Performance Analysis of Effective Retrieval of Kannada Translations in Code-Mixed Sentences using BERT and MPnet

H. P. Rohith
Deptartment of ISE, Nitte Meenakshi Institute of Technology, Bangalore, India
rohith.hp@nmit.ac.in

Lava Kumar
Deptartment of CSE, B.M.S. College of Engineering, Bangalore, India
lavakumar.kushi8197@gmail.com (corresponding author)

Sooda Kavitha
Deptartment of CSE, B.M.S. College of Engineering, Bangalore, India
kavithas.cse@bmsce.ac.in

Rai B. Karunakara
Deptartment of Electronics & Communication, Nitte Meenakshi Institute of Technology, Bangalore, India
karunakara.rai@nmit.ac.in

K. P. Inchara
Deptartment of CSE, B.M.S. College of Engineering, Bangalore, India
inchara.scs22@bmsce.ac.in

Received: 17 September 2024 | Revised: 8 October 2024 | Accepted: 16 November 2024

ABSTRACT

Translating Kannada-English (Kn-En) code-mixed text is a challenging task due to the limited availability of Kannada language resources and the inherent complexity of the dataset. This study evaluates the effectiveness of the sentence transformer model, utilizing pre-trained multilingual MPNet and Bidirectional Encoder Representations from Transformers (BERT) architectures, in generating sentence embeddings to enhance translation accuracy. It encodes both code-mixed sentences and their corresponding Kannada translations into high-dimensional embeddings. By employing cosine similarity, it maps input sentences to their closest translations, encoding 2000 code-mixed sentences and their translations using both the MPNet and BERT models. The findings indicate that the MPNet model proved to be more effective, achieving a model accuracy of 98%, compared to BERT's 88%. Moreover, MPNet outperformed BERT in terms of Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores, attaining 85.0 and 80.0, respectively, while BERT scored 65.3 and 58.7. These results highlight the advanced capabilities of MPNet in translating code-mixed languages and its potential applicability to a broader range of multilingual Natural Language Processing (NLP) tasks.

Keywords-BERT; kannada-english code-mix; MPNet; multilingual; natural language processing; translation

I. INTRODUCTION

India's linguistic and social diversity is immense, encompassing seven hundred eighty languages, twenty-two of which are officially recognized. Furthermore, twenty-one official languages and ten non-scheduled languages are spoken by more than a million people [1, 2]. The Indic languages can be categorized into four primary families: Sino-Tibetan, Austro-Asiatic, Dravidian, and Indo-Aryan. This remarkable diversity underscores the crucial role of translation services for effective inter-state communication. The rise of code-mixed

Kn-En datasets reflects the growing multilingual nature of digital interactions. Code mixing is a common phenomenon among bilingual individuals and emphasizes the need for advanced language processing methods. The sentence transformer model, particularly the pre-trained multilingual MPNet, offers a promising approach to address these complexities by providing effective phrase embedding through masked and permuted language modeling.

The rise of digital communication has led to an increased use of code-mixed languages, where speakers blend vocabulary from multiple languages. Code-mixed Kn-En are common in bilingual areas, but traditional NLP models, designed for monolingual text, struggle with this complexity. Despite the growing prevalence of code-mixed language, resources for effectively handling it, especially for minority languages like Kannada, remain scarce. This challenge results in suboptimal translation performance due to the complex nature of code-mixed sentences. This study aims to convert mixed Kn-En sentences into pure Kannada, with the goal of developing an intuitive system that facilitates easy input and provides accurate Kannada translations.

II. LITERATURE SURVEY

The literature survey reveals a lack of substantial studies addressing the challenges of handling Kn-En code-mixed text, such as dataset scarcity, sentiment analysis, and Language Identification (LI). However, certain methods have been explored for sentiment analysis in other languages, such as in Malayalam-English text, using Deep Learning (DL) and Machine Learning (ML), as well as data augmentation techniques to generate synthetic samples. Additionally, several studies have explored Neural Machine Translation (NMT) challenges for Indic languages, language augmentation for BERT-based models, and pivot-based NMT for Kn-En translation. Furthermore, research has focused on LI and the development of specialized datasets for code-mixed Kn-En text. Offensive language detection in Dravidian languages using MPNet and CNN has been studied, and techniques, like supervised contrastive learning, have been proposed to address the challenges of optimizing pre-trained models for Natural Language Understanding (NLU).

FeaMix has beed introduced, as a novel data augmentation approach that enhances feature mixing in self-consistency learning by leveraging memory batches [3]. This technique uniquely selects samples to maintain the original spatial distribution and extends self-consistency learning to code-related language tasks, achieving state-of-the-art performance on the CoNaLa and CodeTrans benchmarks. However, the BERT model struggles with code-mixed languages, such as Kn-En, due to the limited availability of high-quality annotated datasets. Additionally, LI and normalization techniques for code-mixed text have been considered, outlining various methodologies and highlighting the associated challenges and limitations [4].

NMT and its complexities and advancements have been studied, with a particular focus on Indian languages [5, 6]. The proposed method examines the challenges of translating Indic languages using NMT models and explores linguistic nuances specific to these languages, aiming to enhance translation fluency and accuracy. The current research examines efforts to improve neural machine translation for Indic languages, promoting cross-linguistic communication and understanding through a comprehensive assessment of methodologies and strategies. Additionally, the study aims to analyze sentiment in Kannada text that has been code-mixed, addressing the challenges of sentiment analysis in multilingual environments where Kannada is combined with other languages [7, 8]. The project seeks to devise efficient techniques for comprehending the emotions conveyed in code-mixed languages, taking into account the linguistic complexities involved. The research endeavors to offer insights into the sentiments expressed in code-mixed Kannada text through the utilization of sophisticated sentiment analysis algorithms. The findings have the potential to find applications in a variety of fields, such as opinion mining, social media monitoring, and customer feedback analysis.

A novel NMT approach that does not rely on parallel corpora to translate text from English to Kannada has been investigated [9]. The aim is to bridge the translation gap between the two languages by leveraging pre-trained language models. The project seeks to overcome the lack of parallel data for Kannada through the use of unsupervised algorithms. Pre-trained language models reduce the amount of substantial training data required and enable the creation of accurate translations. This addresses the issue of LI in code-mixed Kn-En writings, where multiple languages are used within the same text [10]. With a focus on sequence labeling problems, where each word is assigned a language label from a preset set, previous work on LI tasks, involving ML, DL, and Transfer Learning (TL), is examined.

Word-level LI in code-mixed Kn-En texts, which is crucial in multilingual contexts like India, has been investigated [11]. The task involves classifying words into six categories and utilizes a dataset reflecting real-world code-mixed usage. The results demonstrate enhancements in identifying languages and word categories within such texts. Additionally, the significance of detecting offensive language on social media, particularly for Dravidian languages, such as Tamil, Malayalam, and Kannada, is explored [12].The challenges of classifying code-mixed comments are examined and the effectiveness of MPNet and Convolutional Neural Networks (CNN) models in offensive language detection is evaluated by comparing their performance using metrics like the weighted average F1-score against baseline models. Furthermore, the benefits of MPNet and Conditional Token Masking (CTM) in capturing context-aware themes from unstructured user feedback have been explored [13]. The particular survey also covers hyperparameter optimization and coherence measures, underscoring improvements in user satisfaction and product development.

Challenges in fine-tuning language models for NLU applications, have been investigated, identifying limitations of the cross-entropy loss function in capturing contrastive information [14]. A Supervised Contrastive Learning (SCL) approach has been introduced, which enhances text classification performance by creating a margin between classes and incorporates text augmentation techniques. The paper further compares the SCL-tuned models to MPNet and DeBERTaLarge, demonstrating advancements in performance and generalizability.

III. IMPLEMENTATION

A. Data Collection and Preprocessing Dataset

The dataset highlights various types of information and their corresponding Kannada translations, which play a crucial role in the conversion of Kn-En code-mixed sentences into complete Kannada sentences, as presented in Table I.

TABLE I. OVERVIEW OF DATASET EXAMPLES AND KANNADA TRANSLATIONS

Category	Examples	Kannada Translation
Units	cm, mm, ltr, kg	ಸೆಂ.ಮೀ, ಮಿ.ಮೀ, ಲೀಟರ್, ಕೆ.ಜಿ
Acronyms	Mr., Dr., CM, DM, PM, GM	ಶ್ರೀ., ಡಾ., ಮುಖ್ಯಮಂತ್ರಿಗಳು, ಜಿಲ್ಲಾಧಿಕಾರಿ, ಪ್ರಧಾನ ಮಂತ್ರಿ, ಜನರಲ್ ಮ್ಯಾನೇಜರ್
Dates	07-Dec-2015	ಏಳನೇ ಡಿಸೆಂಬರ್ ಎರಡು ಸಾವಿರದ ಹದಿನೈದು
Ordinal Numbers	1st, 2nd, 3rd	೧ನೇ, ೨ನೇ, ೩ನೇ
Kn-En Code-Mixed Sentence	Naanu ಸಿನಿಮಾ nodbeku weekend ge	ನಾನೂ ಸಿನಿಮಾ ನೋಡ್ಬೇಕು ವೀಕೆಂಡ್ ಗೆ

The incorporation of a variety of sentence structures strengthens the model's robustness and enhances its capacity to generalize across diverse translation scenarios. The dataset utilized for training and evaluating the translation model comprises Kn-En code-mixed sentences and their corresponding Kannada translations. The dataset was meticulously curated from multiple sources, as detailed below:

- Manual Collection: Sentences and translations were manually curated from real-world sources, such as social media, online forums, or other platforms where code-mixed language is prevalent. This method involves careful annotation to ensure accuracy.
- OpenAI or GPT-based Models: Utilizing Large Language Models (LLMs), such as GPT-4, synthetic data were generated for different phrase structures and situations, with the purpose of translating Kn-En code-mixed words into Kannada.

Text containing emoji's, excessive spaces, needless characters, and other noise is frequently encountered when collecting code-mixed Kn-En data for Kannada translation [16]. To prepare the text, the latter undergoes a process of cleaning and preprocessing. The sentence is then segmented into individual words and punctuation, with a distinction made between Kannada and English terms. Kannada words written in the Latin script are rendered into the native Kannada script, while English words are consistently converted to lowercase. This preprocessing intervention ensures that the text is cohesive and well-structured, thereby enhancing the accuracy and effectiveness of translating code-mixed Kn-En sentences into complete Kannada sentences.

B. Sentence Transformer and MPNet

The sentence transformer library offers readily available pre-trained models for generating sentence embeddings. These high-dimensional vector representations effectively capture the semantic meaning of sentences. The embeddings have numerous applications in NLP, including semantic search, document clustering, and text similarity analysis.

MPNet, also referred to as MPNet understanding, is a transformer-based pre-trained language model that integrates multiple key elements for effective language representation [17]. It combines the strengths of two approaches, namely Masked Language Modeling (MLM) and Permuted Language Modeling (PLM). MLM aids the model in comprehending context by predicting missing words in a sentence, while PLM enhances its ability to capture dependencies between tokens by predicting the next token in a shuffled sequence. This unified approach enables MPNet to outperform other models on various NLP tasks, such as text embedding. By blending the advantages of PLM and MLM, MPNet stands out from other context learning models [18]. This hybrid technique allows MPNet to efficiently capture long-range dependencies and bidirectional context. In contrast to standard models that solely utilize fixed order sequences, MPNet addresses complex and non-linear settings by permuting the input sequence and predicting masked tokens.

$$L(PLM) = \sum_{t=1}^{T} \log p(Word_t | Word_{<t}) \quad (1)$$

$$L(MLM) = \sum_{i=1}^{m} \log p(Word_i(Word - i)) \quad (2)$$

where *Word* represents the masked token at position *t* or *i*, $Word_{<t}$ represents the sequence of tokens before *t*, *Word*(*Word−i*) denotes the sequence of tokens excluding the masked token, *T* is the total number of tokens in the sequence, and *M* is the set of masked positions.

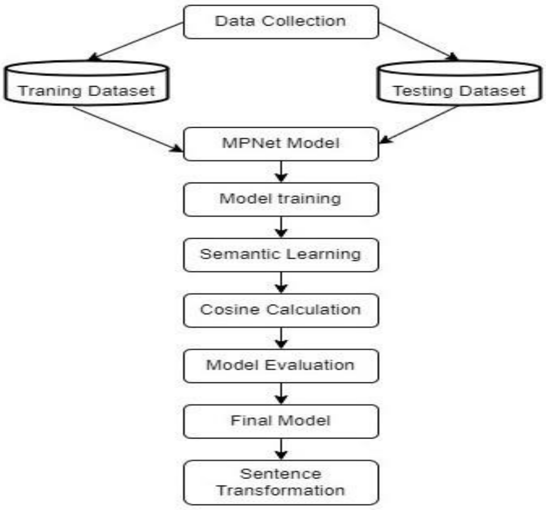

Fig. 1. Methodology flowchart.

When addressing Kn-En code-mixed language, where the structure may be irregular and context-dependent, this capability proves quite valuable. Overall, MPNet's distinctive architecture and training approach enable it to better represent and comprehend the interplay between multiple languages, resulting in superior performance in tasks involving code-mixed language translation. The sentence transformer library is employed to load the paraphrase-multilingual-mpnet-base-v2 model, as depicted in Figure 1. This specific model is a pre-trained variant of MPNet that is designed to manage multiple languages effectively.

TABLE II. SUMMARY OF EMBEDDING'S ROLE AND WORKING FOR PLM AND MLM.

Method	Embedding's Role and Workflow
PLM	Uses embedding analysis to assist in rearranging words into their original arrangement. It creates embedding's/embeddings for every word, and then utilizes a permuted version to predict the correct sequence, as depicted in (1).
MLM	Uses the context of the words around it to fill in the missing words. It converts each word into an embedding to mask a word/it, and predicts using the context provided by other words, as shown in (2).

Table II provides an overview of the function and operation of embeddings in MLM and PLM. Although MLM places more emphasis on predicting masked words by leveraging the context given by surrounding tokens, PLM concentrates on guessing the original word sequence from a permuted version through embedding analysis.

C. Embedding with MPNet and Saving Model

The semantic content and contextual information of text are captured through high-dimensional vector representations, known as embeddings, which are generated by the MPNet model. To comprehend the intricate relationships and nuances inherent in language, the model's layers process text using a combination of masked and permuted language modeling techniques, resulting in these embeddings. The efficient representation of words and sentences provided by MPNet's pre-trained embeddings can be leveraged to enhance a variety of NLP tasks, such as text similarity analysis, machine translation, and text classification. Once the MPNet model is trained, it encodes sentences into embeddings high dimensional vectors capturing the semantic meaning of each sentence:

- Encoding: Both the code-mixed sentences and their Kannada translations are encoded into embeddings using MPNet.
- Saving: These embeddings, along with the Kannada translations, are saved to files mixed_embeddings.pkl and kannada_translations.pkl for efficient retrieval and comparison during inference.

D. Cosine Similarity Calculation

MPNet deploys cosine similarity to measure semantic similarity between sentences and phrases, a crucial step in NLP tasks, like translation and information retrieval, and is computed using (3):

$$Cosine\ Similarity = \frac{A \cdot B}{|A||B|} \quad (3)$$

where A represents the vector of the new input, which might be a word or sentence and B represents the trained or reference input. A encodes the main characteristics or features of the new input into a numerical format and B is essentially a training set of words or phrases. By comparing these two vectors, the cosine similarity metric provides an indication of the level of semantic or content-based resemblance between the new input and the reference input, which can assist in assessing how closely the new input aligns with the reference input. The magnitudes of the two vectors are represented by A and B, while their alignment is reflected in the dot product of A and B. A value of cosine similarity near 1 signifies a high degree of similarity between the vectors, whereas a value close to 0 suggests a lack of similarity.

In the context of MPNet, cosine similarity helps in comparing the contextual embeddings of input sentences with pre-computed embeddings of reference sentences. This comparison enables tasks, such as finding the most similar sentence or translating input sentences into target languages, by selecting the most contextually relevant match from a set of predefined translations.

E. Translating New Inputs

When a new code-mixed sentence is provided, it is translated using the following process:

- Encoding input and similarity calculation: The new sentence is encoded into an embedding using the MPNet model. The similarity between the input embedding and stored embeddings is computed deploying methods like cosine similarity.
- Retrieving translation: The Kannada translation corresponding to the most similar embedding from the dataset is retrieved and presented as the output.

The process begins with a new code-mixed sentence provided by the user. This sentence is then encoded using the pre-trained sentence transformer model. The resulting embedding is compared to precompute embeddings from the dataset to find the closest match. Once the closest match is identified, the corresponding translation is retrieved and presented to the user. This process leverages MPNet's ability to handle code-mixed sentences effectively, ensuring accurate translations by understanding the contextual nuances of the input text.

IV. RESULTS

The sentence transformer model employing a pre-trained MPNet architecture was utilized as the translation tool in this highly effective Kn-En code-mixed text translation approach. By encoding a dataset of 2000 code-mixed Kn-En sentences, which included diverse elements, such as numbers, dates, and years, along with their corresponding Kannada translations into high-dimensional embeddings, a remarkable translation accuracy of 98% was achieved. This performance underscores the MPNet model's ability to effectively capture and represent the semantic meanings of code-mixed sentences, including their contextual and numerical information.

Table III presents translations for different types of inputs, including numbers, dates, and Kn-En code-mixed sentences. It shows each input alongside its expected Kannada translation and the translation provided by the MPNet model. This comparison highlights the accuracy and effectiveness of MPNet in translating code-mixed text.

TABLE III. COMPARISON OF EXPECTED VS. MPNET AND BERT TRANSLATIONS

Code-Mixed Sentence	**Kannada Translation**	**MPNet Translation**	**BERT Translation**
Mr.,The coffee smells good ಕುಡಿಯೋ?	ಶ್ರೀ ,ಕಾಫಿಯ ಪರಿಮಳ ಚೆನ್ನಾಗಿದೆ ಕುಡಿಯೋ?	ಶ್ರೀ ,ಕಾಫಿಯ ಪರಿಮಳ ಚೆನ್ನಾಗಿದೆ ಕುಡಿಯೋ?	Mr., ಕಾಫಿಯ ಪರಿಮಳ ಚೆನ್ನಾಗಿದೆ, ಕುಡಿಯೋ?
2008	ಎರಡು ಸಾವಿರದ ಎಂಟು	ಎರಡು ಸಾವಿರದ ಎಂಟು	೨೦೦೮
7-Dec-15	ಏಳನೇ ಡಿಸೆಂಬರ್ ಎರಡು ಸಾವಿರದ ಹದಿನೈದು	ಏಳನೇ ಡಿಸೆಂಬರ್ ಎರಡು ಸಾವಿರದ ಹದಿನೈದು	ಏಳನೇ ಡಿಸೆಂಬರ್ 2015
I missed the bus today, so late agide office ge.	ನಾನು ಇಂದು ಬಸ್ ಮಿಸ್ ಮಾಡಿದೆ, ಅದಕ್ಕೆ ಆಫೀಸ್ ಗೆ ತಡವಾಗಿದೆ.	ನಾನು ಇಂದು ಬಸ್ ಮಿಸ್ ಮಾಡಿದೆ, ಅದಕ್ಕೆ ಆಫೀಸ್ ಗೆ ತಡವಾಗಿದೆ.	ನಾನು ಇಂದು ಬಸ್ ಮಿಸ್ ಮಾಡಿದೆ, ಆಫೀಸ್ ಗೆ ತಡವಾಗಿದೆ.
12/07/2009	ಹನ್ನೆರಡು ಜುಲೈ 2009	ಹನ್ನೆರಡು ಜುಲೈ ಎರಡು ಸಾವಿರ ಒಂಬತ್ತು	12/07/2009

TABLE IV. MODEL METRIC COMPARISON

Metric	**BERT**	**MPNet**
BLEU Score	65.3	85.0
ROUGE Score	58.7	80.0
Context Handling	Less Effective	More Effective
Training Time	10 hours	8 hours
Fluency	7/10	8/10
Accuracy	7/10	9/10
Naturalness	6/10	9/10

Overall, MPNet significantly outperformed BERT, which attained an accuracy of 88%, achieving an estimated model accuracy of 98%, making it a better choice for translating Kn-En code-mixed sentences into Kannada. The BLEU and ROUGE scores, which measure the overlap between machine-generated and human translations, demonstrated that MPNet outperformed BERT significantly, with values of 85.0 and 80.0 compared to 65.3 and 58.7, respectively. These high scores reflect MPNet's capability to generate translations that are more accurate and closely aligned with human-produced translations. Furthermore, it was found that the translation error rate was notably lower for MPNet, 2.5%, compared to the 12.5% of BERT, indicating fewer mistakes made by the MPNet model.

Additionally, human evaluators rated MPNet's translations higher in terms of fluency, accuracy, and naturalness. Fluency refers to the smoothness and natural flow of the translation, accuracy measures the correctness of the translation, and naturalness assesses how native-like the translation sounds. MPNet's advanced permuted language modeling allows it to better comprehend context and dependencies, enhancing its effectiveness in handling the nuances of code-mixed language.

Figure 2 demonstrates that as the percentage of training data increases so does the model's accuracy. This suggests a positive relationship between the quantity of training data and the model's performance.

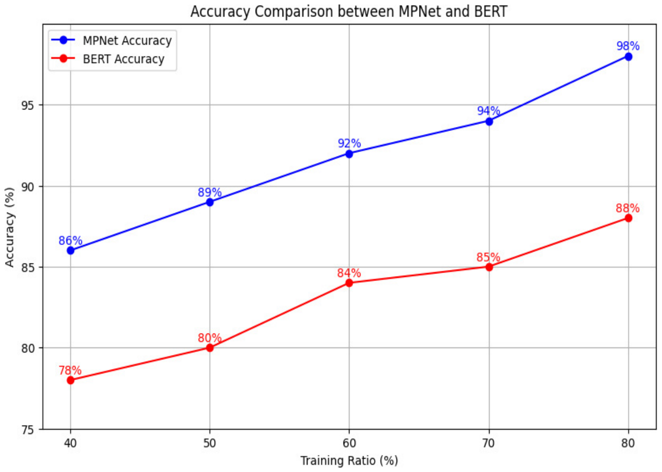

Fig. 2. Model accuracy comparison across training ratio.

V. CONCLUSION AND FUTURE WORK

In this study, MPNet substantially outperforms Bidirectional Encoder Representations from Transformers (BERT) in translating Kannada-English (Kn-En) code-mixed sentences. MPNet attained Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores of 85.0 and 80.0, in contrast to BERT's 65.3 and 58.7, and exhibited a lower translation error rate of 2.5% compared to BERT's 12.5%. MPNet's advanced permuted language modeling enhances its contextual understanding, leading to a model accuracy of 98%, making it a superior choice for this task.

As far as is known, this study represents the first application of MPNet's advanced permuted language modeling to the relatively under-investigated domain of Kn-English code-mixed language translation. By leveraging this modeling approach, MPNet is able to address a limitation in BERT's conventional masked language modeling, namely its inability to effectively capture various contextual connections inherent in code-mixed phrases. Future research will focus on expanding the dataset, exploring domain-specific adaptations, and combining MPNet with other advanced models to further enhance translation accuracy and system efficiency.

REFERENCES

[1] A. R. Jafari, B. Heidary, R. Farahbakhsh, M. Salehi, and N. Crespi, "Language Models for Multi-Lingual Tasks - A Survey," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 6, 2024, https://doi.org/10.14569/IJACSA.2024.01506146.

[2] R. Chundi, V. R. Hulipalled, and J. B. Simha, "SAEKCS: Sentiment Analysis for English – Kannada Code SwitchText Using Deep Learning Techniques," in *Proceeeding of International Conference on Smart Technologies in Computing, Electrical and Electronics*, Bengaluru, India, Oct. 2020, pp. 327–331, https://doi.org/10.1109/ICSTCEE49637.2020.9277030.

[3] S. Zhao, J. Tian, J. Fu, J. Chen, and J. Wen, "FeaMix: Feature Mix With Memory Batch Based on Self-Consistency Learning for Code Generation and Code Translation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–10, 2024, https://doi.org/10.1109/TETCI.2024.3395531.

[4] A. Mangla, R. K. Bansal, and S. Bansal, "Language Identification and Normalization Techniques for Code-Mixed Text," in *Proceeeding of Sixth International Conference on Computational Intelligence and Communication Technologies*, Sonepat, India, Apr. 2024, pp. 435–441, https://doi.org/10.1109/CCICT62777.2024.00077.

[5] S. K. Sheshadri, D. Gupta, and M. R. Costa-Jussà, "A Voyage on Neural Machine Translation for Indic Languages," *Procedia Computer Science*, vol. 218, pp. 2694–2712, Jan. 2023, https://doi.org/10.1016/j.procs.2023.01.242.

[6] G. Takawane, A. Phaltankar, V. Patwardhan, A. Patil, R. Joshi, and M. S. Takalikar, "Language augmentation approach for code-mixed text classification," *Natural Language Processing Journal*, vol. 5, Dec. 2023, Art. no. 100042, https://doi.org/10.1016/j.nlp.2023.100042.

[7] S. Dutta, H. Agrawal, and P. K. Roy, "Sentiment Analysis on Multilingual Code-Mixed Kannada Language," *Forum for Information Retrieval Evaluation*, pp. 908–918, Dec. 2021.

[8] H. Gadugoila, S. K. Sheshadri, P. C. Nair, and D. Gupta, "Unsupervised Pivot-based Neural Machine Translation for English to Kannada," in *Proceedings of 19th India Council International Conference*, Kochi, India, Nov. 2022, pp. 1–6, https://doi.org/10.1109/INDICON56171.2022.10039732.

[9] S. K. Sheshadri, B. Sai Bharath, A. Hari Naga Sree Chandana Sarvani, P. Reddy Vijaya Bharathi Reddy, and D. Gupta, "Unsupervised Neural Machine Translation for English to Kannada Using Pre-Trained Language Model," in *Proceeding of 13th International Conference on Computing Communication and Networking Technologies*, Kharagpur, India, Oct. 2022, pp. 1–5, https://doi.org/10.1109/ICCCNT54827.2022.9984521.

[10] H. L. Shashirekha, F. Balouchzahi, M. D. Anusha, and G. Sidorov, "CoLI-Machine Learning Approaches for Code-mixed Language Identification at the Word Level in Kannada-English Texts." arXiv, Nov. 17, 2022, https://doi.org/10.48550/arXiv.2211.09847.

[11] F. Balouchzahi *et al.*, "Overview of CoLI-Kanglish: Word Level Language Identification in Code-mixed Kannada-English Texts," in *Proceedings of the 19th International Conference on Natural Language Processing: Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, New Delhi, India, Sep. 2022, pp. 38–45.

[12] B. R. Chakravarthi, M. B. Jagadeeshan, V. Palanikumar, and R. Priyadharshini, "Offensive language identification in dravidian languages using MPNet and CNN," *International Journal of Information Management Data Insights*, vol. 3, no. 1, Dec. 2023, Art. no. 100151, https://doi.org/10.1016/j.jjimei.2022.100151.

[13] M. H. Asnawi, A. A. Pravitasari, T. Herawan, and T. Hendrawati, "The Combination of Contextualized Topic Model and MPNet for User Feedback Topic Modeling," *IEEE Access*, vol. 11, pp. 130272–130286, Nov. 2023, https://doi.org/10.1109/ACCESS.2023.3332644.

[14] H. Gao, B. Dong, Y. Zhang, T. Xiao, S. Jiang, and Y. Dong, "An Efficient Method of Supervised Contrastive Learning for Natural Language Understanding," in *Proceeding of 7th International Conference on Computer and Communications (ICCC)*, Chengdu, China, Dec. 2021, pp. 1698–1704, https://doi.org/10.1109/ICCC54389.2021.9674736.

[15] L. Kumar, "kn-En-code-mix-sentence-dataset." GitHub, Jul. 2024, https://github.com/lavakumar7619/kn-En-code-mix-sentence-dataset.

[16] L. Kumar, S. R. Vernekar, D. S. Shreevatsa, T. Srinivas, B. N. Gururaj, and K. Sooda, "Prediction Of Emotions In Kannada Sentence With Homonyms," in *Proceeding of International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications*, Bengaluru, India, Apr. 2024, pp. 1–5, https://doi.org/10.1109/ICETCS61022.2024.10543456.

[17] S. H S, K. Sooda, and B. Karunakara Rai, "EfficientNet-B7 framework for anomaly detection in mammogram images," *Multimedia Tools and Applications*, pp. 1–27, May 2024, https://doi.org/10.1007/s11042-024-18853-1.

[18] N. Sureja, N. Chaudhari, P. Patel, J. Bhatt, T. Desai, and V. Parikh, "Hyper-tuned Swarm Intelligence Machine Learning-based Sentiment Analysis of Social Media," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15415–15421, Aug. 2024, https://doi.org/10.48084/etasr.7818.