Bridging Cultural Nuances in Dialogue Agents through Cultural Value Surveys

Cao, Yong; Chen, Min; Hershcovich, Daniel

Published in: Findings of the Association for Computational Linguistics: EACL 2024

Publication date: 2024

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

Document license: CC BY

Bridging Cultural Nuances in Dialogue Agents through Cultural Value Surveys

Yong Cao¹,², Min Chen³, Daniel Hershcovitch²
¹Huazhong University of Science and Technology
²Department of Computer Science, University of Copenhagen
³School of Computer Science and Engineering, South China University of Technology
yongcao_epic@hust.edu.cn, minchen@ieee.org, dh@di.ku.dk

Abstract
The cultural landscape of interactions with dialogue agents is a compelling yet relatively unexplored territory. It’s clear that various sociocultural aspects—from communication styles and beliefs to shared metaphors and knowledge—profoundly impact these interactions. To delve deeper into this dynamic, we introduce cuDialog, a first-of-its-kind benchmark for dialogue generation with a cultural lens. We also develop baseline models capable of extracting cultural attributes from dialogue exchanges, with the goal of enhancing the predictive accuracy and quality of dialogue agents. To effectively co-learn cultural understanding and multi-turn dialogue predictions, we propose to incorporate cultural dimensions with dialogue encoding features. Our experimental findings highlight that incorporating cultural value surveys boosts alignment with references and cultural markers, demonstrating its considerable influence on personalization and dialogue quality. To facilitate further exploration in this exciting domain, we publish our benchmark publicly accessible at https://github.com/yongcaoplus/cuDialog.

1 Introduction
Culture can be defined as the combinations of beliefs, norms, and customs among groups (Tomlinson et al., 2014). Implicit cultural cues hinted in dialogue utterances reveal different values and beliefs among speakers, which reflects their way of thinking (Nisbett et al., 2001) and emotions (Almuhailib, 2019; Sun et al., 2021; Ma et al., 2022). While pre-trained language models (PLMs) have shown impressive performance on dialogue tasks (Gu et al., 2021; Liu et al., 2021; Sweed and Shafaf, 2021), their cultural bias in terms of values and their inconsistency in many other cultural aspects (Fraser et al., 2022) has severe implications on the prospect of employing them for interaction with speakers of diverse cultural backgrounds (Hershcovitch et al., 2022). This is particularly crucial in the context of culturally-related topics (Zhou et al., 2023a,b), where acknowledging and understanding cultural differences becomes essential. For example, scholars tend to believe that Eastern societies have a more communal or collective orientation compared to that Western societies (Lomas et al., 2023).

Previous studies in the field of cross-cultural NLP (Arora et al., 2023; Hämmerl et al., 2022; Johnson et al., 2022; Santurkar et al., 2023) have primarily utilized probing methods to study the characteristics of models or agents. For instance, Cao et al. (2023) applied the Hofstede Culture Survey (Hofstede, 1984, see §3) to probe ChatGPT, a prominent dialogue system, revealing a distinct disparity between the system and human society. This underscores the need to enhance dialogue agents’ performance by incorporating cultural dimensions. However, developing culturally adaptive dialogue agents poses a significant challenge due to the scarcity of suitable datasets. While there are available multicultural corpora focused on specific domain tasks such as news (Ma et al., 2022)
and image captions (Liu et al., 2021), there is currently a lack of datasets specifically designed for cross-cultural dialogue tasks.

To address this research gap, we introduce cuDialog, an extensive English-language benchmark for multicultural dialogues. Our benchmark covers 13 cultures and 5 genres, specifically designed to mitigate the impact of linguistic variations and emphasize implicit cultural cues. Within cuDialog, we propose two culture understanding tasks and one dialogue generation task, offering a comprehensive framework for evaluating and advancing cultural understanding in dialogue systems.

Specifically, as depicted in Figure 1, we design several baselines on culture classification and regression tasks, showing that cultural attributes behind dialogues can be identified. We leverage the soft cultural knowledge provided by the Hofstede Culture Survey (Hofstede, 1984), which defines six cultural dimensions to measure the cultural attributes of different countries and provides statistical results for numerous nations. To utilize this external knowledge, we present a novel feature fusion mechanism based on an encoder-decoder generation framework, by considering using culture to assist separability in dialog generation. Experimental results reveal that incorporating cultural value representation can improve alignment with references, indicating better cultural representation.

In summary, our contributions are as follows: (1) We introduce cuDialog, a multicultural dialogue benchmark dataset specifically tailored to different genres, enriched with cultural survey annotations. (2) We develop several baseline models that effectively capture cultural nuances and propose three dialogue tasks. (3) We demonstrate the feasibility of capturing cultural nuances and the impact of incorporating cultural representation into dialogue systems, highlighting the significance of considering cultural differences in dialogue modeling.

### 2 Related Work

**Culture-oriented benchmarks.** Researchers have developed a range of culture-oriented benchmarks to investigate the impact of culture on language understanding and generation tasks. These benchmarks involve collecting and annotating multilingual and multicultural corpora to study cultural effects in downstream tasks. For instance, benchmarks have been introduced for news classification across different countries (Ma et al., 2022) and for analyzing user statements reflecting different cultures using text and images (Liu et al., 2021). Other benchmarks focus on detecting culture differences and user attributes, spanning both small-scale (Sweed and Shahaf, 2021) and large-scale (Qian et al., 2021) datasets. Furthermore, recent works have explored in-domain cross-cultural benchmarks, such as multilingual moral understanding and generation (Guan et al., 2022), and culture-specific time expression grounding (Shwartz, 2022). While Zhang et al. (2022) proposed a multilingual conversation dataset, it lacks cultural annotations.

**Cultural attributes learning.** Traditional approaches for capturing cultural differences often rely on probabilistic models, such as Latent Dirichlet Allocation (Pennacchiotti and Popescu, 2011; Al Zamal et al., 2012; Tomlinson et al., 2014). However, the emergence of unsupervised learning and advancements in pre-trained language models (PLMs) have sparked interest in utilizing PLMs to learn cultural attributes and user profiles (Gu et al., 2021; Fraser et al., 2022).

**Culture-sensitive dialogue agents.** Previous studies (Tomlinson et al., 2014; Ma et al., 2022) have demonstrated the benefits of equipping dialogue agents with an understanding of cultural differences for natural language understanding (NLU) and generation (NLG) tasks, even in general natural language processing tasks. For example, Fu et al. (2022) proposed the use of a persona-specific memory network to jointly encode cultural background and user profiles, enhancing the NLG task for dialogue agents. Kanclerz et al. (2021) introduced personalized approaches that respect individual beliefs expressed through user annotations. Additionally, Wu et al. (2021) incorporated user queries, cultural-related comments, and user profiles as encoded features to generate personalized responses, demonstrating the efficacy of leveraging both features in improving dialogue agent satisfaction. Moreover, leveraging external knowledge by retrieving user-related cultural and attribute documents has shown promising improvements, providing additional guidance for model training (Majumder et al., 2021; Guan et al., 2022). These works collectively highlight the value of incorporating cultural aspects into dialogue systems and leveraging personalized approaches for more effective and satisfactory interactions. While recent efforts have incorporated commonsense knowledge (Varshney et al., 2022)
and socio-cultural norms (Moghimifar et al., 2023) into dialogue agents, these approaches have primarily focused on monocultural settings, neglecting the broader context of multicultural dialogue.

3 Cultural Dimensions

The Hofstede Culture Survey (Hofstede, 1984) identifies six cultural dimensions that capture different aspects of cultural values:

- **Power Distance (pdi):** Reflects the acceptance of unequal power distribution within a society.
- **Individualism (idv):** Measures the level of interdependence versus self-definition within a culture.
- **Masculinity (mas):** Examines the emphasis on competition, achievement, and assertiveness versus caring for others and quality of life.
- **Uncertainty Avoidance (uai):** Deals with response to ambiguity and minimizing uncertainty.
- **Long-Term Orientation (lto):** Describes how cultures balance tradition with future readiness.
- **Indulgence (ivr):** Focusses on the control of desires and impulses based on cultural upbringing.

These dimensions offer valuable insights into the beliefs, behaviors, and attitudes that vary across societies. By incorporating these dimensions in our dataset for the corresponding countries, we provide a benchmark for evaluating the ability of dialogue systems to capture and adapt to cultural nuances. This enables researchers to assess the cultural sensitivity and adaptability of dialogue systems in a standardized manner. The survey results, freely available online for 111 countries, serve as a valuable resource for integrating cultural dimensions into dialogue system enhancement and evaluation.

4 Multicultural Dialogue Dataset

In this section, we introduce the collection, benchmarking, and statistics of our proposed multicultural dataset. The cuDialog dataset contains four components: histories, golden predictions, culture label, and culture dimension scores, serving our proposed tasks, including culture classification, cultural alignment and dialogue generation, etc.

**Data source.** We gather multicultural dialogues from the OpenSubtitles 2018 dataset (Lison et al., 2018), which comprises a vast collection of subtitles extracted from movies and television shows. The OpenSubtitles 2018 dataset offers extensive coverage of multiple languages, providing subtitle data in text format that is well-suited for training and evaluating a diverse range of NLP models.

While the dataset has been widely utilized in language identification (Toftrup et al., 2021), domain adaptation (Thompson et al., 2019; Lai et al., 2022), and machine translation (Costa-jussà et al., 2022; Zhang and Ao, 2022), it is essential to recognize that it also contains substantial cultural cues. To our knowledge, our work represents the first application of this dataset for culture-focused research, complemented by cultural annotations.

**Language and culture selection.** Our research aims to explore the cultural differences underlying linguistic variations. We acknowledge that linguistic variations themselves serve as strong cultural features, which can have an impact on aspects such as common grounding and beliefs. To investigate

---


the cultural cues related to beliefs and values, we conducted an analysis using a subset of 500 randomly extracted samples from the OpenSubtitles dataset. These samples were encoded by mBERT and visualized using the t-SNE method (Van der Maaten and Hinton, 2008), with a specific focus on the representation of data from five distinct countries. The visualization revealed distinct separations in the representation space based on different languages, making it challenging to capture cultural cues beyond linguistic variations. This motivated our decision to utilize English subtitles, as they exhibit less trivial separability (Figure 2a). As a result, our benchmark dataset universally employs English subtitles that encompass all cultures. This approach helps to capture cultural cues beyond linguistic variations, making it challenging for dialogue agents to infer the cultural background effectively. Additionally, we eliminate emojis and remove short contexts and responses that provide limited information, making it challenging for dialogue agents to infer the cultural background effectively. Additionally, we eliminate emojis and address encoding errors to enhance overall quality.

**Dataset statistics.** To facilitate comparative analysis and maintain dataset balance, we ensure a consistent number of movies across different genres. Table 1 presents an overview of the cuDialog dataset’s statistics. Each genre comprises approximately 130 to 160 thousand dialogues, with a total of 923 movies and an average sentence length of around 71, considering both the input histories and prediction references. The dataset is divided into train (80%), validation (10%), and test (10%) sets, with no overlap between movies in the test set and those in the train set. This partitioning is performed at the movie and television show level, enabling dialogue-related tasks.

<table>
<thead>
<tr>
<th>Set</th>
<th>Action</th>
<th>Comedy</th>
<th>Drama</th>
<th>Romance</th>
<th>Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>108,934</td>
<td>137,475</td>
<td>109,534</td>
<td>114,467</td>
<td>112,695</td>
</tr>
<tr>
<td>Dev</td>
<td>12,808</td>
<td>17,361</td>
<td>13,256</td>
<td>13,697</td>
<td>14,313</td>
</tr>
<tr>
<td>Test</td>
<td>15,336</td>
<td>18,213</td>
<td>16,179</td>
<td>15,705</td>
<td>16,853</td>
</tr>
</tbody>
</table>

The cuDialog dataset is represented as \( \{d_i \in D \mid d_i = (Q_i, R_i, C_i, S_i)\} \). An example of a dialogue in the cuDialog format is presented in Table 2. To ensure data quality, we remove short contexts and responses that provide limited information, making it challenging for dialogue agents to infer the cultural background effectively. Additionally, we eliminate emojis and address encoding errors to enhance overall quality.

**Dialogue format.** The cuDialog dataset is represented as \( \{d_i \in D \mid d_i = (Q_i, R_i, C_i, S_i)\} \). An example of a dialogue in the cuDialog format is presented in Table 2. To ensure data quality, we remove short contexts and responses that provide limited information, making it challenging for dialogue agents to infer the cultural background effectively. Additionally, we eliminate emojis and address encoding errors to enhance overall quality.

### Table 2: A Romance genre example from cuDialog with four fields: multi-turn history, golden predictions, culture category, and cultural value dimension scores.

<table>
<thead>
<tr>
<th>History: His mortal flesh belonged to the fire, his immortal soul to the flames of Hell.</th>
<th>A gag blocked his mouth.</th>
</tr>
</thead>
<tbody>
<tr>
<td>You’d have thought it was a corpse being led to its grave, “yet it was a living man whose torturers were to gruesomely entertain the people.”</td>
<td>Pardon. I'll break off here.</td>
</tr>
<tr>
<td>Golden Predictions: Will you amuse us now with details of an execution during the Inquisition?</td>
<td>No, I beg your pardon. I’m deeply impressed.</td>
</tr>
<tr>
<td>Culture: Germany.</td>
<td></td>
</tr>
</tbody>
</table>

**Culture Score:** 35, 67, 66, 65, 83, 40.

**Tokens**

<table>
<thead>
<tr>
<th>Vocab</th>
<th>34,883</th>
<th>36,292</th>
<th>32,566</th>
<th>32,666</th>
<th>33,416</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Avg</td>
<td>71.15</td>
<td>70.32</td>
<td>72.38</td>
<td>71.42</td>
<td>72.05</td>
</tr>
</tbody>
</table>

### Table 1: The statistics of cuDialog. Here we split train, dev and test set by movies to avoid data leakage. #Avg is the average number of tokens by mT5 tokenizer. Vocab is the total vocabulary size.

As a result, our benchmark dataset universally em-
5 Cultural Dialogue Tasks

Drawing from the insights gained from previous research (Arora et al., 2023; Cao et al., 2023), which highlighted the challenges faced by pre-trained models and dialogue agents in capturing cultural differences, we aim to analyze cultural attributes and explore effective mechanisms for cultural alignment. We pose the following research questions:

- **RQ1**: Can our cuDialog dataset effectively capture and identify cultural dimensions?
- **RQ2**: How do cultural nuances impact the performance of dialogue agents across cultures?

To address these research questions, we introduce three dialogue tasks, depicted in Figure 4.

To address RQ1, we go beyond the conventional approach and examine whether the dialogues in cuDialog exhibit discernible cultural differences that can be effectively classified. Our first task, **culture classification**, delves into the identification of cultural variations in the dataset. Additionally, we explore the **cultural dimension score regression** task to investigate the feasibility of inferring fine-grained cultural labels. These tasks necessitate capturing cross-cultural differences and exploit the multicultural variety of cuDialog.

To tackle RQ2, we propose a **multi-turn dialogue prediction** task based on the hypothesis of cultural separability. By incorporating cultural features into the dialogue agent framework, we aim to enhance the performance of dialogue agents by considering the influence of cultural nuances. This task provides valuable insights into how culture impacts dialogue systems and sheds light on the role of cultural factors in improving the overall performance and adaptability of dialogue agents.

5.1 Culture Classification

In the culture classification task, depicted in Figure 4(a), the goal is to predict the correct culture label $C_i$ among the 13 countries, given a dialogue history $Q_i$ and golden prediction $R_i$. The task involves predicting $\mathcal{P}_c(c|h_i, r_i)$, where $c \in C_i$, $h_i \in Q_i$, and $r_i \in R_i$. Notably, the input contains the query and response as a combined context. We specifically choose the multi-turn dialogue format instead of single-turn dialogues due to the short and limited information present in OpenSubtitles sentences. By ensuring longer text, we aim to capture and learn the cultural cues effectively. This task can be modeled using encoder-only models and does not involve generation or address cultural dimensions.

5.2 Cultural Dimension Regression

In cultural dimension regression, we leverage the cultural dimensions obtained from the Hofstede Culture Survey (§3) as fine-grained cultural labels. As depicted in Figure 4(b), we employ a regression layer that operates on the encoder hidden states to predict the six-dimensional cultural scores for each dialogue. Specifically, we aim to predict $\mathcal{P}_c(\hat{c}|h_i)$, where $\hat{c}$ represents the six-dimensional cultural vectors and $\mathcal{P}_c$ denotes the prediction. In this task we use only the history text instead of concatenating the history and golden predictions. This adjustment allows us to effectively capture the cultural dimensions and assess their impact on dialogue systems’ performance, using encoder-decoder models.

5.3 Multi-Turn Dialogue Prediction

Culture plays a crucial role in dialogue generation, as it influences the choice of words, expressions, and behaviors in conversations. To capture the cultural nuances and ensure culturally appropriate
responses, we propose a multi-turn dialogue prediction task that incorporates cultural value representations. In our approach, we utilize the cultural dimensions (§3) as representations of cultural values. These dimensions serve as contextual cues that guide the dialogue generation process by integrating them into the encoder-decoder framework.

In this task, we employ an encoder-decoder framework, where the encoder processes the dialogue history \( h_i \) to obtain the hidden states \( H^{(1)}, \ldots, H^{(L)} \). We consider the cultural dimensions \( \hat{c} \) obtained from a culture regression model (§5.2) as representations of cultural values. To incorporate these dimensions into the dialogue generation process, we extend each dimension to match the length of the hidden states, resulting in \( \hat{c}_d \). We concatenate \( \hat{c}_d \) with the hidden states at each layer:

\[
H^{(1)}_d, \ldots, H^{(L)}_d = \mathcal{D}(H^{(1)}, \ldots, H^{(L)}, \hat{c}_d) \tag{1}
\]

Finally, the decoder generates the predicted response by utilizing the concatenated hidden states.

This approach requires the model to consider cultural dimensions, ensuring that the generated responses align with the underlying cultural values.

6 Experiments

6.1 Evaluated Models

To extensively evaluate the performance of currently available models, we select various models for evaluation, encompassing both encoder and encoder-decoder frameworks, as well as monolingual and multilingual models. Specifically, we evaluate the following baselines for culture classification tasks: BERT (Devlin et al., 2019), multilingual BERT, RoBERTa (Liu et al., 2019), and XLM-RoBERTa (Conneau et al., 2020). For the culture regression task, we evaluate T5 (Raffel et al., 2020), mT5 (Xue et al., 2021), BART (Lewis et al., 2020), and mBART50 (Tang et al., 2020). For dialogue prediction, we evaluate mT5 on five genres.

6.2 Experimental Setup

Using pre-trained models from HuggingFace (Wolf et al., 2020), we use one A100 GPU for culture classification and regression and two A100 GPUs for multi-turn dialogue prediction. As hyperparameters, we set the batch size to 128, 256, and 64 for culture classification, regression, and prediction tasks, respectively. We use an early stopping strategy with a patience of 2 or 3. For generation, we employ beam search with a width of 3, temperature of 0.7, and repetition penalty of 1.2.5

6.3 Evaluation Metrics

The evaluation metrics used in our study depend on the task at hand. For classification tasks, we employ recall, precision, and F1 score. Regression tasks are evaluated using the Spearman correlation coefficient, R2 score, and root mean squared error (RMSE). For generation, we use BLEU measuring n-gram overlap, ROUGE-L considering the longest common subsequence, BERTScore assessing similarity using contextualized embeddings, and Distinction evaluating distinctiveness in terms of diversity and uniqueness. These metrics align with the approach proposed by Zhang et al. (2022).

6.4 Main Results

Culture Classification. Table 3 presents the results for culture classification, comparing the performance of monolingual models (BERT and RoBERTa) with multilingual models (mBERT and XLM-R).6 Interestingly, we observe that the monolingual models demonstrate superior performance in this task, suggesting a slight disadvantage for multilingual models within the context of an English corpus encompassing all cultures. It is noteworthy that the action and crime genres exhibit a higher suitability for culture classification, aligning with our expectations. This can be attributed to the significant cultural variations in the interpretation of criminal activities, such as the legality of firearm possession (Boine et al., 2020).

In contrast, the comedy corpus performs relatively poorly in culture classification, which can be attributed to the challenges of translation. Prior research (Jiang et al., 2019) has indicated the existence of cultural differences in humor usage between Eastern and Western societies. Western cultures tend to associate humor with positivity and view it as a natural form of amusement expression (Martin and Ford, 2018), whereas Eastern cultures often hold contrasting attitudes towards humor (Dong Yue, 2010). However, we contend that during the translation process, a significant number of comedic elements lose their impact, resulting in diminished distinction for the models.

Culture Regression. Table 4 presents the results for culture regression using T5, mT5, BART and

5More details for reproducibility are in Appendix D.
6Additional scores for each culture are in Appendix F.
Table 3: F1 scores of dialogue culture classification models for 13 cultural categories. The English-only models RoBERTa and BERT outperform the multilingual models mBERT and XLM-R.

<table>
<thead>
<tr>
<th>Model</th>
<th>Action</th>
<th>Comedy</th>
<th>Drama</th>
<th>Romance</th>
<th>Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa</td>
<td>87.93</td>
<td>75.43</td>
<td>82.39</td>
<td>83.20</td>
<td>85.29</td>
</tr>
<tr>
<td>XLM-R</td>
<td>86.50</td>
<td>75.39</td>
<td>80.69</td>
<td>79.29</td>
<td>84.27</td>
</tr>
<tr>
<td>BERT</td>
<td>88.49</td>
<td>76.80</td>
<td>83.77</td>
<td>82.60</td>
<td>85.70</td>
</tr>
<tr>
<td>mBERT</td>
<td>86.05</td>
<td>76.26</td>
<td>82.62</td>
<td>80.48</td>
<td>81.21</td>
</tr>
</tbody>
</table>

Table 4: Regression results aligned with human society surveys. Statistically significant values with $p \leq 0.001$ are marked with *. All correlations of multilingual models are positive and outperform monolingual.

<table>
<thead>
<tr>
<th>Method</th>
<th>Action</th>
<th>Comedy</th>
<th>Drama</th>
<th>Romance</th>
<th>Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman correlations (COR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>0.8849*</td>
<td>0.1002*</td>
<td>0.7432*</td>
<td>0.0848*</td>
<td>0.0784*</td>
</tr>
<tr>
<td>mT5</td>
<td>0.7776</td>
<td>0.0763</td>
<td>0.7361</td>
<td>0.0759</td>
<td>0.7825*</td>
</tr>
<tr>
<td>BART</td>
<td>0.6506</td>
<td>0.5229</td>
<td>0.5994</td>
<td>0.4697</td>
<td>0.5810</td>
</tr>
<tr>
<td>mBART</td>
<td>0.7031</td>
<td>0.6094</td>
<td>0.6383</td>
<td>0.5599</td>
<td>0.8725*</td>
</tr>
<tr>
<td>Coefficient of Determination ($R^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>0.52</td>
<td>0.1045</td>
<td>0.0970</td>
<td>0.0942</td>
<td>0.1088</td>
</tr>
<tr>
<td>mT5</td>
<td>0.56</td>
<td>0.0750</td>
<td>0.0928</td>
<td>0.0943</td>
<td>0.1116</td>
</tr>
<tr>
<td>BART</td>
<td>0.63</td>
<td>0.1043</td>
<td>0.0963</td>
<td>0.0942</td>
<td>0.1116</td>
</tr>
<tr>
<td>mBART</td>
<td>0.70</td>
<td>0.1043</td>
<td>0.0963</td>
<td>0.0942</td>
<td>0.1116</td>
</tr>
<tr>
<td>Root Mean Squared Error (RMSE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>0.2218</td>
<td>0.2196</td>
<td>0.2180</td>
<td>0.2218</td>
<td>0.2219</td>
</tr>
<tr>
<td>mT5</td>
<td>0.2571</td>
<td>0.2443</td>
<td>0.2113</td>
<td>0.2154</td>
<td>0.1364</td>
</tr>
<tr>
<td>BART</td>
<td>0.2190</td>
<td>0.2195</td>
<td>0.2192</td>
<td>0.2217</td>
<td>0.2222</td>
</tr>
<tr>
<td>mBART</td>
<td>0.1002</td>
<td>0.1239</td>
<td>0.1079</td>
<td>0.1089</td>
<td>0.1044</td>
</tr>
</tbody>
</table>

7 Discussion

In our investigation regarding culture identification, we strive to explore the extent to which models can effectively capture cultural attributes within the context of cuDialog (RQ1). Additionally, we examine the integration of these identified cultural attributes into the demanding task of multi-turn dialogue prediction, thereby yielding outcomes that are both more satisfactory and diverse. This empirical analysis provides compelling evidence that incorporating cultural considerations can improve the performance of dialogue agents, validating the notion that cultural awareness plays a crucial role in enhancing their effectiveness (RQ2).

Multi-Turn Dialogue Prediction. Table 5 presents the results of our proposed cultural enhancement approach for multi-turn dialogue prediction. Pre-trained models without fine-tuning on cuDialog mBART$_{25}$ and mT5$_{25}$ exhibit weaker capabilities in dialogue prediction, resulting in lower values and shorter sentence length than fine-tuned models mBART$_{99}$ and mT5$_{99}$. This can be attributed to their pre-training tasks, which primarily focus on machine translation rather than dialogue generation. However, with cultural enhancement mBART$_{cul}$ and mT5$_{cul}$, dialogue prediction on most genres achieves better alignment with references and produces more diverse results, as evidenced by enhancements in both BLEU and Distinction metrics. Thus, it can be inferred that integrating cultural dimensions into dialogue agents leads to enhanced performance across various genres. Despite the improvements observed, there is still a need for further enhancement to improve the model’s ability to comprehend and generate coherent responses in long-term dialogues, as supported by lower BLEU values consistent with prior work.

Furthermore, we can find that the outcomes of mBART align consistently with that of mT5 model, which demonstrate enhanced metrics across the Action, Comedy, Drama, and Crime genres, except for Romance. Notably, improvements on mBART is more significant than mT5, which is consistent with the regression task in Table 4. Our findings confirm the effectiveness of our cultural enhancement approach in improving dialogue prediction, aligning with references. To illustrate how the cultural attributes boost model performance, we provide the illustrative example of our generation results of mBART in Appendix C.
Table 5: Prediction results for the multi-turn dialogue prediction task, demonstrating the impact of our proposed cultural enhancement on various genres. It reveals improvements in four genres, while one genre experienced a decrease. #Avg is the average number of tokens by mT5 tokenizer. Vocab is the total vocabulary size.

![Image](image.png)

Figure 5: mT5 t-SNE before (left) and after (right) fine-tuning on regression. For clarity, we only select five countries as an example.

Figure 6: mBERT classification results, revealing clear distinctions in the classification capabilities of models across different cultures and genres.

8 Conclusion

We introduced cuDialog, utilizing OpenSubtitles 2018 for cultural identification and enhancing dialogue tasks. Our approach goes beyond dialogue texts by introducing culture classification and regression tasks, capturing both coarse-grained and fine-grained cultural knowledge. By leveraging cues from cultural value surveys, we bridge the cultural gap that requires empirical investigation (Hershcovich et al., 2022). To validate our perspective, we randomly select 2,500 samples from five distinct cultures and visualize their representations using t-SNE based on the mT5 model. Figure 5(a) shows that zero-shot models struggle to differentiate between different cultural cases effectively. However, after fine-tuning the models with cuDialog, Figure 5(b) demonstrates a significant improvement in the separability of the representations. This indicates that incorporating cultural dimensions as guidance during fine-tuning facilitates the injection of implicit cultural features into language models.

Cultural cues in cuDialog. Figure 6 illustrates the significant variation in mBERT F1 scores for classification across cultures and genres. Notably, mBERT demonstrates a strong ability to identify American, Australian, and Canadian cultures, with particularly high performance in identifying American culture. These findings align with previous studies (Arora et al., 2023; Cao et al., 2023). The dominance of the English training corpus (Ouyang et al., 2022) contributes to a strong cultural embedding that may overshadow other cultures. Interestingly, Crime and Action dialogues consistently exhibit strong classification across all cultures, indicating a strong cultural component in these genres. This highlights the presence of cultural cues in cuDialog, resulting in distinct cultural representations.
cultural nuances between dialogue agents and human society, resulting in effective dialogue prediction adaptation. Further research in this area will advance the design of culturally aware dialogue systems that better meet user expectations.

**Limitations**

While our work has achieved good performance and shown promising results in enhancing dialogue tasks through incorporation of cultural cues, there are still limitations in our work.

The reference-based approach for multi-turn dialogue prediction evaluation is limited due to the subjectivity and variability of the task. A coherent and appropriate continuation may receive low scores simply because it diverges from the single reference in our dataset.

The OpenSubtitles 2018 English corpus we used has inherent artifacts as it is a combination of human translations and machine-generated translations. Although we acknowledge that human translations tend to adapt to target cultures, we believe that distinct cultural differences can still be captured based on our observations.

Furthermore, we recognize that determining the cultural norm to align with remains an unresolved issue, as extensively discussed in Gabriel (2020). Our approach continues to be grounded in the premise that Chatbots should align to meet the needs of the majority of users, thereby aligning with individuals from diverse cultural backgrounds.

We adopt human survey dimensions as cultural representations, despite its extensively aligned with human society, the intensity of the intervention is relatively soft. However, we believe that this study is still useful in highlighting the challenges of boosting the performance of dialogue agents by cultural considerations. In the future, we plan to explore the feasibility of collecting paired multicultural dialogues from conversation bots and utilizing structural cultural knowledge to guide the adaptation of cultural dialogues, which can be potential to provide further insights into incorporating cultural understanding into dialogue systems.

**Ethics Statement**

Given the current gap in cross-cultural dialogue datasets within existing research, we have proposed constructing such datasets using existing dialogue corpora. However, obtaining paired cultural annotations for each dialogue presents a unique and open challenge, especially for benchmarking purposes. Ensuring the quality and accuracy of our multicultural dataset is crucial.

Our cultural dimension scores are derived from survey results obtained from a comprehensive sample of 117,000 matched employees across various countries, encompassing all the cultures of interest in our study. Furthermore, in terms of genre labels, we utilize the annotations provided by OpenSubtitles, which are included in the original resource and annotated by its creators. Our utilized datasets, including OpenSubtitles and Hofstede Cultural Survey, are publicly available and do not raise any privacy concerns. We have maintained the integrity of the data and adhered to privacy standards by not introducing any additional corpus or cultural annotations. The OpenSubtitles is released with the GNU General Public License v3.0. We will release our processed version with the same license.

We acknowledge that our analysis is based on the assumption that language accurately represents culture. However, we recognize that this notion may not be entirely congruent, as culture is complex, dynamic and highly diverse within countries and languages. This is especially true in cases where multiple official languages exist in a country, or where a language is spoken in multiple countries. Despite this limitation, our research still holds value as we identify a promising combination of existing corpora for our work.

Despite the above ethical considerations, this paper represents one of the initial endeavors in addressing cultural identification and cross-cultural dialogue enhancement, making it a pioneering effort in exploring the cultural adaptability of dialogue agents. We believe this research direction has the potential to mitigate cultural biases and facilitate honest, respectful and informative cross-cultural communication between humans, with the assistance of AI.

**9 Acknowledgement**

Thanks to the anonymous reviewers for their helpful feedback. The authors gratefully acknowledge financial support from China Scholarship Council. (CSC No. 202206160052).

---

7 https://en.wikipedia.org/wiki/Hofstede%27s_cultural_dimensions_theory
9 https://www.gnu.org/licenses/gpl-3.0.en.html
References


Xiao Dong Yue. 2010. Exploration of chinese humor: Historical review, empirical findings, and critical reflections.


Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?


Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. 2022. Mdla: A benchmark for
multilingual dialogue generation in 46 languages. 


A Hofstede Cultural Survey

This survey is one of the most commonly used cross-cultural tools developed by Dutch social psychologist, Geert Hofstede, aiming to measure cultural distinctions among countries. Six cultural dimensions are proposed by this survey, including:

- **Power Distance (pdi).** It measures the acceptance of unequal power distribution within organizations and institutions.
- **Individualism (idv).** It explores the extent to which individuals are integrated into groups.
- **Uncertainty Avoidance (uai).** It assesses the individuals’ attitude to something unexpected, unknown, or away from the status quo.
- **Masculinity (mas).** It measures individuals’ preference in society for achievement, heroism, assertiveness, and material rewards for success.
- **Long-term Orientation (lto).** It measures the focus on traditions and steadfastness (short-term) versus adaptability and pragmatic problem-solving (long-term).
- **Indulgence (ivr).** It measures the degree of societal norms in allowing individuals to freely fulfill their desires.

This survey will ask participants to answer 24 questions and drive each dimension scores \( S_i \) based on four related questions \( Q_j \) by:

\[
S_i = \lambda_i^0(Q_i^0 - Q_i^1) + \lambda_i^1(Q_i^2 - Q_i^3) + C_i \tag{2}
\]

where \( \lambda_i \) is the hyper-parameter and \( C_i \) is a constant. Detailed values for \( \lambda_i \) and \( Q_i \) are listed in Table 6. The results of our used cultures are listed in Table 7. Besides, given Hofstede scores, we tabulated all the cases in our proposed cuDialog in Table 11.

B Significance Check

To ascertain the non-trivial nature of our experimental findings, we pass our experiment results of multi-turn dialogue prediction task through a statistical significance test, aiming to show the effectiveness of our improvements. To achieve this, we have employed a widely recognized tool as outlined in Dror et al. (2018a) and Dror et al. (2018b). Specifically, we format our predictions of each case and baseline’s as required by Dror et al. (2018b)
and then conduct Anderson-Darling (ad) with the desirable significance level (alpha=0.05) and t-test. By comparing the BLEU metrics derived from the aforementioned mBART generation table, we have obtained the results presented in Table 8 (Yes denotes significant, Not denotes not significant). Notably, a substantial portion of the BLEU scores exhibit statistical significance when compared to the baseline outcomes.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Anderson-Darling BLEU-1</th>
<th>Anderson-Darling BLEU-2</th>
<th>Anderson-Darling BLEU-4</th>
<th>t-test BLEU-1</th>
<th>t-test BLEU-2</th>
<th>t-test BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comedy</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Drama</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Romance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Crime</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Action</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 8: Statistical significance test for multi-turn dialogue prediction task in mBART model, where ✓ denotes significant and × denotes not significant.

**C Case Study**

To illustrate how the cultural attributes boost model performance, we present an illustrative example within the Crime genre test set below, which is from an Italian film titled "Pasolini Un Delitto Italiano - Marco Tullio Giordana (1995)", as is shown in Table 9.

In this instance, it is evident that the history sentences are talking about politics and crimes and the attitude of golden sentences is “refusing to tell the truth”, whose topic and attitude aligns closely with our generated sentences, but quite different from the baseline model’s output. Moreover, for the Romance genre, studies show that romance can indeed deviate from accepted social and cultural norms, agreeing with our obtained results. For example, Asian romance movies often challenge values such as obeying authority, adherence to cultural norms and putting society before self (Rahman, 2013).

**D Hyper Parameter Setting**

To facilitate the reproducibility of our training process for culture classification, culture regression, and multi-turn dialogue prediction tasks, we provide a comprehensive list of the hyper-parameters used to achieve the best results on our proposed datasets, as demonstrated in Table 10.

**E Pre-trained Models Download**

All BERT pre-trained models adopted in Table 3, 4 and 5 are published by (Wolf et al., 2020). In order to help reproduce our work and use our code easily, we summarize the download links of the pre-trained models as follows.

**Culture Classification.**
- BERT
  https://huggingface.co/bert-base-uncased
- multilingual BERT
  https://huggingface.co/bert-base-multilingual-cased
- RoBERTa
  https://huggingface.co/roberta-base
- XLM-RoBERTa
  https://huggingface.co/xlm-roberta-base

**Culture Regression & Dialogue Prediction.**
- T5
  https://huggingface.co/t5-base
- mT5
  https://huggingface.co/mt5-base

Table 9: Case study for multi-turn dialogue prediction.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Classification</th>
<th>Regression</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate</td>
<td>3e⁻⁵</td>
<td>1e⁻⁴</td>
<td>1e⁻⁴</td>
</tr>
<tr>
<td>Batch size</td>
<td>128</td>
<td>128</td>
<td>64</td>
</tr>
<tr>
<td>Epochs</td>
<td>30</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>Num Labels</td>
<td>13</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>GPU Num</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Warmup Steps</td>
<td>156</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Early Stop</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Early Stop Patience</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Repetition Penalty</td>
<td>-</td>
<td>-</td>
<td>1.2</td>
</tr>
<tr>
<td>Num Beams</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 10: The hyper-parameter settings of the best results on our proposed three tasks.
Table 11: Detailed statistics of cuDialog, consisting of 13 cultural backgrounds and 5 conversation genres. The dataset includes movie subtitles between individuals from different cultures discussing various genres such as comedy, romance, etc. The 13 cultural backgrounds represented in the dataset include but are not limited to American, Chinese, Indian, and Japanese cultures.

<table>
<thead>
<tr>
<th>Culture</th>
<th>Action</th>
<th>Comedy</th>
<th>Drama</th>
<th>Romance</th>
<th>Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA (US)</td>
<td>15,221</td>
<td>15,110</td>
<td>11,820</td>
<td>14,081</td>
<td>11,154</td>
</tr>
<tr>
<td>Britain (UK)</td>
<td>11,233</td>
<td>16,076</td>
<td>11,260</td>
<td>10,336</td>
<td>11,771</td>
</tr>
<tr>
<td>France (FR)</td>
<td>10,598</td>
<td>12,953</td>
<td>8,771</td>
<td>9,403</td>
<td>12,021</td>
</tr>
<tr>
<td>Japan (JA)</td>
<td>7,601</td>
<td>11,097</td>
<td>8,695</td>
<td>7,778</td>
<td>8,311</td>
</tr>
<tr>
<td>Germany (GM)</td>
<td>10,163</td>
<td>12,459</td>
<td>11,009</td>
<td>10,106</td>
<td>11,169</td>
</tr>
<tr>
<td>Canada (CA)</td>
<td>10,171</td>
<td>13,795</td>
<td>9,010</td>
<td>11,269</td>
<td>10,004</td>
</tr>
<tr>
<td>Italy (IT)</td>
<td>8,873</td>
<td>17,378</td>
<td>15,890</td>
<td>11,810</td>
<td>13,056</td>
</tr>
<tr>
<td>South Korea (KS)</td>
<td>7,128</td>
<td>7,487</td>
<td>9,070</td>
<td>8,787</td>
<td>9,349</td>
</tr>
<tr>
<td>India (IN)</td>
<td>13,783</td>
<td>16,164</td>
<td>14,268</td>
<td>15,407</td>
<td>13,278</td>
</tr>
<tr>
<td>Spain (SP)</td>
<td>10,350</td>
<td>13,861</td>
<td>9,833</td>
<td>10,029</td>
<td>12,180</td>
</tr>
<tr>
<td>Australia (AS)</td>
<td>12,107</td>
<td>12,953</td>
<td>10,114</td>
<td>14,117</td>
<td>10,872</td>
</tr>
<tr>
<td>China (CH)</td>
<td>11,202</td>
<td>12,020</td>
<td>10,751</td>
<td>10,262</td>
<td>11,111</td>
</tr>
<tr>
<td>Sweden (SZ)</td>
<td>8,648</td>
<td>11,696</td>
<td>8,478</td>
<td>10,484</td>
<td>9,585</td>
</tr>
</tbody>
</table>

Figure 7: mBERT classification results, showing cultural features vary among countries and genres.

- BART
  [https://huggingface.co/facebook/bart-base](https://huggingface.co/facebook/bart-base)

- mBART
  [https://huggingface.co/facebook/mbart-large-50](https://huggingface.co/facebook/mbart-large-50)

### F Classification Results

The results of the culture classification task, including recall, precision, and F1 scores, are presented here, including BERT (Table 12), mBERT (Table 13), Roberta (Table 14), and mRoberta (Table 15). Additionally, for enhanced clarity and visual representation, we offer a comprehensive comparison of F1 scores for all cultures and topics of mBERT in Figure 7, with the complete version depicted in Figure 6. These findings provide valuable insights into the performance and effectiveness of different models in accurately classifying cultures and topics, contributing to the advancement of the field.
<table>
<thead>
<tr>
<th>Cul</th>
<th>Action</th>
<th>Comedy</th>
<th>Drama</th>
<th>Romance</th>
<th>Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>96.03</td>
<td>92.82</td>
<td>94.40</td>
<td>92.66</td>
<td>92.71</td>
</tr>
<tr>
<td>UK</td>
<td>77.42</td>
<td>94.81</td>
<td>85.24</td>
<td>91.12</td>
<td>92.53</td>
</tr>
<tr>
<td>FR</td>
<td>84.18</td>
<td>91.96</td>
<td>83.06</td>
<td>56.19</td>
<td>60.75</td>
</tr>
<tr>
<td>JA</td>
<td>79.78</td>
<td>91.73</td>
<td>85.34</td>
<td>74.34</td>
<td>79.38</td>
</tr>
<tr>
<td>GM</td>
<td>93.80</td>
<td>89.10</td>
<td>91.39</td>
<td>59.52</td>
<td>82.64</td>
</tr>
<tr>
<td>CA</td>
<td>83.44</td>
<td>92.53</td>
<td>87.75</td>
<td>88.78</td>
<td>65.82</td>
</tr>
<tr>
<td>IT</td>
<td>81.69</td>
<td>71.87</td>
<td>76.47</td>
<td>63.64</td>
<td>77.00</td>
</tr>
<tr>
<td>KS</td>
<td>93.73</td>
<td>91.42</td>
<td>93.92</td>
<td>93.69</td>
<td>75.25</td>
</tr>
<tr>
<td>IN</td>
<td>90.07</td>
<td>94.59</td>
<td>92.28</td>
<td>84.52</td>
<td>86.58</td>
</tr>
<tr>
<td>SP</td>
<td>95.38</td>
<td>84.94</td>
<td>89.86</td>
<td>57.68</td>
<td>64.06</td>
</tr>
<tr>
<td>AS</td>
<td>95.78</td>
<td>92.05</td>
<td>93.88</td>
<td>91.49</td>
<td>75.38</td>
</tr>
<tr>
<td>CH</td>
<td>96.11</td>
<td>92.39</td>
<td>94.21</td>
<td>77.17</td>
<td>73.84</td>
</tr>
<tr>
<td>SE</td>
<td>84.10</td>
<td>81.10</td>
<td>82.57</td>
<td>77.50</td>
<td>75.48</td>
</tr>
</tbody>
</table>

**AVG**: 88.58 88.77 88.49 77.56 77.03 76.80 84.07 83.99 83.77 82.44 83.22 82.60 85.83 85.96 85.70

Table 12: Recall (Rec), Precision(Pre) and F1 Performance of Dialogue Culture Classification Model based on BERT. The performance indicators are reported for 13 different cultural categories.

<table>
<thead>
<tr>
<th>Cul</th>
<th>Action</th>
<th>Comedy</th>
<th>Drama</th>
<th>Romance</th>
<th>Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>92.06</td>
<td>95.36</td>
<td>93.68</td>
<td>91.29</td>
<td>97.28</td>
</tr>
<tr>
<td>UK</td>
<td>71.67</td>
<td>98.03</td>
<td>82.80</td>
<td>90.94</td>
<td>89.22</td>
</tr>
<tr>
<td>FR</td>
<td>80.16</td>
<td>81.16</td>
<td>80.66</td>
<td>61.47</td>
<td>59.25</td>
</tr>
<tr>
<td>JA</td>
<td>75.51</td>
<td>88.89</td>
<td>81.65</td>
<td>70.11</td>
<td>78.40</td>
</tr>
<tr>
<td>GM</td>
<td>92.86</td>
<td>92.96</td>
<td>92.91</td>
<td>61.26</td>
<td>71.36</td>
</tr>
<tr>
<td>CA</td>
<td>85.30</td>
<td>88.11</td>
<td>86.68</td>
<td>90.30</td>
<td>58.35</td>
</tr>
<tr>
<td>IT</td>
<td>72.39</td>
<td>74.22</td>
<td>73.29</td>
<td>64.24</td>
<td>77.32</td>
</tr>
<tr>
<td>KS</td>
<td>94.15</td>
<td>88.21</td>
<td>91.08</td>
<td>83.84</td>
<td>83.21</td>
</tr>
<tr>
<td>IN</td>
<td>88.54</td>
<td>88.16</td>
<td>88.35</td>
<td>84.38</td>
<td>77.55</td>
</tr>
<tr>
<td>SP</td>
<td>91.54</td>
<td>78.83</td>
<td>84.71</td>
<td>56.27</td>
<td>72.36</td>
</tr>
<tr>
<td>AS</td>
<td>97.66</td>
<td>83.96</td>
<td>90.29</td>
<td>89.47</td>
<td>84.01</td>
</tr>
<tr>
<td>CH</td>
<td>96.59</td>
<td>87.51</td>
<td>91.82</td>
<td>75.00</td>
<td>71.94</td>
</tr>
<tr>
<td>SE</td>
<td>81.97</td>
<td>79.48</td>
<td>80.71</td>
<td>79.25</td>
<td>76.79</td>
</tr>
</tbody>
</table>

**AVG**: 86.18 86.53 86.05 76.76 76.70 76.26 83.27 82.59 82.62 80.58 80.97 80.48 81.52 81.44 81.21

Table 13: Recall (Rec), Precision(Pre) and F1 Performance of Dialogue Culture Classification Model based on mBERT. The performance indicators are reported for 13 different cultural categories.

<table>
<thead>
<tr>
<th>Cul</th>
<th>Action</th>
<th>Comedy</th>
<th>Drama</th>
<th>Romance</th>
<th>Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>95.20</td>
<td>92.36</td>
<td>93.76</td>
<td>89.50</td>
<td>94.16</td>
</tr>
<tr>
<td>UK</td>
<td>77.58</td>
<td>94.03</td>
<td>85.01</td>
<td>89.80</td>
<td>84.61</td>
</tr>
<tr>
<td>FR</td>
<td>82.38</td>
<td>79.01</td>
<td>80.66</td>
<td>59.94</td>
<td>54.76</td>
</tr>
<tr>
<td>JA</td>
<td>76.85</td>
<td>92.93</td>
<td>84.13</td>
<td>67.55</td>
<td>77.06</td>
</tr>
<tr>
<td>GM</td>
<td>92.66</td>
<td>94.22</td>
<td>93.43</td>
<td>56.04</td>
<td>70.52</td>
</tr>
<tr>
<td>CA</td>
<td>85.65</td>
<td>86.34</td>
<td>85.99</td>
<td>84.08</td>
<td>63.48</td>
</tr>
<tr>
<td>IT</td>
<td>80.08</td>
<td>75.36</td>
<td>77.64</td>
<td>71.36</td>
<td>69.09</td>
</tr>
<tr>
<td>KS</td>
<td>93.82</td>
<td>93.10</td>
<td>93.46</td>
<td>91.41</td>
<td>87.65</td>
</tr>
<tr>
<td>IN</td>
<td>92.16</td>
<td>93.65</td>
<td>92.90</td>
<td>84.72</td>
<td>78.06</td>
</tr>
<tr>
<td>SP</td>
<td>87.81</td>
<td>89.75</td>
<td>88.77</td>
<td>50.42</td>
<td>79.93</td>
</tr>
<tr>
<td>AS</td>
<td>94.77</td>
<td>91.48</td>
<td>93.10</td>
<td>86.86</td>
<td>79.17</td>
</tr>
<tr>
<td>CH</td>
<td>97.00</td>
<td>87.93</td>
<td>92.24</td>
<td>80.25</td>
<td>74.59</td>
</tr>
<tr>
<td>SE</td>
<td>86.00</td>
<td>78.29</td>
<td>81.96</td>
<td>76.43</td>
<td>74.14</td>
</tr>
</tbody>
</table>

**AVG**: 87.84 88.34 87.93 76.03 75.86 75.43 82.67 83.25 82.39 83.39 83.37 83.20 85.78 85.57 85.29

Table 14: Recall (Rec), Precision(Pre) and F1 Performance of Dialogue Culture Classification Model based on Roberta. The performance indicators are reported for 13 different cultural categories.
<table>
<thead>
<tr>
<th>Cul</th>
<th>Action</th>
<th>Comedy</th>
<th>Drama</th>
<th>Romance</th>
<th>Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rec</td>
<td>Pre</td>
<td>F1</td>
<td>Rec</td>
<td>Pre</td>
</tr>
<tr>
<td>US</td>
<td>93.09</td>
<td>94.60</td>
<td>93.84</td>
<td>87.27</td>
<td>97.63</td>
</tr>
<tr>
<td></td>
<td>85.45</td>
<td>91.92</td>
<td>88.57</td>
<td>81.23</td>
<td>93.39</td>
</tr>
<tr>
<td>UK</td>
<td>70.53</td>
<td>97.28</td>
<td>81.77</td>
<td>91.20</td>
<td>87.26</td>
</tr>
<tr>
<td></td>
<td>68.37</td>
<td>74.47</td>
<td>71.29</td>
<td>88.72</td>
<td>80.39</td>
</tr>
<tr>
<td>FR</td>
<td>81.72</td>
<td>75.30</td>
<td>78.38</td>
<td>71.77</td>
<td>46.49</td>
</tr>
<tr>
<td></td>
<td>74.94</td>
<td>87.04</td>
<td>80.54</td>
<td>76.72</td>
<td>81.75</td>
</tr>
<tr>
<td>JA</td>
<td>80.45</td>
<td>86.06</td>
<td>83.16</td>
<td>78.66</td>
<td>63.58</td>
</tr>
<tr>
<td></td>
<td>92.27</td>
<td>76.31</td>
<td>84.48</td>
<td>76.72</td>
<td>81.75</td>
</tr>
<tr>
<td>GM</td>
<td>93.38</td>
<td>91.12</td>
<td>92.24</td>
<td>55.96</td>
<td>69.26</td>
</tr>
<tr>
<td></td>
<td>65.55</td>
<td>79.07</td>
<td>71.68</td>
<td>68.96</td>
<td>79.24</td>
</tr>
<tr>
<td>CA</td>
<td>85.12</td>
<td>88.90</td>
<td>86.97</td>
<td>86.73</td>
<td>60.66</td>
</tr>
<tr>
<td></td>
<td>70.33</td>
<td>67.62</td>
<td>68.95</td>
<td>93.54</td>
<td>89.88</td>
</tr>
<tr>
<td>IT</td>
<td>77.13</td>
<td>75.56</td>
<td>76.34</td>
<td>61.12</td>
<td>78.65</td>
</tr>
<tr>
<td></td>
<td>70.48</td>
<td>76.05</td>
<td>73.16</td>
<td>88.78</td>
<td>76.12</td>
</tr>
<tr>
<td>KS</td>
<td>95.03</td>
<td>86.62</td>
<td>90.63</td>
<td>87.37</td>
<td>84.60</td>
</tr>
<tr>
<td></td>
<td>93.66</td>
<td>76.27</td>
<td>84.08</td>
<td>76.41</td>
<td>78.31</td>
</tr>
<tr>
<td>IN</td>
<td>90.44</td>
<td>88.38</td>
<td>89.40</td>
<td>87.04</td>
<td>80.45</td>
</tr>
<tr>
<td></td>
<td>86.26</td>
<td>88.81</td>
<td>87.52</td>
<td>78.75</td>
<td>93.57</td>
</tr>
<tr>
<td>SP</td>
<td>89.68</td>
<td>82.46</td>
<td>85.92</td>
<td>53.57</td>
<td>74.96</td>
</tr>
<tr>
<td></td>
<td>70.97</td>
<td>72.91</td>
<td>71.93</td>
<td>86.96</td>
<td>86.96</td>
</tr>
<tr>
<td>AS</td>
<td>93.36</td>
<td>91.30</td>
<td>92.32</td>
<td>87.75</td>
<td>79.93</td>
</tr>
<tr>
<td></td>
<td>93.10</td>
<td>87.59</td>
<td>90.26</td>
<td>92.38</td>
<td>90.68</td>
</tr>
<tr>
<td>CH</td>
<td>94.68</td>
<td>89.66</td>
<td>92.10</td>
<td>73.25</td>
<td>83.63</td>
</tr>
<tr>
<td></td>
<td>82.50</td>
<td>83.30</td>
<td>82.90</td>
<td>84.38</td>
<td>80.87</td>
</tr>
<tr>
<td>SE</td>
<td>82.19</td>
<td>80.57</td>
<td>81.37</td>
<td>72.69</td>
<td>79.82</td>
</tr>
<tr>
<td></td>
<td>90.89</td>
<td>72.01</td>
<td>80.36</td>
<td>84.22</td>
<td>87.70</td>
</tr>
</tbody>
</table>

Table 15: Recall (Rec), Precision(Pre) and F1 Performance of Dialogue Culture Classification Model based on mRoberta. The performance indicators are reported for 13 different cultural categories.