to be replaced with the mask token. If both approaches yield no result, the token is manually replaced in the target sentence based on the authors’ language understanding and using online translators.

**Language selection** In total, we investigate 13 languages, mapped to one country each as outlined in Table 1, according to criteria further detailed below. One of the limitations of this one-to-one mapping is that the languages are spoken in wider regions and not specifically in one country (disregarding also e.g. diaspora communities). This allows for the closest match to the value theories we work with, which operate on country level. The definition of culture by country has been criticised by, e.g., Nasif et al. (1991).

We select the languages as follows: We first include the countries covered in both the surveys of WVS and Hofstede. We limit to languages which are official languages of the countries observed in the studies of both WVS and Hofstede. We further select languages for which the distribution of speakers is primarily localized to a country or relatively narrow geographical region. To ensure the language models will be able to have (potentially) sufficient amount of training data, from the set of languages, only those are selected which have at least 10,000 articles on Wikipedia and are present in mBERT.

6 Methodology

6.1 Models

We conduct the probing experiments on three widely used multilingual PTLMs: the multi-lingual, uncased version of BERT base (mBERT) (Devlin et al., 2018), the 100 language, MLM version of XLM (Conneau and Lample, 2019), and the base version of XLM-RoBERTa (XLM-R) (Conneau et al., 2020) available in the Transformers (Wolf et al., 2020) library. mBERT was trained with a Masked Language Modelling (MLM) and Next Sentence Prediction objective, on Wikipedia articles in 102 languages with the highest number of articles on them. The XLM model builds on top of mBERT, only using the MLM objective but with modifications to the selection and truncation of training text fed to the model at each training step. It was also trained on Wikipedia texts, including 100 languages. The XLM-R model uses the RoBERTa architecture (Liu et al., 2019) and is trained with an MLM objective on 2.5 TB of filtered CommonCrawl corpus data in 100 languages. It shows strong multilingual performance across a range of benchmarks and is commonly used for extracting multilingual sentence encodings.

6.2 Mask Probing

For each model \( M \), we run inference on the created cloze-style question probes (QPs, described in Section 5) using an MLM head producing the log probabilities for the [MASK] tokens in the QPs over the entire vocabulary \( V \) of the respective model: 

\[
\log P_M(w_i, t|W_i^{\backslash i}, \Theta_M) \in \mathbb{R}^{|V|}, \quad \text{where } t \text{ is the position of the [MASK] token in the text } W_i \in QP, \text{ and } \Theta_M \text{ are the parameters of the corresponding Language Model } M.
\]

Since the survey respondents have to answer the questions with a choice between a range of values, for instance 1-10 with 1 representing *democratic* and 10 representing *effective*, in order to replicate a similar setting with PTLMs, we subtract the predicted logit for the response label with the highest score \( w_i^+ \) with the predicted
logit for the lowest score \( w_i^- \). This normalises the predicted logits for the responses on opposing ends of the survey question and is then used as a score for that question.

\[
\log P_M(w_i) = \log P_M(w_i^+) - \log P_M(w_i^-) 
\]

Finally, in order to collapse the World Value Survey responses per category, within which many questions have different scales, we normalize the aggregate survey responses per the corresponding question scale, so that \( y_{i,c} \in [0,1] \), \( c \in C \). We then take the mean of the responses across all the questions of the category to arrive at the aggregated score of the category for each country:

\[
y_i = \frac{1}{|C|} \sum_{c \in C} y_{i,c} \in [0,1].
\]

6.3 Evaluation

We calculate Spearman’s \( \rho \) – a rank correlation coefficient between the values predicted by the language models and values calculated through the surveys: \( \rho(\log P_M(w_i,t|W,M,\Theta_M),y_i). \) For the World Value Survey, we do this per question, as well as per category. For Hofstede, we limit this calculation to value level correlations due to lack of access to individual or aggregate survey response data per question.\(^4\) We further calculate correlations per country. Spearman’s \( \rho \) works on relative predicted ranks to each variable, ignoring the individual predicted values. Our choice of using a rank correlation was motivated by the fact that we are working with population level aggregate responses and our aim of assessing whether language models pick up on relative differences in values across cultures, rather than on exact values.

7 Results

7.1 RQ1: Model Predictions

We show the predicted scores for the XLM-R model in Figure 1. As is clear from the figure, there are substantial differences in the predicted scores for the cultural dimensions across cultures. On average, predicted values for power distance (pdi) are high, whereas ones for masculinity (mas) and indulgence (ivr) are relatively low. The model considers Greece and South Korea as places with high power distance, Pakistan, Germany as more masculine. Indulgence (ivr) has the lowest scores across all values with only Vietnam and Malaysia having positive values, indicating high restraint in these cultures according to the model.

To understand whether LMs can preserve cross-cultural differences in values, we plot the results of the probing for Hofstede’s and WVS’ survey in Figures 2 and 3 respectively. As is visible in these plots, there is a variety in the values, i.e.,

\[\text{Table 1: Predicted value scores on Hofstede’s survey questions per country across the three models. Values were normalized to } [0,1] \text{ using min-max normalisation.}\]

\[
\begin{array}{cccccccc}
\text{Country} & \text{pdi} & \text{idv} & \text{mas} & \text{uai} & \text{lto} & \text{ivr} \\
\hline
\text{Turkey} & 13.602711 & 18.629017 & 12.053948 & -104.650377 & 18.629017 & -26.212504 \\
\text{Philippines} & 69.900500 & 32.454040 & -36.899588 & 68.000574 & 29.341779 & 127.777239 \\
\text{Romania} & 44.320077 & 29.049334 & 1.349567 & -44.142810 & 11.191644 & -98.112277 \\
\text{Malaysia} & 35.838907 & 0.000000 & 0.000000 & 35.838907 & 32.649895 & 45.570506 \\
\text{Korea South} & 86.411911 & -14.096250 & 9.024229 & 43.393946 & 3.383976 & -38.241958 \\
\text{Greece} & 54.289859 & -2.447076 & -2.989553 & 58.395254 & 5.342546 & -95.347674 \\
\text{Iran} & 45.620954 & 28.837315 & 30.049334 & 27.243457 & 40.234542 & 74.347725 \\
\text{Germany} & 87.777131 & 29.341779 & 36.899588 & 68.000574 & 29.341779 & -74.347725 \\
\text{Indonesia} & 39.817128 & 0.000000 & 0.000000 & 39.817128 & 28.297720 & -95.310727 \\
\text{Pakistan} & 94.257019 & -0.935359 & 44.619987 & 154.199166 & 18.802991 & -48.476206 \\
\text{Serbia} & -61.367068 & -56.727120 & -91.214805 & -75.667432 & -7.394542 & -38.726997 \\
\text{Bangladesh} & 53.279862 & 70.191660 & -31.600899 & 36.495058 & 25.490227 & -42.405676 \\
\end{array}
\]

\[\text{Figure 1: Heatmap of scores predicted per value for XLM-R mask probing on Hofstede’s survey questions}\]

\[\text{Figure 2: Predicted value scores on Hofstede’s survey questions per country across the three models. Values were normalized to } [0,1] \text{ using min-max normalisation.}\]

\[\text{Figure 3: Predicted value scores on World Value Survey questions per country across the three models. Values were normalized to } [0,1] \text{ using min-max normalisation.}\]

\[\text{Appendix C.}\]
the models seem to place different importance on different values across cultures, displaying cross-cultural differences in the values. We quantify these differences among the prediction scores by testing for statistical significance between the model’s predictions by culture, seeing how they capture cross-cultural differences. For XLM-R’s predictions for the WVS, 42.31% of the country pairs have a statistically significant difference, meaning the model preserves cross-cultural differences. For the other two models, the share of significantly different country pairs are 51.28% and 46.15% for mBERT and XLM respectively. For XLM-R’s predictions of Hofstede’s survey, only 10.26% of cultures have \( p < 0.05 \). For the other two models, the share of significantly different country pairs are none and 6.41% for mBERT and XLM respectively. We attribute these low percentages to the fact that we conduct the test over the six value dimensions only, while it is on over 200 questions for WVS.

### 7.2 RQ2: Model Agreement

To further study whether scores across values and categories are consistent across the three models, we check for correlation between the predicted scores between the three models and outline them in Tables 2 and 3. We can see that predictions are not consistent across the models, indicating differences in the cross-cultural values models pick up on. mBERT and XLM share the same architecture and are both trained on Wikipedia, yet the correlations across values are low, indicating the large effect that changes to the model training can have on the cultural values picked up by the model, such as Next Sentence Prediction, and the truncation of the input sentences.

### 7.3 RQ3: Alignment with Surveys

Finally, we investigate whether the models’ predictions for the value questionnaire are consistent with existing value survey scores.

**Hofstede** We outline the results of correlations between each of the models’ predictions for mask probing per value in Table 4. We find no statistically significant alignment between the models’ predictions and survey value scores provided by Hofstede but given the low sample size, this is to be expected (Sullivan and Feinn, 2012). We also check for correlations between the predicted values
Table 3: Pairwise correlations in model predictions for masked probing on WVS survey questions. Statistically significant values with p <= 0.05 are marked with *.

Table 4: Correlation per value between masked prediction scores and Hofstede’s value survey. Statistically significant values with p <= 0.05 are marked with *

Table 5: Correlation per country between masked prediction scores and Hofstede’s value survey. Statistically significant values with p <= 0.05 are marked with *

We find weak correlations among some of the values between the models’ predicted scores and the value survey suggesting the disparity in cultural values outlined by Hofstede and the ones picked up by PTLMs.

WVS Tables 6 and 7 similarly show the correlations between the models’ predicted scores and the world values survey scores per category and per country respectively. Here too, we find no statistically significant correlation between the predicted and the survey scores outlining the difference in values picked up by the language models and those quantified in the surveys.

8 Discussion

Our experiments show that there are sizable differences in the cultural values picked up by the different multilingual models which are widely used for a number of language tasks, even when they are trained on data from the same source. This is in line with previous results (Stanczak et al., 2021) and hints at the sensitivity of model design, training choices and their downstream effect on models biases towards certain groups. While the values picked up by the models vary, the bias in the models are not in line with values outlined in existing large scale value surveys of people from different cultures. A possible reason for this is lack of diversity in training data. Wikipedia articles in different languages are written by a small subset of editors that are not representative of the populations in those countries. This shows the need for including more diverse and emic sources of data, in order for the models to better reflect the cultural values of those populations. Joseph et al. (2021) suggest that people express themselves differently online on Twitter compared to survey responses. This is another potential reason for this mis-alignment.

Pre-Trained Language Models are used for a variety of different NLP tasks in different countries and hence to accommodate the usage of a variety of people from diverse backgrounds and cultures, it is not just important to have linguistic and typological diversity in training data, but also cultural diversity (Liu et al., 2021).

9 Limitations

There are several limitations of our approach in trying to assess cultural diversity and alignment.
<table>
<thead>
<tr>
<th></th>
<th>mBERT</th>
<th>XLM</th>
<th>XLM-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science and Technology</td>
<td>0.50</td>
<td>0.09</td>
<td>0.19</td>
</tr>
<tr>
<td>Security</td>
<td>0.38</td>
<td>-0.22</td>
<td>0.09</td>
</tr>
<tr>
<td>Social Values, Attitudes and Stereotypes</td>
<td>-0.34</td>
<td>-0.30</td>
<td>-0.07</td>
</tr>
<tr>
<td>Political Culture and Regimes</td>
<td>0.29</td>
<td>0.15</td>
<td>-0.05</td>
</tr>
<tr>
<td>Political Interest and Political Participation</td>
<td>0.25</td>
<td>0.02</td>
<td>0.10</td>
</tr>
<tr>
<td>Migration</td>
<td>0.19</td>
<td>0.26</td>
<td>0.21</td>
</tr>
<tr>
<td>Social Capital, Trust and Organisational Member...</td>
<td>0.17</td>
<td>0.06</td>
<td>-0.38</td>
</tr>
<tr>
<td>Religious Values</td>
<td>0.14</td>
<td>0.13</td>
<td>-0.37</td>
</tr>
<tr>
<td>Corruption</td>
<td>0.07</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Happiness and Well-being</td>
<td>-0.07</td>
<td>0.37</td>
<td>0.07</td>
</tr>
<tr>
<td>Ethical Values and Norms</td>
<td>0.04</td>
<td>-0.02</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 6: Correlation per question between masked prediction scores and Hofstede’s value survey. Statistically significant values with p <= 0.05 are marked with *

<table>
<thead>
<tr>
<th></th>
<th>mBERT</th>
<th>XLM</th>
<th>XLM-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greece</td>
<td>0.78*</td>
<td>-0.26</td>
<td>0.01</td>
</tr>
<tr>
<td>Philippines</td>
<td>0.67*</td>
<td>0.53</td>
<td>0.36</td>
</tr>
<tr>
<td>Turkey</td>
<td>-0.56</td>
<td>-0.34</td>
<td>-0.86*</td>
</tr>
<tr>
<td>Malaysia</td>
<td>0.44</td>
<td>0.31</td>
<td>0.23</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>0.43</td>
<td>-0.36</td>
<td>0.10</td>
</tr>
<tr>
<td>Vietnam</td>
<td>0.28</td>
<td>0.46</td>
<td>0.26</td>
</tr>
<tr>
<td>Iran</td>
<td>-0.24</td>
<td>-0.43</td>
<td>-0.09</td>
</tr>
<tr>
<td>Korea South</td>
<td>-0.20</td>
<td>-0.36</td>
<td>-0.06</td>
</tr>
<tr>
<td>Romania</td>
<td>0.20</td>
<td>0.14</td>
<td>-0.18</td>
</tr>
<tr>
<td>Indonesia</td>
<td>0.18</td>
<td>0.34</td>
<td>0.03</td>
</tr>
<tr>
<td>Germany</td>
<td>0.13</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>Serbia</td>
<td>0.03</td>
<td>-0.01</td>
<td>-0.14</td>
</tr>
<tr>
<td>Pakistan</td>
<td>-0.01</td>
<td>0.17</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Table 7: Correlation per country between masked prediction scores and world value survey. Statistically significant values with p <= 0.05 are marked with *

of the values picked up by PTLMs. While our methodology of probing models using cloze style questions gives us some insight into token level biases picked up by the language models, it is limited in its approach to only show static and extrinsic biases at inference time using output probabilities. There are intrinsic measures for quantifying bias, but those do not always correlate with extrinsic measures (Goldfarb-Tarrant et al., 2021). As discussed earlier, a major limitation that comes with quantifying cultural values is the mapping of countries to cultures and in our case, also to languages. Since this is an imperfect mapping, it is a difficult task to accurately quantify and assess cultural bias and values embedded in the models. We partially addressed this by restricting our study to languages which are mostly geographically restricted to one country. This is a limitation faced by cross-cultural research in general, where countries are often used as surrogates for cultures (Nasif et al., 1991). Finally, surveys and aggregate responses are also imperfect tools to evaluate and quantify cultural disparity, though the best ones currently in use. They are tasked with collapsing individual values into a set of questions. Individuals answering those questions from different backgrounds may perceive the questions differently. Further, there are several confounding factors affecting the survey responses and problems relating to seeing populations as a monolithic homogeneous whole. While these limitations pose important questions around how one should be careful in interpreting these values, we believe our study makes important contributions and provides a first step in assessing alignment between PTLMs and cultural values, which we argue is necessary for models to faithfully work in a cross-cultural context.

10 Conclusion

In this study, we propose a methodology for probing of cultural values embedded in multilingual pre-trained language models and assessing differences among them. We measure alignment of these values amongst the models and with existing value surveys and theories. We find that PTLMs capture marked differences in values between cultures, though these in turn are only weakly correlated with values surveys. Besides the training data, we discuss the impact training and modelling choices
can have on cultural bias picked up by the models. We further discuss the importance of this alignment when developing models in a cross-cultural context and offer suggestions for more inclusive ways of diversifying training data to incorporate these values.

Acknowledgements

The authors thank Pepa Atanasova for taking part in preliminary discussions. We would further like to thank Dustin Wright and Zeerak Talat for providing constructive feedback on the paper. This research is funded by DFF Sapere Aude grant No 0171-00034B ‘Learning to Explain Attitudes on Social Media (EXPANSE)’.

References


Michael I. Jones. 2007. Hofstede-culturally questionable?


Table 8: Hofstede Ablation: Correlation and p-values of the three different model for simple probing

<table>
<thead>
<tr>
<th></th>
<th>mBERT</th>
<th>XLM</th>
<th>XLM-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>mas</td>
<td>0.48</td>
<td>-0.18</td>
<td>0.06</td>
</tr>
<tr>
<td>uai</td>
<td>0.37</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>idv</td>
<td>-0.34</td>
<td>0.51</td>
<td>0.10</td>
</tr>
<tr>
<td>ivr</td>
<td>-0.25</td>
<td>0.51</td>
<td>0.39</td>
</tr>
<tr>
<td>lto</td>
<td>-0.13</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>pdi</td>
<td>-0.02</td>
<td>-0.05</td>
<td>-0.11</td>
</tr>
</tbody>
</table>

Table 9: WVS Ablation: Category wise correlation of the three different model for simple probing using only high response token

<table>
<thead>
<tr>
<th>Category</th>
<th>mBERT</th>
<th>XLM</th>
<th>XLM-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>0.51</td>
<td>0.40</td>
<td>-0.13</td>
</tr>
<tr>
<td>Social Val</td>
<td>-0.44</td>
<td>-0.50</td>
<td>0.16</td>
</tr>
<tr>
<td>Political Cul</td>
<td>0.43</td>
<td>0.33</td>
<td>0.08</td>
</tr>
<tr>
<td>Corruption</td>
<td>0.39</td>
<td>0.42</td>
<td>-0.11</td>
</tr>
<tr>
<td>Ethical</td>
<td>-0.24</td>
<td>0.10</td>
<td>0.29</td>
</tr>
<tr>
<td>Religious</td>
<td>-0.16</td>
<td>-0.06</td>
<td>0.36</td>
</tr>
<tr>
<td>Migration</td>
<td>0.14</td>
<td>0.17</td>
<td>0.08</td>
</tr>
<tr>
<td>Political Int</td>
<td>0.06</td>
<td>0.16</td>
<td>-0.21</td>
</tr>
<tr>
<td>Security</td>
<td>-0.06</td>
<td>-0.09</td>
<td>-0.12</td>
</tr>
<tr>
<td>Happiness</td>
<td>-0.06</td>
<td>-0.09</td>
<td>0.21</td>
</tr>
<tr>
<td>Social Cap</td>
<td>-0.06</td>
<td>-0.55*</td>
<td>0.22</td>
</tr>
</tbody>
</table>


A Models

All models were run in Python using PyTorch (Paszke et al., 2019) and the Transformers library (Wolf et al., 2020). When speaking about XLM-R, mBERT, XLM we refer to the models with the names xlm-roberta-base, bert-base-multilingual-uncased, xlm-mlm-100-1280 respectively.

B Ablations

In order to assess the effect of subtraction of logit for label token with the lower response score in the survey question from the one for higher score, we calculate correlations with just the high response label token $y^+$. We report our results for Hofstede in Table 8 and WVS in Table 9.

C Hofstede Value Calculation

Following are the six formulas used to calculate the values based on the questions provided by Hofstede at [https://geerthofstede.com/research-and-vsm/vsm-2013/](https://geerthofstede.com/research-and-vsm/vsm-2013/).

**Power Distance**

$$pdi = 35(m_{07} - m_{02}) + 25(m_{20} - m_{23}) + C(pd)$$ (1)

**Individualism**

$$idv = 35(m_{04} - m_{01}) + 35(m_{09} - m_{06}) + C(ic)$$ (2)

**Uncertainty Avoidance**

$$mas = 35(m_{05} - m_{03}) + 35(m_{08} - m_{10}) + C(mf)$$ (3)

**Masculinity**

$$uai = 40(m_{18} - m_{15}) + 25(m_{21} - m_{24}) + C(ua)$$ (4)

**Long term orientation**

$$lto = 40(m_{13} - m_{14}) + 25(m_{19} - m_{22}) + C(ls)$$ (5)

**Indulgence**

$$ivr = 35(m_{12} - m_{11}) + 40(m_{17} - m_{16}) + C(ir)$$ (6)