IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Language Identification For English-Punjabi (Romanized) Code-Mixed Social Media Text Using Transformers

[1]Sunita, [2]Ajit Kumar, [3]Neetika Bansal

[1]Research Scholar,[2]Associate Professor, [3]Assistant Professor.

[1]Computer Science

[1]Punjabi University, Patiala, Punjab, India

Abstract: Code-mixed language is the language of social media, where speakers switch between two or more languages ever more frequently. In a multilingual environment such as Punjab, the users usually write in code-mixed form, i.e., English mixed with Romanized Punjabi, which makes processing a bit difficult for Natural Language Processing (NLP). This paper presents a Transformers-based model for word-level language identification (LI) in English-Punjabi (Romanized) code-mixed text, with a training corpus of 150,000 annotated tokens. We conducted experiments that achieved an accuracy of 97.09%, outperforming traditional methods such as Conditional Random Fields (CRF) and Hidden Markov Models (HMM). Such a system could be utilized to develop efficient preprocessing tools for a wide range of NLP tasks, including part-of-speech (POS) tagging, sentiment analysis, and machine translation in code-mixing scenarios.

Keywords: Code-Mixed Text, Language Identification, English-Punjabi, Romanized Punjabi, BERT, Transformer Models, Social Media NLP

1.INTRODUCTION

The way communication has evolved in multilingual communities has been greatly influenced by the swift rise of social media platforms like Facebook, YouTube, and WhatsApp. One notable trend in this space is code-mixing. In India, especially in Punjab, social media users frequently blend English with Romanized Punjabi in their posts and conversations [1]. These code-mixed texts differ from traditional bilingual texts, as they often display erratic spelling, inconsistent transliteration, and unconventional grammar, which can present challenges for NLP systems [2].

Segmenting an utterance into words is critical, and once that task is completed, assigning a language to each word is vital in LI. It enables more advanced transformations for downstream tasks, including, but not limited to, POS tagging, sentiment analysis, and translation. More traditional techniques, such as CRF and HMM, as well as word-based style classifiers, are not inherently designed to handle even simple code-mixing problems [3]. As a result, these models are often poorly designed and heavily dependent on socially engineered features. Deep learning has made remarkable strides, particularly with the advent of transformer technologies like BERT. BERT is adept at tackling intricate contextual semantics in sequence labelling tasks, showcasing its strength in navigating language ambiguity and code-mixing thanks to its bidirectional self-attention mechanisms [4].

This paper presents an innovative approach utilizing Transformer frameworks to accurately identify individual words in code-mixed texts featuring Romanized Punjabi and English. To cater to our specific training and testing requirements, we created a dataset of 150,000 tokens, sourced from publicly available social media datasets. This curated dataset effectively captures the diverse spelling variations and syntactic structures

commonly found in everyday online interactions. Our experimental results demonstrate that this method achieves over 97.09% accuracy, surpassing all traditional machine learning techniques we evaluated.

This work makes the following contributions:

1. Introducing a comprehensive annotated dataset for English–Punjabi (Romanized) code-mixed social media text, featuring 150,000 tokens—an invaluable resource for research and applications in this vibrant linguistic space.
2. Creating a well-optimized Transformer-based model for identifying languages at the word level.
3. Achieving an outstanding accuracy of 97.09%, this approach clearly outshines traditional CRF and HMM-based methods.
4. Emphasizing the effectiveness of transformer models as valuable preprocessing tools for various downstream tasks in code-mixed NLP

The structure of this paper is as follows: In Section 2, we review existing research on LI for code-mixed text. Section 3 outlines our proposed approach, including the collection, preprocessing, and annotation of the dataset. In Section 4, we delve into our methodology, focusing on the implementation of the Transformer model. Section 5 presents our experimental results, highlighting the performance of our approach on a dataset of English-Punjabi code-mixed social media texts. Finally, in Section 6, we conclude from our findings and discuss potential directions for future research. We aim for this work to enhance the understanding of code-mixing on social media and its significance for NLP tasks.

2.LITERATURE REVIEW

Conventional LI models were primarily developed for use in monolingual or more structured bilingual settings. However, their effectiveness diminishes significantly when faced with informal, noisy, and transliterated texts commonly found on social media platforms. This challenge is further compounded by the frequent switching between languages and the lack of clear linguistic markers within the text. Considerable research has focused on the task of identifying languages within code-mixed texts, examining various language combinations, including Hindi-English, Spanish-English, by [5], and Bengali-English [6]. While traditional methodologies—such as HMMs, CRFs, and SVMs—have seen moderate success in differentiating between languages in code-mixed datasets, the unique properties of informal social media text present ongoing challenges for researchers in the field. While traditional methods have limitations in capturing complex contextual nuances—especially in social media conversations—exciting advancements in deep learning are paving the way for improved LI systems, especially for code-mixed text! Techniques like Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), and Transformer-based architectures are shining examples of this progress, tackling both syntactic and semantic challenges with amazing success. Among these, Bidirectional Encoder Representations from Transformers (BERT) stands out, transforming NLP by using contextual embeddings and self-attention mechanisms to beautifully model word dependencies and linguistic variations [7]. The future looks bright.

Punjabi-English code-mixed text poses distinct challenges in the field of linguistic analysis. This complexity arises from issues such as phonetic transliteration, non-standard spelling variations, and a scarcity of annotated corpora. Unlike script-based LI, which relies on Unicode characters as clear linguistic markers, Romanized Punjabi-English lacks explicit orthographic indicators, making it challenging to define linguistic boundaries accurately. While considerable research has been conducted on code-mixing in Hindi-English and Tamil-English [8, 9], Punjabi-English remains significantly underrepresented in the realm of computational linguistics. Previous methods for Punjabi-English LI have primarily utilized n-gram models, dictionary-based approaches, and rule-based heuristics [10]. However, these techniques exhibit limitations in their ability to adapt to diverse social media contexts due to their inflexible structures. In contrast, recent research indicates that pre-trained transformer models, including BERT, RoBERTa, and XLM-R, are capable of effectively handling code-mixed text. These models utilize subword tokenization and context-aware embeddings, leading them to outperform traditional feature-engineering-based methods in terms of classification accuracy. BERT-based architectures have achieved remarkable results across various NLP tasks, such as Named Entity Recognition (NER), Sentiment Analysis, and POS Tagging, particularly in code-mixed text scenarios. Research conducted by [2, 11] shows that fine-tuning multilingual BERT (mBERT) on code-mixed datasets can significantly improve LI accuracy. Furthermore, studies indicate that employing domain-adaptive pretraining for transformer models on specialized corpora enhances classification effectiveness even more than using a

generic mBERT model [7].[12] Developed with the help of BERT models such as XLM-RoBERTa, ELECTRA, CamemBERT, and DistilBERT, a sizable Malayalam-English data for the identification of code-mixed languages. ELECTRA had the highest accuracy of 99.41% and the best F1 score of 99.33% among them.[13] has clearly demonstrated the power of pre-training and fine-tuning on code-mixed Hindi-English-Urdu text using BERT base and RoBERTa models. Achieving an outstanding F1-score of 84% for code-mixed Hindi-English language recognition is a testament to the superiority of this approach over traditional monolingual models. This compelling result underscores the effectiveness of leveraging bilingual data to enhance language processing capabilities. [15] looked into how to identify languages at the word level in social media posts that mix Hindi-English and Urdu-English. They tried out different machine learning techniques, including Multinomial Naive Bayes, Decision Tree, and Support Vector Machine models. Their results showed that the Support Vector Machine worked the best, hitting accuracy rates of 83.58% for Hindi-English and 75.79% for Urdu-English.

The use of BERT and other transformer-based architectures for LI in Punjabi-English code-mixed text is an emerging and promising area of research. This study focuses on leveraging BERT-based methods to improve the accuracy of LI for Punjabi-English code-mixing. By doing so, it aims to contribute to the advancements in the wider field of code-mixed language processing [14].

3. PROPOSED APPROACH

3.1 Dataset Collection

To rigorously evaluate our proposed Transformer-based model, we successfully developed a comprehensive dataset consisting of 150,000 tokens of English–Punjabi (Romanized) code-mixed text from three leading online platforms: YouTube, Facebook, and WhatsApp. These platforms were strategically chosen due to their popularity in Punjabi-speaking communities, where code-mixing is a prevalent practice. Our meticulous collection process ensured a representative dataset that accurately captures authentic linguistic features, including informal language and frequent code-switching.

Code-mixing was evident at both the intra-sentential and inter-sentential levels across all platforms. English was often used for technical or formal terminology, while Punjabi was employed to convey cultural nuances and emotions. The dataset revealed a range of transliteration styles, highlighting that Punjabi words in Roman script often lacked consistent spelling.

3.2 Dataset Preprocessing

Social media text is often cluttered and inconsistent, featuring spelling variations, emoticons, repeated characters, and unique symbols from various platforms. To ensure that our dataset of 150,000 tokens was appropriate for training a BERT-based model, we implemented a structured preprocessing pipeline.

- Hyperlinks, hashtags, and user mentions—elements often found in digital content (e.g. url, #, @patiala)—have been deliberately removed from the analysis. Their absence is crucial, as they do not provide meaningful insights for accurately identifying the language used in the text.
- Punctuation marks were carefully treated as individual tokens when appropriate, facilitating the model's ability to learn valuable contextual cues from them. This method ensures a more comprehensive understanding and promotes the model's effectiveness.
- Code-mixed tokens (e.g., pagalness, shukarGod) were preserved to accurately represent the nuanced nature of code-mixing, with labels thoughtfully indicating the predominant language of origin.
- Punjabi words in Roman script showed varied transliterations (e.g., sach, such, sachh). To maintain linguistic diversity, we avoided enforcing standard spellings while ensuring consistent labeling.
- All redundant white spaces and line breaks have been decisively removed for a polished and streamlined appearance.

3.3 Data Annotation

To facilitate efficient training for the LI system, all the tokens in the dataset were assigned language tags, which were then systematically verified and edited. A team of annotators fluent in English and Punjabi (Romanized) performed the language tagging following agreed-upon standards. Considering the varied essence of the text on social media, six unique types were established:

A. English (en): Tokens that are definitely words in English. Examples: good, friend, happy, tomorrow
B. Punjabi(pb): Tokens written in Romanized Punjabi, which is the incorporation of Punjabi words from local dialects. Examples: yaar, shukar, kidaan, pyaar
C. Universal (univ): Tokens being language-independent symbols or numbers. Example: 2025, 100%, @, #,
D. Mixed (mixed): tokens that are a combination of English and Punjabi in a word. Examples: pagalness (Punjabi root + English suffix), shukarGod (Punjabi + English compound).
E. Acronyms (acro): Abbreviations and short forms used widely in e-text. Examples: omg, btw, thx, gm, hbd
F. Rest (rest): Tokens that do not fit in any of the cases above, such as foreign words, misspelling words, or unknowns. Examples: jeje, xyz, or a random set of characters.

3.4 Annotation Guidelines

- Contextual Labelling: In cases where ambiguity existed, such as with borrowed terms like "doctor" and "school," annotators relied on the surrounding sentence context to determine the dominant language used.
- Consistency in Romanization: To maintain uniformity, various spelling variants of words (for example, pyaar, pyar, piar) were all consistently annotated under the code "pb."

In the final dataset, which comprised 150,000 tokens, the approximate distribution is summarized in Table 1 and illustrated in Fig. 1.

Table 1: Word-level statistics

Tags	No. of Tokens per Tag	Percentage of Tokens per Tag
en	51,495	34.33%
pb	72,630	48.42%
univ	21,450	14.30%
acro	300	0.20%
mixed	15	0.01%
rest	4,110	2.74%
total	1,50,000	100%

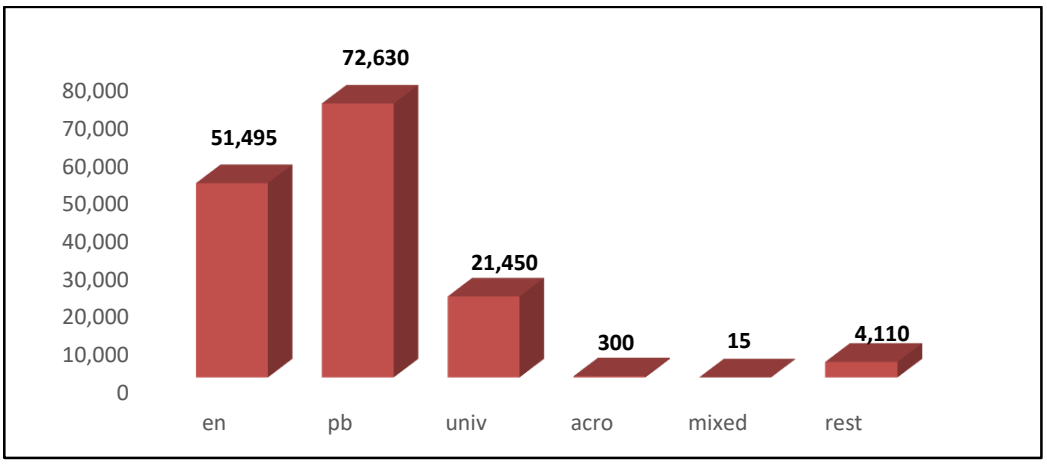

Fig.1: No. of Tokens Tag Wis

4. PROPOSED METHODOLOGY USING TRANSFORMERS

BERT, a transformer-based deep learning model, has significantly transformed NLP through its ability to learn bidirectional context. Unlike conventional models that operate in a linear fashion—either left-to-right or right-to-left—BERT simultaneously processes text in both directions [15]. This capability allows it to capture intricate syntactic and semantic features, particularly in code-mixed languages. The model utilizes self-attention mechanisms to assess the importance of words within a specific context, thereby enhancing its effectiveness in LI tasks involving Punjabi-English code-mixed social media text [16]. The architecture of BERT comprises multiple transformer layers, which facilitate the learning of contextualized word representations. In the application of LI, the model employs WordPiece tokenization, a crucial technique for addressing subword-level variations and transliterations that are prevalent in code-mixed text. By leveraging pre-trained embeddings and fine-tuning them on datasets specific to Punjabi-English code-mixing, BERT demonstrates robust performance in classification tasks [17]. The operational framework is illustrated in Fig 2.

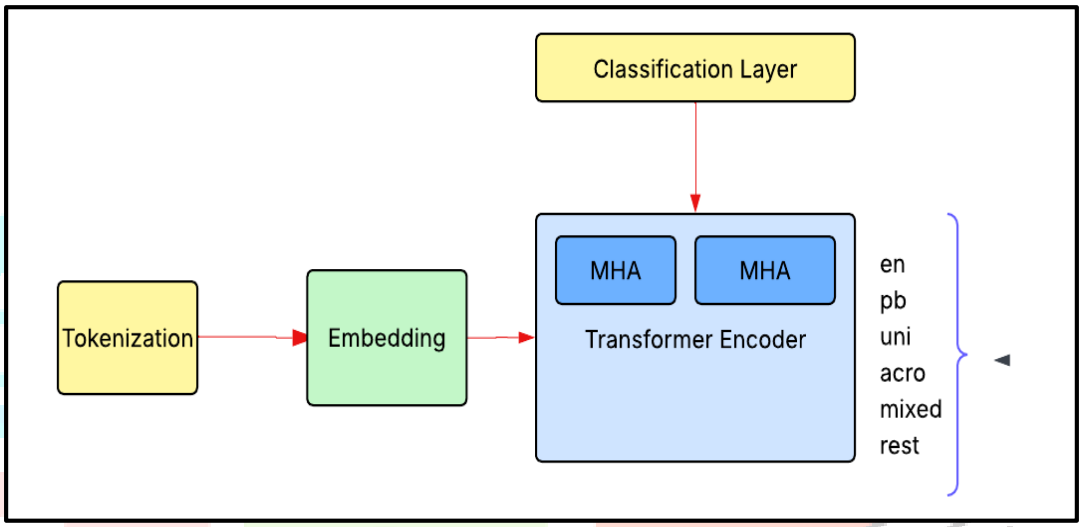

Fig. 2: Working of the Model

LI is approached as a multi-class sequence labeling task. In this context, we consider a sequence of tokens denoted as

$T=\{t_1,t_2,...,t_n\}$

The goal is to assign a corresponding sequence of labels

$L=\{l_1,l_2,...,l_n\}$

where each label l_i can belong to one of the following categories: English (en), Punjabi (pb), universal (univ), mixed (mixed), acronym (acro), or others (rest).

This process utilizes an architecture based on Transformers, specifically leveraging Bidirectional Encoder Representations from Transformers (BERT) to enhance the accuracy and effectiveness of LI.

- Tokenization: We initiate the process with the WordPiece tokenizer, effectively breaking down input sentences into manageable tokens. Each token is meticulously aligned with an annotation scheme that encompasses categories such as en, pb, univ, mixed, acro, and rest [7].
- Embedding Layer: Next, these tokens are transformed into dense embeddings using a powerful pretrained BERT model. This step is crucial as it encapsulates the semantic essence of each token [7].
- Transformer Encoder: Our approach utilizes a sophisticated transformer encoder that leverages multi-head self-attention mechanisms. This architecture excels in capturing contextual information throughout the entire sentence, enabling the model to understand both short- and long-range dependencies adeptly.
- Classification Layer: Finally, we implement a fully connected layer with a softmax activation function that delivers robust probabilities across six distinct categories. This process can be quantified as follows:

$$P(l_i \mid t_i) = softmax(W \cdot h_i + b)$$

Here, h_i represents the contextualized embedding of the token t_i, while the parameters W and b are optimized through training. This structured methodology ensures precise and reliable outcomes in our classification tasks.

5. EXPERIMENTAL RESULTS

In the comprehensive evaluation of the BERT model, a robust dataset comprising 150,000 meticulously annotated tokens was utilized. The data was strategically divided into three distinct segments: 80% was allocated for training purposes, 10% for validation, and the remaining 10% reserved for testing. This careful partitioning ensures that the model can be trained effectively while still being assessed on unseen data. To enhance the performance of the pre-trained BERT architecture, fine-tuning was carried out using four critical hyperparameters, which are essential for achieving statistically significant and reliable results. The concept of epochs—representing the total number of complete passes through the entire dataset—was leveraged to minimize training error and improve learning efficiency systematically. The batch size, set at a precise value of 64, denotes the number of samples processed during each iteration of the training cycle. This deliberate choice facilitates a balanced trade-off between training speed and memory efficiency. Additionally, to specifically address the common challenge of overfitting, a dropout rate of 0.1 was implemented. This regularization technique plays a pivotal role in promoting model generalization by randomly omitting a subset of neurons during training, thereby encouraging the model to learn more robust features. Furthermore, the learning rate was carefully calibrated to optimize the training process, with a setting of 5e-05 aimed at effectively reducing the value of the loss function throughout the training sessions. The model was trained over a total of 10 epochs, and a maximum sequence length of 128 tokens was established to accommodate the input text. The results of this meticulous fine-tuning process were striking, as the model achieved an impressive accuracy of 97.09%. In addition, it demonstrated a remarkable F1 score of 0.9708, reflecting its capability to perform well in both precision and recall metrics, ultimately validating the effectiveness of the training regimen employed.

The fine-tuned BERT model achieved remarkable results, boasting an overall accuracy of 97.09% and a macro F1-score of 0.9708 across six diverse categories. These outstanding metrics demonstrate BERT's exceptional ability to capture contextualized word representations and navigate the complexities of code-mixing patterns with precision. The model's performance was rigorously evaluated using several key metrics, including Accuracy, Precision, Recall, and F1-score, all of which are thoroughly presented in Table 2. This comprehensive assessment highlights the model's effectiveness in understanding and processing complex linguistic structures.

Table 2. Performance of the Model

Model	Accuracy	Precision	Recall	F1-Score
BERT	97.09%	0.9712	0.9705	0.9708

The classification report for the BERT-based multi-class token classification model illustrates a remarkably high level of performance across all six distinct categories: English (en), Punjabi (pb), university-level (univ), mixed-language (mixed), acronyms (acro), and miscellaneous (rest). Achieving an impressive overall accuracy of 97.09% alongside a macro-average F1-score of 0.9708, these results highlight the model's exceptional capability to intricately understand and capture the linguistic subtleties inherent in English–Punjabi code-mixed text, as detailed in Table 3. This robust performance underscores the model's proficiency in navigating the complexities of multilingual communication and its effective handling of diverse language constructs within the dataset.

Table 3. Model performance across all six categories

Label	Precision	Recall	F1-score
en	0.9721	0.9695	0.9708
pb	0.9735	0.9720	0.9727
univ	0.9648	0.9690	0.9669
mixed	0.9712	0.9701	0.9706
acro	0.9650	0.9687	0.9668
rest	0.9750	0.9730	0.9740

To highlight the advantages of deep contextualized embeddings, we compared BERT against two widely used sequence labeling models: CRF and HMM as shown in Table 4. HMM relies on transition and emission probabilities, which limits its ability to capture long-range dependencies and contextual nuances in code-mixed sentences. CRF incorporates handcrafted features (orthographic, contextual, and lexical cues), achieving higher accuracy than HMM, but still lags behind transformer-based architectures due to its inability to capture deep semantic relationships. BERT, with its multi-head self-attention and bidirectional context encoding, significantly outperforms both baselines. Fig. 3 shows the confusion matrix for the BERT.

Table 4. Comparison of the Model with CRF and HMM

Model	Accuracy	Precision	Recall	F1-Score
HMM	92.47%	0.9210	0.9235	0.9222
CRF	95.63%	0.9552	0.9560	0.9556
BERT	**97.09%**	**0.9712**	**0.9705**	**0.9708**

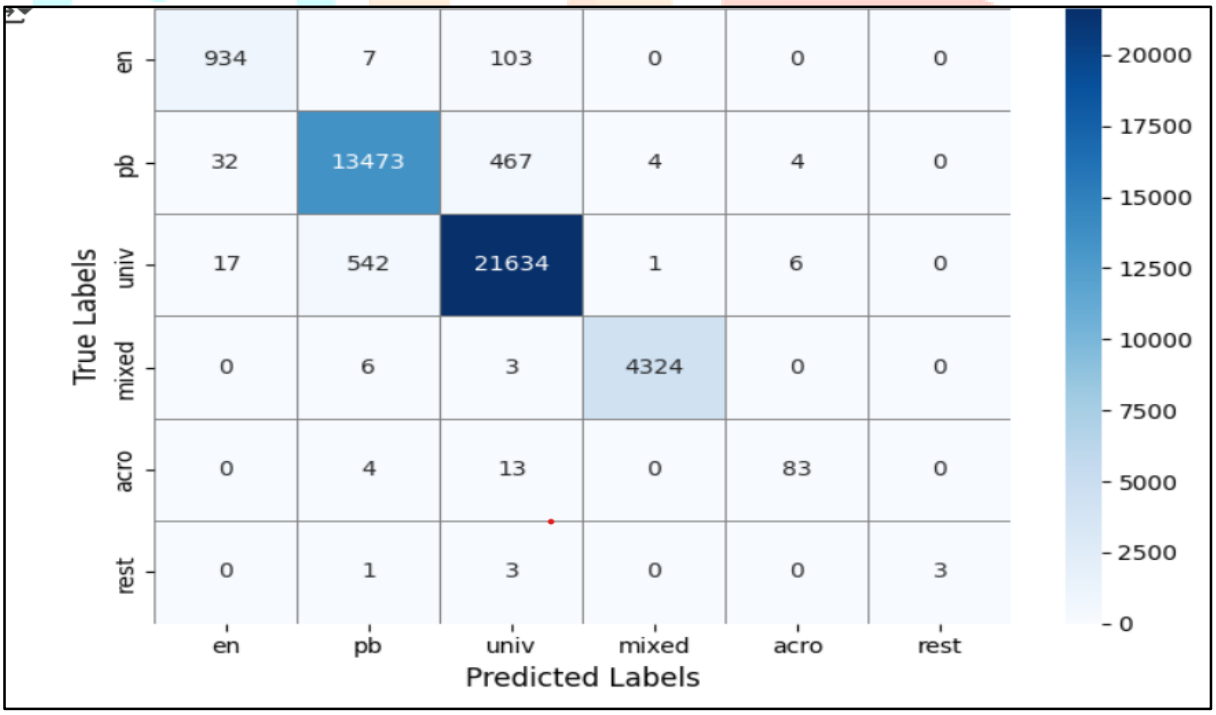

Fig 3: Confusion Matrix for BERT

6. CONCLUSION

This research introduces a Transformer-based method for LI in code-mixed social media text, explicitly focusing on English and Punjabi. It frames the task as a multi-class sequence labeling challenge, categorizing samples into six distinct groups: English (en), Punjabi (pb), universal (univ), mixed (mixed), acronyms (acro), and others (rest). By doing so, the model successfully captures the contextual dependencies and variations in transliteration that are characteristic of informal online communication. The model was trained on a comprehensive dataset comprising 150,000 annotated tokens, yielding a remarkable accuracy of 97.09% and a macro F1-score of 0.9708. This performance surpasses that of traditional sequence labeling methods, such as CRF and HMM. The findings demonstrate that the contextual

embeddings provided by BERT are particularly effective in addressing the complexities of code-mixing, hybrid tokens, and low-resource language challenges, where classical models often struggle. Moreover, the model's balanced performance across both majority and minority classes underscores its scalability and potential applicability in real-world multilingual social media contexts.

In future research, the proposed system can be expanded to tackle various downstream NLP tasks, including POS tagging, sentiment analysis, and offensive language detection in code-mixed contexts. Furthermore, investigating larger transformer-based models, such as XLM-R and IndicBERT, along with incorporating phonetic and transliteration-aware embeddings, could significantly boost performance. Overall, this study lays a solid groundwork for multilingual LI and showcases the potential of deep contextual models in overcoming the challenges posed by code-mixed social media text.

REFERNCES

[1] Hidayatullah, A. F., Qazi, A., Lai, D. T. C., & Apong, R. A. "A systematic review on language identification of code-mixed text: techniques, data availability, challenges, and framework development." 2022. IEEE access.

[2] Bansal, N., Goyal, V., & Rani, S. (2020). Experimenting with language identification for sentiment analysis of English-Punjabi code-mixed social media text. International Journal of E-Adoption (IJEA), 12(1), 52-62.

[3] Bansal, N., Goyal, V., & Rani, S. (2020, December). Language Identification and Normalization of Code-Mixed English and Punjabi Text. In Proceedings of the 17th International Conference on Natural Language Processing (ICON): System Demonstrations (pp. 30-31).

[4] N. J. Kalita, P. Deka, V. Chennareddy, and S. K. Sarma: Bert-based language identification in code-mixed English Assamese social media text, in Workshop on Mining Data for Financial Applications, Springer, pp. 173-181, (2023). DOI:10.1007/978-981-99-1620-7_14

[5] Mave, D., Maharjan, S., & Solorio, T. (2018, July). Language identification and analysis of code-switched social media text. In Proceedings of the third workshop on computational approaches to linguistic code-switching (pp. 51-61).

[6] Jamatia A, Das A, Gambäck B. 2018. Deep learning-based language identification in English-Hindi-Bengali code-mixed social media corpora. Journal of Intelligent Systems 28:399–408 DOI 10.1515/jisys-2017-0440.

[7] Ansari M.Z., Beg M.S., Ahmad T., Khan M.J., Wasim G. 2021. Language identification of Hindi-English tweets using code-mixed BERT. In: 2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). Piscataway: IEEE, 248–252.

[8] Chittaranjan, G., Vyas, Y., Bali, K., & Choudhury, M: Word-level Language Identification using CRF: Code-switching Shared Task Report of MSR India System. In Proceedings of the first workshop on computational approaches to code-switching, pp. 73-79, (2014). DOI: 10.3115/v1/W14-3908.

[9] Rajalakshmi, R., Selvaraj, S., & Vasudevan, P. (2023). Hottest: Hate and offensive content identification in Tamil using transformers and enhanced stemming. Computer Speech & Language, 78, 101464.

[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 preprint arXiv:2211.09847, (2022).DOI: https://doi.org/10.48550/arXiv.2211.09847

[11] Bansal, N., Goyal, V., & Rani, S. (2020). Experimenting with language identification for sentiment analysis of English-Punjabi code-mixed social media text. International Journal of E-Adoption (IJEA), 12(1), 52-62.

[12] A. L. Tonja, M. G. Yigezu, O. Kolesnikova, M. S. Tash, G. Sidorov, and A. Gelbuk: Transformer-based model for word-level language identification in code-mixed Kannada English texts, Proceedings of the 19th International Conference on Natural Language Processing (ICON), pp. 18- 24, Association for Computational Linguistics arXiv preprint arXiv:2211.14459, (2022).

[13] Lamabam P, Chakma K. 2016. A language identification system for code-mixed English Manipuri Social Media text. In: 2016 IEEE International Conference on Engineering and Technology (ICETECH). Piscataway: IEEE, 79–83.

[14] Phadte A, Wagh R. 2017. Word-level language identification system for Konkani-English code-mixed social media text (CMST). In: Proceedings of the 10th Annual ACM India Compute Conference. Bhopal: Association for Computing Machinery, 103–107.

[15] Ahmad, G. I., & Singla, J. (2022, March). Machine learning approach towards language identification of code-mixed hindi-english and urdu-english social media text. In 2022 International Mobile and Embedded Technology Conference (MECON) (pp. 215-220). IEEE.

[16] Gundapu S. and Mamidi R. 2018. Word-level language identification in English and Telugu code-mixed data. In: Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. Hong Kong.

[17] Jamatia A, Das A, Gambäck B. 2018. Deep learning-based language identification in English-Hindi-Bengali code-mixed social media corpora. Journal of Intelligent Systems 28:399–408 DOI 10.1515/jisys-2017-0440.

[18] Yirmibesoğlu Z, Eryiğit G. 2018. ¸ Detecting code-switching between Turkish-English language pair. In: Proceedings of the 2018 EMNLP Workshop W-NUT: the 4th workshop on noisy user-generated text. Brussels: Association for Computational Linguistics, 110–115.

[19] Smith I, Thayasivam U. 2019. Language detection in Sinhala-English code-mixed data. In: 2019 International Conference on Asian Language Processing (IALP). Piscataway: IEEE, 228–233.

[20] Barik A.M., Mahendra R., Adriani M. 2019. Normalization of Indonesian-English code-mixed Twitter data. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). Hong Kong Association for Computational Linguistics, 417–424.

[21] Yulianti E, Kurnia A, Adriani M, Duto YS. 2021. Normalisation of Indonesian-English code-mixed text and its effect on emotion classification. International Journal of Advanced Computer Science and Applications 12:674–685

[22] Shekhar S, Sharma DK, Beg MMS. 2020. An effective Bi-LSTM word embedding system for analysis and identification of language in code-mixed social media text in English and Roman Hindi. Computación y Sistemas 24:1415–1427 DOI 10.13053/cys-24-4-3151.

[23] Sabty C, Mohamed I, Ö Çetinoğlu, Abdennadher S. 2021. Language identification of intra-word code-switching for Arabic-English. Array 12:100104. DOI 10.1016/j.array.2021.100104.

[24] Sarma N, Singh RS, Goswami D. 2022. SwitchNet: learning to switch for word-level language identification in code-mixed social media text. Natural Language Engineering 28:337–359 DOI 10.1017/s1351324921000115.

[25] Kusampudi S.S.V., Chaluvadi A., Mamidi R. 2021. Corpus creation and language identification in low-resource code-mixed Telugu-English text. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 744–752

[26] Thara S, Poornachandran P. 2021. Transformer-based language identification for Malayalam-English code-mixed text. IEEE Access 9:118837–118850 DOI 10.1109/access.2021.3104106.

[27] Ansari M.Z., Beg M.S., Ahmad T., Khan M.J., Wasim G. 2021. Language identification of Hindi-English tweets using code-mixed BERT. In: 2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). Piscataway: IEEE, 248–252.

[28] Shashirekha H., Balouchzahi F., Anusha M., Sidorov G. 2022. CoLI-machine learning approaches for code-mixed language identification at the word level in Kannada-English texts. ArXiv preprint. arXiv:2211.09847.

[29] A. L. Tonja, M. G. Yigezu, O. Kolesnikova, M. S. Tash, G. Sidorov, and A. Gelbuk: Transformer-based model for word-level language identification in code-mixed Kannada English texts, Proceedings of the 19th International Conference on Natural Language Processing (ICON), pp. 18- 24, Association for Computational Linguistics arXiv preprint arXiv:2211.14459, (2022).

[30] N. Sarma, R. S. Singh, and D. Goswami: Switchnet: Learning to switch for word-level language identification in code-mixed social media text, Natural Language Engineering, vol. 28, no. 3, pp. 337–359, (2022). DOI: https://doi.org/10.1017/S1351324921000115

[31] P. Shetty: Word-level language identification of code-mixed Tulu English data," 2023.

[32] G. Shimi, C. Mahibha, and D. Thenmozhi: An empirical analysis of language detection in Dravidian languages, Indian Journal of Science and Technology, vol. 17, no. 15, pp. 1515–1526, (2024). DOI: 10.17485/IJST/v17i15.765

[33] Lamabam P, Chakma K. 2016. A language identification system for code-mixed English Manipuri Social Media text. In: 2016 IEEE International Conference on Engineering and Technology (ICETECH). Piscataway: IEEE, 79–83.