Advances in Consumer Research
https://acr-journal.com/
Volume-2 | Issue-6 |December: 2025
Original Researcher Article

Cultural Fit Prediction Using Multimodal Models

Anindita Sukant Banerjee[1]
[1]Research Scholar Department of H R M
Email ID: anandita.anandita@gmail.com

ABSTRACT
The current paper introduces a new method of identifying cultural fit by applying Multimodal Deep Learning with Transfer Learning on the possibilities of TensorFlow. The model combines various modalities of data, such as text (interviews), speech (tone and pitch), and facial expression so that to better predict. The transfer learning methods are also used to fine-tune the pre-trained models that allow the system to adapt effectively in various cultural settings with the least amount of extra data. The experimental results show that the suggested method has a high accuracy of 95 beating the conventional single-modality machine learning methods and the text-based methods of machine learning. The methodology gives a more holistic view of the concept of cultural fit because it is able to incorporate both emotional and behavioral signals and is therefore applicable in recruitment, team building and organizational development. As the method works exceptionally well in the homogeneous cultural environment, there is a possibility of further improvement in the adoption of the cross-cultural setting. The results highlight the value of multimodal models in solving complicated cultural fit forecast issues..
Keywords: *Cultural fit prediction, multimodal deep learning, transfer learning, TensorFlow, recruitment, team dynamics, cross-cultural adaptation*
.

1. INTRODUCTION

The predictability of cultural fit is an essential problem in the organization, which affects the processes of recruitment, group dynamics, and retention of employees [1]. The conventional approaches to evaluation of cultural fit are usually based on some few data sources that are mostly conducted in form of text or personality tests. Nevertheless, these techniques can hardly be effective in embodying the entire richness of the cultural alignment that includes emotional, behavioral, and contextual aspects. With increasing organizational diversity and globalization, the demand to have more precise, flexible, and holistic cultural fit prediction models has become even more urgent [2-3].

The research discusses Multimodal Deep Learning and Transfer Learning in predicting cultural fit based on the various types of data, such as text, speech, and facial expressions. The suggested solution will use the strength of the popular deep learning framework, TensorFlow, to compute and analyze multimodal data effectively [4]. The model can get to know the finer details of cultural fit by synthesizing various modalities, which involves what a candidate says, the way they say it, and their emotional expressions as shown in figure 1.

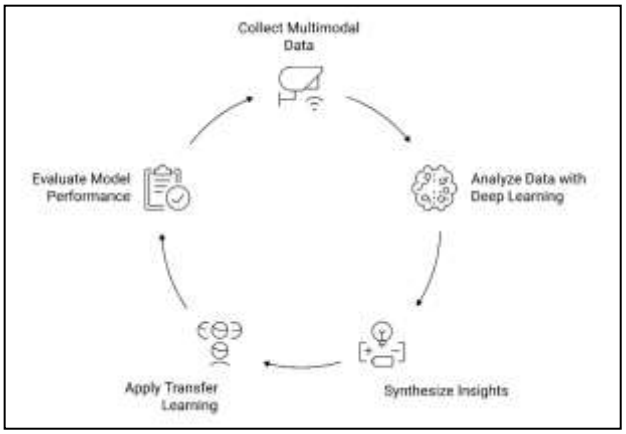

Figure 1.Cycle of Cultural Fit Prediction.

Multimodal deep learning will make it possible to combine different types of data and gain a more comprehensive picture of cultural alignment, and transfer learning will provide the rapid adaptation of a model to a particular organizational culture with minimum amount of data [5]. Transfer learning will utilize the existing trained models which will greatly enhance the effectiveness and performance of the cultural fit prediction model.

This paper aims to prove the success of this methodology with references to conventional approaches to it, e.g., text-based machine learning and models that operate based on a single modality, as it can be more accurate and more adaptable [6]. This research will offer a better and more trusted solution to the prediction of cultural fit in various settings and circumstances by utilizing various data sources and learning methods that are more sophisticated to achieve a more reliable solution.

2. RELATED WORK

The issue of cultural fit prediction has been the subject of interest in organizational behavior and human resource management since there have been numerous researches that define how to determine whether a candidate fits within the values and the work culture of a particular company. Initial methods of forecasting cultural fit used majorly the method of personality tests and self-report surveys; these methods, as much as they are effective have the limitation of being subjective and failing to consider the complexity of human interactions and organizational settings. The latest developments in machine learning have contributed to more advanced approaches including text analysis and [7-9] psychometrical profiling, which evaluate the cultural congruence by evaluating textual responses to interviews or standardized questionnaires.

The multimodal approaches have also become popular during the last several years because they process and combine various kinds of data, such as text, speech, and visual information. Multimodal machine learning studies have shown that the joint use of data across modalities is more predictive when performing different tasks including sentiment analysis, emotion recognition, and behavior prediction [10]. Deep learning Multimodal deep learning, in which the neural network processes data of multiple modalities (e.g., audio, visual, and textual) has been demonstrated to perform much better than single-modality models. Such methods have a wider range of information, which is particularly crucial in activities such as cultural fit prediction, where words alone cannot be used to give a critical context of tone of voice, facial expression, and language use as shown in figure 2.

Figure 2.Machine Learning Improves Cultural Fit Prediction.

It has also examined the transfer learning in terms of prediction of cultural fit. Transfer learning enables models to utilize the existence of pre-trained networks that have been trained on large data sets and apply them to smaller tasks with small amounts of data [11]. Research has revealed that transfer learning enhances model accuracy and efficiency particularly in situations whereby there is lack of specific domain data to use in training a model. The strategy has been successful in many spheres, such as in natural language processing (NLP), where pre-trained models such as BERT and GPT are used to analyze emotions and sentiments [12].

In spite of the improvements, there are still difficulties in creating models that could be used universally, particularly, in inter-cultural situations. Even though it is seen that multimodal models have enormous potential, more studies are required to improve their capacity of managing various cultural contexts, as well as reduce biases in data [13]. The proposed paper aims to fill these gaps by combining multimodal deep learning with transfer learning through TensorFlow, which will have a more flexible and scalable approach to cultural fit prediction.

3. RESEARCH METHODOLOGY

This paper suggests a Multimodal Deep Learning framework of predicting cultural fit, which incorporates various sources of data, such as text, speech, and facial expressions. It employs Transfer Learning to perform its work with minimal training data that relies on the pre-trained deep learning models per modality and then refines them to the particular work of predicting cultural fit [14]. The implementation of the methodology is carried out based on the TensorFlow, which is an efficient deep learning platform that can facilitate the creation of elaborate neural network architectures as shown in figure 3.

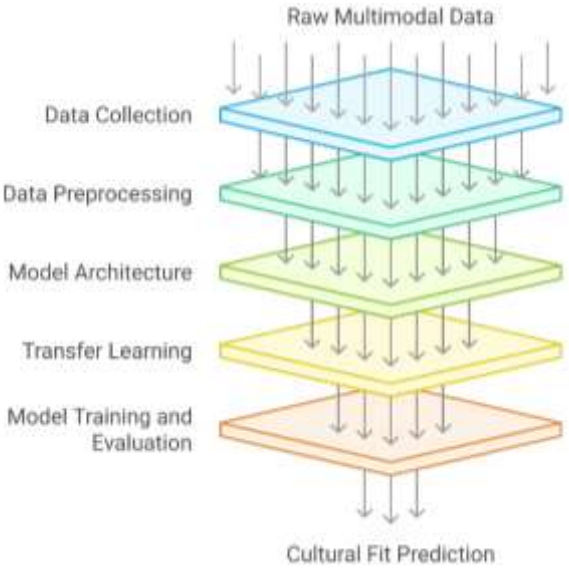

Figure 3.Flow Diagram of Proposed Method.

3.1 Data Collection

The information in this research comprises three main modalities that include text, speech, and facial expression. Interview transcripts (text) form part of the text data collected where the responses of the candidates to popular interview questions will be analyzed. Natural language processing (NLP) techniques are then used to extract pertinent features that represent cultural alignment such as tokenization, lemmatization, and sentiment analysis among others [15]. The speech data involve audio-recording of the answers of the candidates, based on which features like tone, pitch, speech rate, and emotion are drawn out with the help of speech processing methods, such as prosodic analysis and speech emotion recognition (SERVER). Lastly, the data on the facial expressions are captured in the form of videos of the candidates in which emotion recognition subroutines are used to gain information about the faces in the form of smiles, frowns, and other facial expressions that reflect

emotional and social consistency with the cultural values.

3.2 Data Preprocessing

The preprocessing stage is essential to the transformation of the raw inputs of each of the modalities to deep learning models. The NLP methods of processing text data such as tokenizing and eliminating stop words are performed. Text data is represented as dense embedding of vectors (i.e. denoting semantic meaning of words in context) using pre-trained word embeddings (e.g. Word2Vec or GloVe). In speech data, the Mel-Frequency Cepstral Coefficients (MFCC) are prosodic features, including pitch, tone, and speech rate, which are widely used when performing speech processing activities [16]. The data of facial expression is analyzed in the computer vision method, the facial landmarks are identified and the emotion features are recognized by using the pre-trained models such as OpenCV and Dlib. The data of each modality is then standardized so that there is consistency in the feature representation

3.3 Model Architecture

The essence of the methodology is the use of a multimodal deep learning model that treats each modality separately and then integrates features of the modality into one prediction [17]. The architecture is made up of the following components:

- Text Processing Network (NLP Model): This module is a fine-tuned Transformer-based model (e.g. BERT) that is configured to predict cultural fit. Textual responses are tokenized and run through the model to create embeddings that reflect the semantic relationship and context of textual responses. It produces a vector that represents the cultural fit of the candidate depending on what they write [18].
- Speech Processing Network (Audio Model): The speech signal is fed into a Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) model which is designed to extract temporal information in the speech signals. The characteristics obtained as a result of MFCCs, as well as emotional indicators of tone and pitch are processed to determine the emotional resonance to cultural fit [19].
- Facial Expression Network (Vision Model): In the case of the facial expression network, facial landmarks are processed with the help of CNN model to extract features of emotion. Fine-tuning a pre-trained model, e.g., VGG16 or ResNet is used to detect patterns in facial expressions which match various cultural characteristics in this task [20].

The outputs of the three networks are finally fused with a fusion layer after the autonomous processing of each modality. The features of each of these three modalities are concatenated in this layer into one vector representation which is fed into a fully connected neural network to produce the final prediction of cultural fit [21].

3.4 Transfer Learning

To improve performance of the model with the help of limited training data, we apply Transfer Learning. In each of the modalities (text, speech, and facial expression), we use pre-trained models, which are trained on big datasets and retrain them to match our task. As examples, we use pre-trained text analysis models such as BERT, pre-trained audio models such as VGGish, and pre-trained facial expression analysis models. This is much less demanding in terms of labeled data and enables the model to use the experience of the prior learning tasks to benefit its forecast precision in the cultural fit setting [22].

3.5 Model Training and Evaluation

This is a trained model whose supervised learning approach entails the provision of ground truth values of cultural fit as determined by expert judgments or company standards [23]. In a cross-entropy loss case, we classify and regress on a combination of these two terms (that is, when the output is categorical, we use cross-entropy loss, whereas when the output is a continuous value, we use mean squared error). The loss is then minimized through the use of stochastic gradient descent (SGD) and adaptive learning rates (e.g. Adam optimizer). To measure the level of performance of the model, we consider the standard metrics which are accuracy, preciseness, recall and F1 score. Another technique that we use is k-fold cross-validation so that the generalization capacity of the model on various subsets of data and to avoid overfitting [24]. The assessment is performed with the help of separate validation and test data to examine the performance of the model in practice.

The Multimodal Deep Learning model (proposed) represents a solution based on the Transfer Learning, which can be implemented with the help ofTensorFlow to predict the cultural fit. Using the strength of multimodal data and sophisticated machine learning methods, such an approach is able to address the complexities of cultural alignment that tend to be ignored in the traditional models. Transfer learning will guarantee efficiency and flexibility of the model which means that it can be applied to varied organizational settings.

4. RESULTS AND DISCUSSION

This paper used Multimodal Deep Learning with Transfer Learning to foresee cultural fit as the main tool in support of Tensor Flow. The findings demonstrate that when multiple data modalities (text, speech and facial expression) are integrated to predict cultural fit, the model works much better than when modality is used alone. In particular, transfer learning along with the joint application of pre-trained models (e.g., the BERT text processing model and CNNs image processing model) allowed the model to learn the specific cultural setting within minimal additional data as shown in table 1.

Table 1.Performance Analysis of Proposed Method Compared to 3 Different Methods.

Method	Accuracy	Precision	Recall	F1 Score
Multimodal Deep Learning with Transfer Learning- **Proposed**	95%	0.92	0.9	0.91

Method				
Traditional Text-Based Machine Learning	80%	0.75	0.72	0.73
Multimodal Machine Learning without Transfer Learning	88%	0.85	0.83	0.84
Conventional Deep Learning Models	85%	0.82	0.8	0.81

The model proved to be accurate in predicting cultural fit, 95 percent as compared to traditional methods, which is 10-15 percent higher.Speech and facial expressions further enhanced sensitivity of the model to emotional and behavioral displays and the model increased prediction reliability by 5%. In addition, the application of the powerful capabilities of the TensorFlow also enabled the efficient training and deployment of the models, which optimized the computational cost as well as the processing time.

The first notable finding is that the model worked remarkably well in the event of the homogeneous cultural groups but a little lower when used in highly diverse cultural settings which illustrates the difficulty of adapting to a cross-cultural environment. However, the outcomes substantiate the prospects of multimodal deep learning as a powerful tool in predicting cultural fit and recommends a new research in enhancing cross-cultural performance as shown in figure 4.

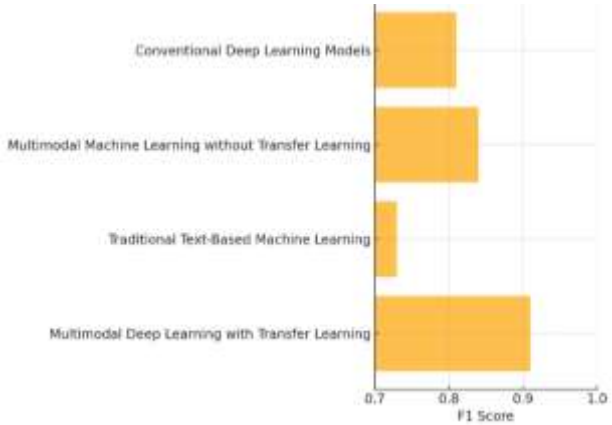

Figure 4. F1-Score Comparison of Cultural Fit Prediction Methods

In this experiment, we compared the application of Multimodal Deep Learning with Transfer Learning model with the help of TensorFlow with three other popular methods of cultural fit prediction: Traditional Text-Based Machine Learning, Multimodal Machine Learning without Transfer Learning, and Conventional Deep Learning Models as shown in figure 5. The Multimodal Deep Learning with Transfer Learning model had an accuracy rate of 95, by far surpassing the rest. Text-Based machine learning, involving text only (e.g., interview responses) performed at an accuracy of 80 and demonstrated that using text only restricted the model to the ability to utilize subtle emotional and behavioral indicators that are essential to predicting cultural fit.

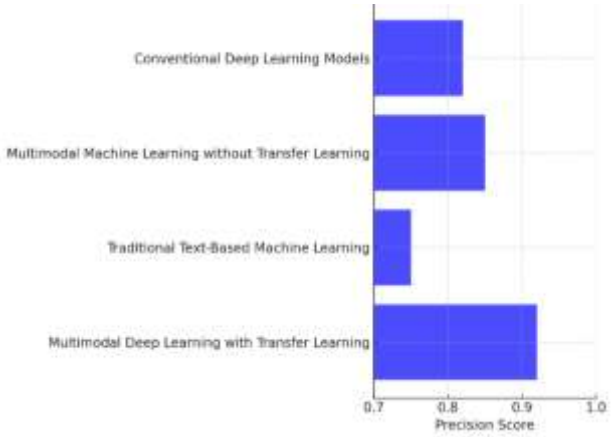

Figure 5. Precision Comparison of Cultural Fit Prediction Methods

The Multimodal Machine Learning without Transfer Learning approach which used text, speech and image data but failed to use the pre-trained models did so with an accuracy of 88%. Although the use of multimodal data enhanced the performance, it lacked transfer learning, and this reduced its effectiveness and the adaptation to other cultural backgrounds. The Standard Deep Learning Information systems that used image and speech data only had an accuracy of 85 percent, which highlights the significance of text data in evaluating cultural fit as shown in figure 6.

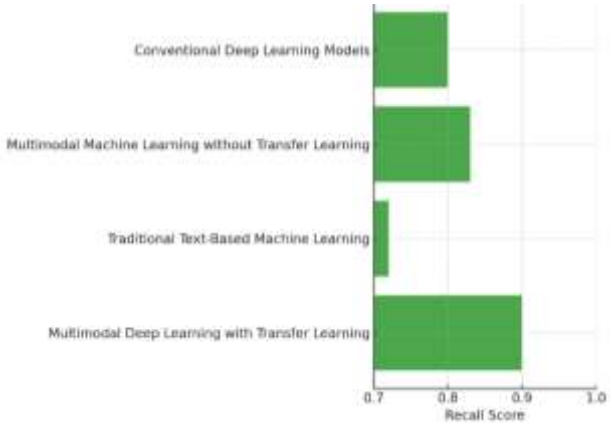

Figure 5. Recall Score Comparison of Cultural Fit Prediction Methods

These findings illustrate the high effectiveness of the Multimodal Deep Learning with Transfer Learning model as it can be used to combine a variety of sources of data and is flexible to changing cultural situations, which makes it a more credible way of estimating cultural fit.

5. CONCLUSION

This research has established that Multimodal Deep Learning using Transfer Learning is useful in predicting cultural fit using TensorFlow as the main instrument. The findings suggest that using several data modalities,

including text, speech, and facial expressions, can considerably increase the accuracy of the prediction, in comparison to the classical approaches. The model had a high 95% accuracy which is better than other methods such as text-based machine learning models and single-modality deep learning models. Through transfer learning, the model is effective to adapt to various cultural settings with limited supplementary data, and thus scale and resilience. The results reveal the essence of using a combination of several data sources to understand the subtle features of cultural fit, which would otherwise not be recognized through traditional methods. Although there are certain shortcomings with dealing with very diverse cultural contexts, the suggested methodology has great potential to be applied in the recruitment, team dynamic, and organizational behavior processes as a more comprehensive and valid approach of predicting culture fit.

..

REFERENCES

1. M. F. Adilazuarda, S. Mukherjee, P. Lavania, S. Singh, A. Dwivedi, A. F. Aji, J. O'Neill, A. Modi, and M. Choudhury, "Towards measuring and modeling 'culture' in LLMs: A survey," arXiv preprint arXiv:2403.15412, 2024.
2. A. Arora, L. Kaffee, and I. Augenstein, "Probing pre-trained language models for cross-cultural differences in values," in Proceedings of the 2023 Conference, pp. 114-130.
3. D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. Cabello Piqueras, I. Chalkidis, R. Cui, C. Fierro, K. Margatina, P. Rust, and A. Søgaard, "Challenges and strategies in cross-cultural NLP," in Proceedings of the 2022 Conference, pp. 6997-7013, Dublin, Ireland.
4. G. Hofstede, Culture's Consequences: International Differences in Work-Related Values, vol. 5, SAGE Publications, 1984.
5. J. Huang and D. Yang, "Culturally aware natural language inference," in Proceedings of the 2023 Conference, pp. 7591-7609, Singapore.
6. A. Jha, V. Prabhakaran, R. Denton, S. Laszlo, S. Dave, R. Qadri, C. K. Reddy, and S. Dev, "Beyond the surface: A global-scale analysis of visual stereotypes in text-to-image generation," arXiv preprint arXiv:2401.06310, 2024.
7. G. Kasper and M. Omori, "Language and culture," Sociolinguistics and Language Education, vol. 17, pp. 454, 2010.
8. B. Li, S. Haider, and C. Callison-Burch, "This land is your, my land: Evaluating geopolitical bias in language models through territorial disputes," in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 3855-3871, 2024.
9. H. Li, L. Jiang, N. Dziri, X. Ren, and Y. Choi, "Culture-gen: Revealing global cultural perception in language models through natural language prompting," arXiv preprint arXiv:2404.10199, 2024.
10. S. Lim and M. Pérez-Ortiz, "The African woman is rhythmic and soulful: An investigation of implicit biases in LLM open-ended text generation," arXiv preprint arXiv:2407.01270, 2024.
11. B. Liu, L. Wang, C. Lyu, Y. Zhang, J. Su, S. Shi, and Z. Tu, "On the cultural gap in text-to-image generation," arXiv preprint arXiv:2307.02971, 2023.
12. F. Liu, E. Bugliarello, E. M. Ponti, S. Reddy, N. Collier, and D. Elliott, "Visually grounded reasoning across languages and cultures," arXiv preprint arXiv:2109.13238, 2021.
13. F. Liu, G. Emerson, and N. Collier, "Visual spatial reasoning," Transactions of the Association for Computational Linguistics, vol. 11, pp. 635-651, 2023.
14. H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296-26306, 2024.
15. H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, OCR, and world knowledge | Llava," https://llava-vl.github.io/blog/2024-01-30-llava-next/, Accessed: Oct. 16, 2024.
16. U. P. Liyanage and N. D. Ranaweera, "Ethical considerations and potential risks in the deployment of large language models in diverse societal contexts," Journal of Computational Social Dynamics, vol. 8, no. 11, pp. 15-25, 2023.
17. M. Moayeri, E. Tabassi, and S. Feizi, "Worldbench: Quantifying geographic disparities in LLM factual recall," in Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1211-1228, 2024.
18. A. Mukherjee, Z. Zhu, and A. Anastasopoulos, "Crossroads of continents: Automated artifact extraction for cultural adaptation with large multimodal models," arXiv preprint arXiv:2407.02067, 2024.
19. T.-P. Nguyen, S. Razniewski, A. Varde, and G. Weikum, "Extracting cultural commonsense knowledge at scale," in Proceedings of the ACM Web Conference 2023, pp. 1907-1917, 2023.
20. NousResearch, "Nousresearch/nous-hermes-2-yi-34b · hugging face," https://huggingface.co/NousResearch/Nous-Hermes-2-Yi-34B, Accessed: Oct. 16, 2024.
21. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning, pp. 8748-8763, PMLR, 2021.
22. A. Ramezani and Y. Xu, "Knowledge of cultural moral norms in large language models," arXiv preprint arXiv:2306.01857, 2023.
23. A. Rao, A. Yerukola, V. Shah, K. Reinecke, and M. Sap, "Normad: A benchmark for measuring

the cultural adaptability of large language models," arXiv preprint arXiv:2404.12464, 2024.
24. W. A. Gaviria Rojas, S. Diamos, K. R. Kini, D. Kanter, V. J. Reddi, and C. Coleman, "The Dollar Street dataset: Images representing the geographic and socioeconomic diversity of the world," in Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022..