Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 2** : 1–9.
ISSN-0037-8712
doi:10.5269/bspm.80479

Leveraging Cross-Lingual Transfer Learning for Low-Resource Natural Language Processing

Ushashree P., Vansh Badani, Nikilesh and B. Nagamani

ABSTRACT: The field of natural language processing (NLP) is growing quickly, yet, many languages are still under-represented because there is a dearth of labelled data. This study investigates the transfer of knowledge from resource-rich to low-resource languages through cross-lingual transfer learning as a way to overcome this difficulty. We test multilingual models like mBERT and XLM-R on tasks including machine translation, named entity recognition, and sentiment analysis. These models are refined utilising task-specific datasets from low-resource languages after being pre-trained on a variety of languages. Significant gains are demonstrated by the results, particularly in tasks with little labelled data and in languages that are closely linked to those used in the pre-training. These results demonstrate how multilingual models can help close the performance disparities between languages.Overall, this study offers useful information and shows how well cross-lingual transfer learning is in low- resource environments. These findings highlight the potential of multilingual models to reduce perfor- mance gaps across languages.Overall, this work demonstrates the effectiveness of cross-lingual transfer learning in low-resource situations and gives practical insights for designing inclusive NLP systems that better reflect global linguistic variety.

Keywords: Cross-lingual NLP, multilingual models, trans-fer learning, low-resource language, BERT.

Contents

1. Introduction

Research in Natural Language Processing (NLP) is grow- ing, with applications in business, education, healthcare, and communication. Computers can now comprehend and produce human language more effectively thanks to recent devel- opments in machine learning and deep learning. Improved performance has been observed in tasks including machine translation, named entity recognition (NER), and sentiment

2020 *Mathematics Subject Classification:* 68T50.
Submitted November 29, 2025. Published March 14, 2026

analysis. However, the unequal dissemination of data across languages continues to be a significant obstacle. Large labelled datasets are necessary for training the majority of NLP mod- els.While such resources exist for widely spoken languages.

like English, Chinese, and Spanish, they are scarce for many others. These low-resource languages often lack sufficient data to support reliable model development. As a result, NLP tools are unavailable or less effective for many linguistic communities. This limits the usefulness of NLP in real-world areas such as healthcare or education, where accurate language tools could improve access to essential services. To reduce reliance on large datasets, researchers have ex- plored cross- lingual transfer learning, where knowledge from resource- rich languages is used to improve performance in languages with limited data. Pre-trained multilingual mod- els such as multilingual BERT (mBERT) and XLM-Roberta (XLM-R) are trained on text from multiple languages and can later be fine- tuned for specific tasks. Because these models capture shared language patterns, they provide a practical way to extend NLP support to languages with limited resources. This study focuses on evaluating how cross-lingual transfer learning can improve NLP performance for low-resource lan- guages, specifically Hindi and Telugu, by transferring knowl- edge from high-resource English. We examine mBERT and XLM-R across three tasks: sentiment analysis, named entity recognition, and machine translation. The study also con- siders how similarities between languages influence transfer performance. The goal is to provide both empirical results and practical guidance for applying transfer learning methods in low-resource settings. The contributions of this work are three parts: First, it adds evidence on the performance of multilingual models when applied to languages with limited data. Second, it offers researchers and developers working on NLP applications in these kinds of settings useful information. Third, it highlights how crucial it is to develop inclusive language technologies that promote linguistic diversity and increase the number of people who can access digital services.

This paper's remaining sections are arranged as follows: Section 2 examines relevant research on multilingual NLP and cross-lingual transfer learning. The models, datasets, and techniques are described in Section 3. The experimental results are shown in Section 4, and Section 5 discusses them. The key conclusions and recommendations for further research are presented in Section 6.

2. Literature Survey

Natural Language Processing (NLP) has seen significant growth in recent years, enabling machines to analyze, interpret, and generate human language for tasks such as sentiment analysis, named entity recognition (NER), machine translation, and text summarization [1] [2]. Most advances rely on large annotated datasets, which are widely available for major languages but scarce for many others. This scarcity limits the development of NLP tools for low-resource languages, making it difficult to deploy applications effectively across diverse linguistic communities.

Several studies have explored ways to leverage existing resources for low-resource NLP. [3]proposed multi-task deep neural networks for natural language understanding, showing that training a single model across multiple tasks improves generalization and efficiency. [4]analyzed multilingual BERT (mBERT) and demonstrated that while cross-lingual trans- fer works well for linguistically similar languages, performance drops for typologically distant ones. [5] addressed low- resource named entity recognition using neural machine trans- lation to transfer knowledge from resource-rich languages, highlighting practical strategies for limited-data settings.

Based on these concepts, [6] proposed the InfoXLM, which is a framework that complements cross-lingual pre-training using information-theoretic goals. This method enhances the sharing of the representation between the languages and has a high performance in the classification and translation process. [7] specifically considered the low-resource NER and demonstrated that multilingual pre-training with task-specific fine-tuning can achieve a large improvement in performance in low-annotation languages.

Cross-lingual models such as mBERT ([8]) and XLM-R ([9]) have become a key component of cross-lingual tasks. XLM-R, which is trained on more than 100 languages, has good per- formance even when fine-tuning is needed on low-resource languages, but it consumes large amounts of computational resources. Multilingual corpora supporting cross-lingual research and analysis are offered by the Universal Depen- dencies Project [9], but annotated data is still unavailable in many languages.

The practical NLP applications also shed light on the issues of limited data working. To illustrate this, in the preprocessing and feature extraction of noisy real-world text, sentiment analysis to detect depression in social media users [10] is shown. In like manner, the AI-Driven Health [11] represents a web-based application, which deciphers the healthcare inquiries and nutrition data with the assistance of NLP. These studies highlight the importance of means that can be used to make cross-linguistic cross-domain generalizations even in the face of sparse annotated data. The latest studies have come very far in the achievement of natural language pro- cessing in various languages and in particular, languages with fewer resources.

As demonstrated in [12], multilingual denoising pre-training can be used to teach models to learn strong repre- sentations through corrupted text reconstruction, which enhances translation quality in language pairs with limited resources significantly. [13] investigated the cross-lingual capabilities of mBERT and the results indicate that it can store useful multilingual information, especially related languages but transfering to very distant languages is difficult. Similarly, [13] emphasized the remarkable zero-shot performance of BERT across languages by indicating its applicability even when its training is exclusively on one language. Based on that, [14] proposed an al- ternating language modeling, where models are trained to alternate between languages, to better reflect common linguistic patterns, and showed better results on multilingual bench-marks. Complementing these advances, [15] created MLQA, a benchmark for cross-lingual question answering, revealing how model performance can vary across language pairs and providing a practical tool for evaluating multilingual NLP systems Together, these studies highlight the potential and difficulties of cross-lingual models and provide insightful information for implementing multilingual pre-training in low- resource, real-world settings. Even though current cross-lingual techniques have showed promise, they frequently require a lot of fine-tuning or big parallel corpora and struggle with low-resource languages, par- ticularly for distant language pairs. By utilising mBERT and XLM-R in conjunction with typological traits and linguistic similarities, our suggested method overcomes these difficulties and provides reliable performance across a variety of low- resource languages. Because of this, our approach is both scalable and useful for tasks like machine translation, named entity recognition, and sentiment analysis.

3. Proposed Method

A.The step by step description of the proposed methodology of this study is supported through the system architecture presented in Figure 4. It is aimed at studying the effectiveness of multilingual models in cross-lingual transfer learning of low-resource languages on a range of natural language processing (NLP) problems, including machine translation, named entity recognition (NER), and sentiment analysis. To ensure comparable and correct findings, the methodology pays a high value to precise dataset preparation, preprocessing and model selection and fine-tuning.

3.1. Data Collection

To carry out this research, datasets were carefully edited to encompass the NLP tasks of sentiment analysis, named entity recognition (NER) and machine translation. This study is concerned with two low resource Indian languages, Hindi and Telugu. English was also introduced in the experiments to use as a high resource baseline on performance comparison and as a source language in our machine translation activities. Each task was performed on specific datasets, which are important to guarantee the presence of a strong and repeatable analysis.The sentiment analysis task was performed on the IIT Patna Product Review Sentiment Analysis dataset in Hindi consisting of 1,134 sentences that are labeled. In the case of Telugu we used the the ACTREC-Telugu Movie Review Dataset, which has 5,420 reviews of movies that are labeled. The high-resource English baseline was the standard IMDB Large Movie Review Dataset which is 50,000 reviews.

To test Named Entity Recognition, we obtained dataset which has standard entity annotations. In Hindi, we used the CoNLL-2008 Hindi-Urdu Treebank of the Universal Dependencies Project that includes about 150,000 tokens. The Telugu NER data came by taking the FIRE 2013 Shared Task south Asian languages, and it had more than 75,000 annotated tokens. English was done using the well-known English dataset, the CoNLL- 2003, which contains more than 200,000 tokens.The English machine translation tasks were performed with large-scale parallel corpus. The English-Hindi activity was based on the use of

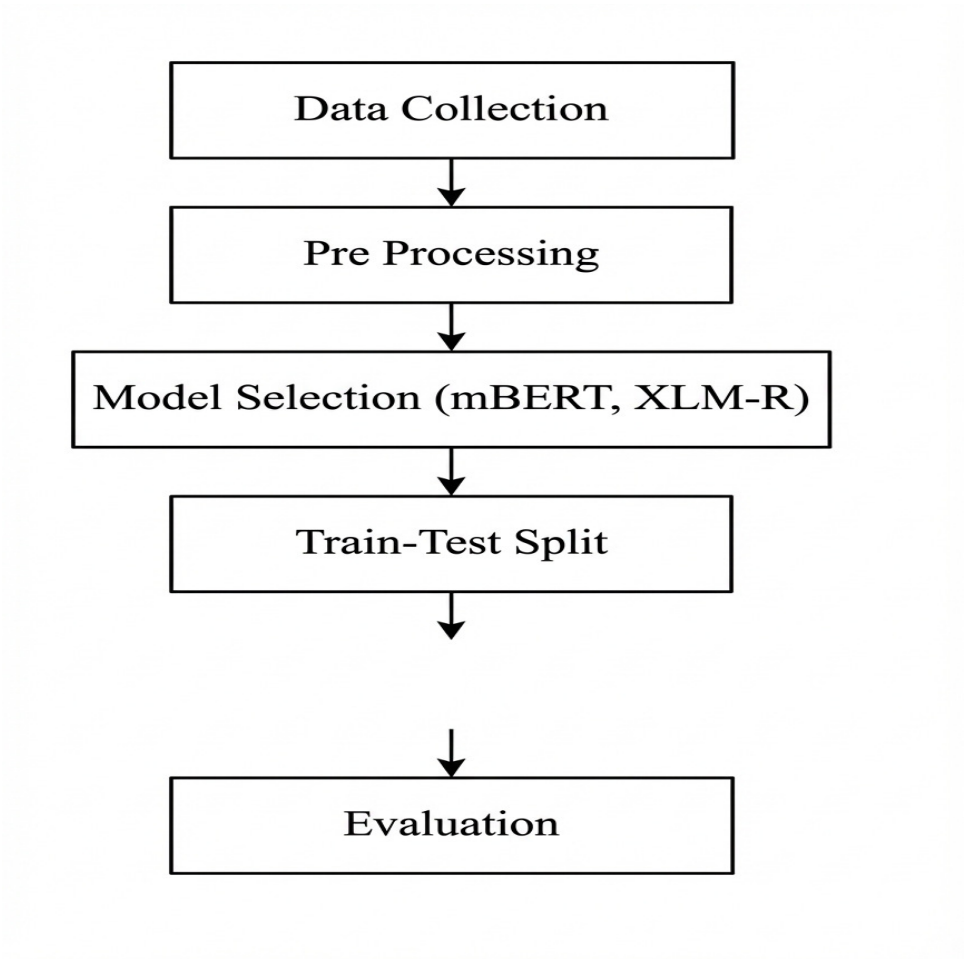

Figure 1: Proposed method

the English-Hindi Parallel Corpus of the English-Hindi-sentences consisting of the total number of about 1.5 million pairs of sentences. In the case of English-Telugu language pair, we used the Samanantar Corpus where we obtained more than 450,000 parallel sentences.

3.2. Pre-processing Techniques

The datasets that had been collected prior to training and fine-tuning were first taken through a stringent pre-processing to normalize it and improve its quality. This started with the tokenization, and we used the AutoTokenizer class in the Hugging Face transformers library. This action is necessary to make sure the text is divided by the same WordPiece tokenization scheme that mBERT and XLM-R models were conditioned on. The procedure that took place was the normalization process that entailed changing all texts to lowercase. In the sentiment analysis task, to cleanse our data, we similarly applied regular expressions to drop URLs, email addresses, and social media handles like those beginning with the character @ like a Twitter mention. Lastly, noise removing was done to deprive any left-over HTML tags or artifacts that were not linguistic.

Any sentences less than three tokens following cleaning were dropped in order to guarantee the low data quality. This pre- processing was necessary so that there would be consistency between datasets and also reduce noise and improve the effectiveness of the further training of the model.

3.3. Model Selection

The object of the investigation is two multilingual models (mBERand XLM- R) that are pre-trained. mBERT([1]) also captures common linguistic patterns by using a single BERT architecture extended to support multiple languages si- multaneously. Having more scalability and strength, XLM-R ([9]) enhances the possibilities of mBERT and is highly efficient in cross-lingual knowledge transfer. The two models are famously known due to their cross lingual generalisation of sentences which makes them suitable in the instances that demand low-resource languages.

3.4. Rationale for Model Choice

In this case, mostly mBERT and XLM-R were selected mostly due to the pre-training on large multilingual corpora, as they can employ learnt language representations to target tasks during the optimisation. This capacity is quite beneficial considering that there is not a lot of labelled data in low-resource languages. These models are able to maintain reliable performance without huge task specific data due to the use of pre -trained information.

3.5. Training and Fine-Tuning Approach

After pre-processing, the chosen models are refined with the help of the datasets on specific NLP tasks. The founda- tion models are supplemented with task-specific layers to address sentiment analysis, NER, and machine translation independently. Fine-tuning parameters such as batch size and learning rate are carefully adjusted to maximise performance. The implementation will make sure that the models are able to fit in any job and language without fitting the sparse data. In conclusion, as a solution, the multi-lingual pre-trained models, standardised pre-processing and data curation were proposed to address low-resource NLP problems. The method also targets the effective utilisation of cross-lingual transfer learning but with the upholding of the model generalisation and the data consistency.

4. Results

The research will show that multilingual models like mBERT and XLM-R are applicable in any of the low-resource languages in a proposed way by following these steps sequentially. The datasets that were selected to test the methodology to be proposed included the language structure and the semantic information of the languages. The Universal Dependencies (UD) Project, a resource that provides syntactic and morphological annotations to a variety of languages, was helpful in the training and evaluation of multilingual NLP models. Also, labelled datasets for sentiment analysis, which refer to sentiment in various languages, were acquired from trustworthy sources in the academic literature.

4.1. Dataset Description

Datasets including both linguistic structure and semanticunderstanding across several languages were chosen in order to assess the suggested methodology. Multilingual NLP models can be trained and evaluated using the Universal Dependen- cies (UD) Project, which offers syntactic and morphological annotations for a large number of languages. Furthermore, labeled sentiment analysis datasets with sentiment annotations in many languages were obtained from reliable academic sources.Together, these datasets enable comprehensivensive experimentation, allowing the models to be assessed on both cross-lingual syntactic parsing and sentiment prediction tasks.This selection ensures robust evaluation, particularly for low- resource language scenarios, demonstrating the versatility and effectiveness of the proposed approach.

4.2. Experimental Setup

Each task dataset was stratified sampled into a 70 percent training sample, a 15 percent validation sample and a 15 percent test sample to have a uniform distribution of labels between splits. The hyper-parameter of the models were optimized by making the grid search and cross-validation on the validation set and then the fine-tuning process is performed with the following parameters; mBERT learning rate 5×105 and XLM-R learning rate 1×105; mBERT and XLM-R batch size 32 and 16 respectively. To prevent overfitting, a dropout of 0.1 was used on both models. The mBERT and XLM-R pre-trained

weights were adapted to the task-specific training data to sentiment analysis, NER, and machine translation using the iterative parameter optimization process to tune the models to the target languages and target task.

4.3. Evaluation Metrics

To obtain a clear and accurate assessment, we made separate models on each language-task condition (e.g., a particular model on the sentiment analysis of Hindi, a different model on the sentiment analysis of Telugu, etc.). This strategy will make the model to not aver- age its performance but will enable the evaluation of its functionality on each of the languages that has plenty of resources separately, i.e. sentiment analysis (I) and Named Entity Recognition (NER) (II). For machine translation, the Bilingual Evaluation Understudy (BLEU) scores are presented in III, reflecting the quality of translation from English to each target language.

Table 1: Performance on Sentiment Analysis

Language	Model	Accuracy	Precision	Recall	F1-score
Hindi	mBERT	87.9	89.5	88.4	88.9
Hindi	XLM-R	88.6	90.1	89.2	89.6
Telugu	mBERT	86.2	88.1	87.0	87.5
Telugu	XLM-R	87.1	88.9	87.9	88.4

Table 2: Performance on Named Entity Recognition (NER)

Language	Model	Precision	Recall	F1-score
Hindi	mBERT	85.8	84.5	85.1
Hindi	XLM-R	86.7	85.2	85.9
Telugu	mBERT	83.9	82.6	83.2
Telugu	XLM-R	84.8	83.3	84.0

Table 3: BLEU Scores for Machine Translation

Language Pair	Model	BLEU Score
English to Hindi	mBERT	31.2
English to Hindi	XLM-R	32.9
English to Telugu	mBERT	29.8
English to Telugu	XLM-R	31.3

4.4. Example Results

Figures 2 illustrate sample output from the language processing system. The screenshots display the original sentence, translated sentence, sentiment prediction, named entities, and audio output.

4.5. Interpretation of Results

The experiments demonstrate the robust performance of mBERT and XLM-R on the target low-resource languages of Hindi and Telugu. I has both Hindi and Telugu, XLM-R achieved higher scores in accuracy, precision, recall, and F1-score. The performance on Hindi was slightly better than on Telugu for both models, which may be due to Hindi's larger representation in the pre-training corpora of these multilingual models. II in the NER task, XLM-R again showed a clear advantage, yielding a higher F1-score for both languages.The performance gap between Hindi and Telugu was more pronounced in

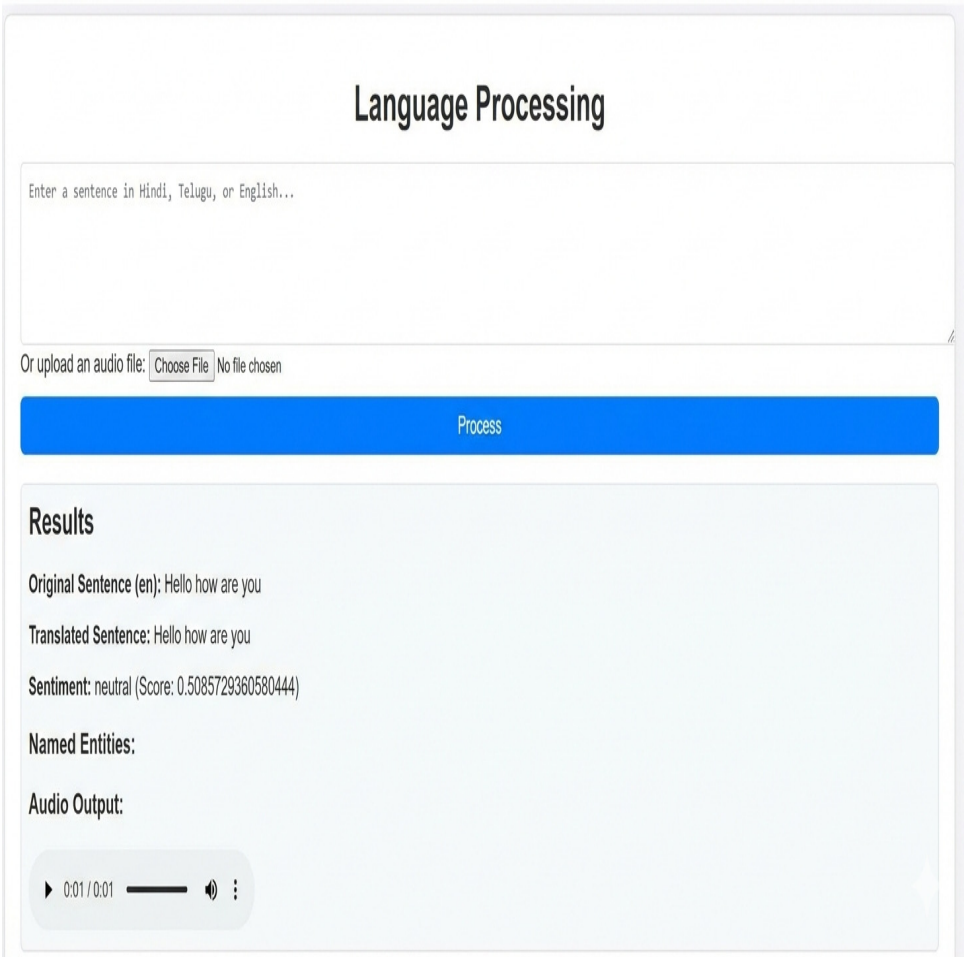

Figure 2: Example English sentence processing result showing neutral sentiment.

this task, suggesting that NER is more sensitive to the amount of linguistic information captured during pre- training. III has BLEU scores confirm the trend, with XLM-R producing higher-quality translations for both English-to-Hindi and English-to-Telugu pairs. The score of 32.9 for the English- to-Hindi pair with XLM-R is a strong result for a low-resource language pair.

4.6. F.Practical Implications

The findings highlight the applicability of pre-trained multilingual models in quick implementation of NLP applications in low-resource languages that do not require large amounts of labeled data. The model expedites the growth process and democratizes access to state-of-the-art language technologies on underserved linguistic groups.

This research offers valuable information on the efficiency of cross-lingual transfer learning, specifically with the latest models, including mBERT and XLM-R to overcome the difficulties presented by the lack of data in low-resource language. The empirical evidence of the per- formance of these models in various linguistic contexts makes us emphasize the effect of linguistic similarities and typological aspects on the results of transfer learning. The paper will aim to demonstrate the successful implementation of multilingual models like mBERT and XLM-R on various low-resource languages through a systematic series of steps. Datasets including both linguistic structure and semantic understanding across several languages were chosen in order to assess the suggested methodology. Multilingual NLP models can

be trained and evaluated using the Universal Dependencies (UD) Project, which offers syntactic and morphological annotations for a large number of languages. Furthermore, labelled sentiment analysis datasets with sentiment annotations in many lan- guages were obtained from reliable academic sources.

Datasets including both linguistic structure and semantic understanding across several languages were chosen in order to assess the suggested methodology. Multilingual NLP models can be trained and evaluated using the Universal Dependencies (UD) Project, which offers syntactic and morphological anno- tations for a large number of languages. Furthermore, labelled sentiment analysis datasets with sentiment annotations in many languages were obtained from reliable academic sources.

5. Conclusion

This study investigated how multilingual mod- els—specifically mBERT and XLM-R—improve natural language processing (NLP) performance for the low-resource languages of Hindi and Telugu, primarily by leveraging knowledge transferred from high-resource English. Through extensive testing, we discovered that these models consistently enhance important performance metrics, such as BLEU scores in machine translation jobs and accuracy, precision, recall, and F1-score in sentiment analysis and named entity recognition (NER).Our findings demonstrate the robustness of cross-lingual transfer learning by showing how it may be used to improve natural language processing (NLP) skills in languages with little data by utilising knowledge from languages with abundant resources. These results not only support the usefulness of mBERT and XLM-R in practice, but they also provide insightful information for further study and applications in multilingual NLP, especially for low-resource.

6. Future Scope

Future research should aim to broaden the range of low- resource languages studied and explore more efficient fine- tuning strategies to maximize model performance when data is limited. Examining alternative multilingual models, as well as hybrid approaches, could further improve the adaptability and scalability of cross-lingual transfer learning across diverse NLP tasks. Additionally, investigating the role of linguistic typology and its influence on transfer learning effectiveness can offer valuable insights for designing model architectures tailored to specific language families or dialects. These directions hold promise for advancing both the theoretical understanding and practical applications of multilingual NLP in low-resource settings.

Acknowledgments

We think the referee by your suggestions.

References

1. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Proc. NAACL-HLT, pp. 4171–4186, (2019).
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., *Attention Is All You Need*, Advances in Neural Information Processing Systems, Vol. 30, pp. 5998–6008, (2017).
3. Liu, X., He, P., Chen, W., and Gao, J., *Multi-task Deep Neural Networks for Natural Language Understanding*, arXiv preprint arXiv:1901.11504, (2019).
4. Pires, T., Schlinger, E., and Garrette, D., *How Multilingual is Multilingual BERT?*, arXiv preprint arXiv:1906.01502, (2019).
5. Ahmad, W. U., Zhang, Z., Ma, J., Hovy, E., Chang, K.-W., and Peng, N., *Cross-lingual Named Entity Recognition with Minimal Resources Using Neural Machine Translation*, arXiv preprint arXiv:2010.12496, (2020).
6. Wang, Y., Li, C., Liu, X., He, P., Chen, W., and Gao, J., *InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training*, arXiv preprint arXiv:2007.07834, (2021).
7. Ahmad, W. U., and Hovy, E., *Cross-lingual Transfer Learning for Low-resource Named Entity Recognition*, Proc. 12th Int. Conf. on Natural Language Processing (ICON-2022), (2022).
8. Google Research, *XLM-R: Unsupervised Cross-lingual Representation Learning at Scale*, (2022).
9. Universal Dependencies Project, *Multilingual Corpora for Dependency Parsing*, (2021).
10. Ushashree, P., *Sentiment Analysis to Detect Depression in Social Media Users: Overview and Proposed Methodology*, Emerging Research in Computing, Information, Communication and Applications, (2021).

11. Ushashree, P., Naik, A., and Sri, P. A. S., *AI-driven Health: A Web App for Enhanced Healthcare Queries and Nutrition Analysis*, Proc. 5th Int. Conf. on Smart Electronics and Communication, (2024).

12. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L., *Multilingual Denoising Pre-training for Neural Machine Translation*, Transactions of the Association for Computational Linguistics, 8, 726–742, (2020).

13. Pires, T., Schlinger, E., and Garrette, D., *How Multilingual is Multilingual BERT?*, Proc. ACL 57th Annual Meeting, pp. 4996–5001, Florence, Italy, (2019).

14. Wu, S., and Dredze, M., *Beto, Bentz, Becas: The Surprising Cross-lingual Effectiveness of BERT*, Proc. EMNLP–IJCNLP, pp. 833–844, Hong Kong, (2019).

15. Yang, J., Ma, S., Zhang, D., Wu, S., Li, Z., and Zhou, M., *Alternating Language Modeling for Cross-Lingual Pre-Training*, Proc. AAAI-34, (2020).

Ushashree P.,
Department of CSE Geethanjali College of Engineering and Technology Hyderabad,
India.
E-mail address: ushashree.sgs@gmail.com

and

Vansh Badani,
Department of CSE Geethanjali College of Engineering and Technology Hyderabad,
India.
E-mail address: 23r11a0593@gcet.edu.in

and

Nikilesh,
Department of AIML Geethanjali College of Engineering and Technology Hyderabad,
India.
E-mail address: 23r11a66h7@gcet.edu.in

and

B. Nagamani,
Department of Freshman Engineering Geethanjali College of Engineering and Technology Hyderabad,
India.
E-mail address: nagamani.english@gcet.edu.in