Transformer Driven Word Level Classification of Dravidian Languages

C.Jerin Mahibha[1,*], Wordson Robert[2], Gersome Shimi[3] and Durairaj Thenmozhi[4]

[1]Meenakshi Sundararajan Engineering College, Chennai
[2]Indian Institute of Science Education and Research, Kolkotta, India
[3]Madras Christian College,Chennai, India
[4]Sri Sivasubramaniya Nadar College of Engineering, Chennai

Abstract

Language detection is the process of automatically identifying the language used in a text, even when that text is not always coherent or grammatically correct. Language detection has become an essential tool in today's digital landscape focused on consumers, where businesses rely on user-generated content to tailor advertisements, products, and services more effectively. The challenge becomes much tougher when dealing with code-mixed or multilingual text, which is common in linguistically diverse regions like South India. These texts often include multiple languages, sometimes written in non-native scripts, and frequently involve code-switching at various levels, making it difficult for models trained only on monolingual data to perform well. To identify the language associated with a word, a high-performance model has been proposed for the CoLI-Dravidian@FIRE 2025 shared task for word-level identification, focused on identifying Dravidian languages such as Tamil, Telugu, Malayalam, Kannada, and Tulu. The proposed system uses a language-agnostic model to identify languages associated with words using the data sets provided by the organizers of CoLI-Dravidian@FIRE 2025. The results of the proposed system are encouraging with an Macro F1 score of 0.8995 for Kannada, 0.7434 for Tamil, 0.8271 for Malayalam, 0.9515 for Telugu and 0.8224 for Tulu. We were ranked 1 in the leaderboard for the languages Tamil, Malayalam, and Telugu and ranked 2 for the language Tulu. These results show the strength of the proposed model in word-level language identification of Dravidian languages.

1. Introduction

Language detection is a process by which we detect the language of an unknown text. Often when we deal with languages that have a similar origin or a source, then we can use various information we know about these languages to guess with a reasonable amount of error which language the text is from. The information used here is that languages, and more specifically languages of Dravidian origin, have very similar synonyms [1]. Most Dravidian languages, like Tamil, Tulu, and Malayalam use this extensively. The shared task by CoLI-Dravidian @ FIRE 2025 has organized a shared task on identifying word-level languages of the most prominent Dravidian languages.

Language detection is usually a tedious and rigorous process. It relies heavily on the vectorization of the most commonly used phrases, and the model has to be statistically trained on that data [2]. The datasets also need to be standardized to reduce any misunderstandings that the models, like BERT (Bidirectional Encoder Representations from Transformers), might have. Transformer models can make high-precision, reasoning-backed decisions from the data that is provided. It's truly a high-performing model that can understand nuance and context in the datasets [3]. It also uses its pretrained library to connect the dataset to the grammar patterns of languages it already understands, thus giving a much more precise prediction. It also uses smaller details like punctuation, word length, and word order to provide a much more informed and effective result.

Forum for Information Retrieval Evaluation, December 17-20, 2025, India.

*Wordson Robert

*Corresponding author.

†These authors contributed equally.

jerinmahibha@msec.edu.in (C.Jerin Mahibha); wordsonrobert@gmail.com (W. Robert); gshimi2022@gmail.com (G. Shimi); theni_d@ssn.edu.in (D. Thenmozhi)

For most languages, models like XLM-BERT are highly effective because they are user-efficient and come pretrained on over 100 languages. However, for Dravidian languages, models like RoBERTa and BERT often perform better. In the case of low-resource languages like Tamil and Tulu [4], when working with unlabeled data, semi-supervised learning techniques like pseudo-labeling are particularly effective [5]. For similar languages, generalization tends to work well, which boosts the performance of these models. BERT, in particular, excels in both generalization and data expansion [6]. Retraining the model on the pseudo-labeled dataset can lead to even better results for transformer-based models like BERT.

The paper is organized with Section 2 and 3 discussing on related works and datasets, Section 4 on system description, Section 5 discussing results, Section 6 and 7 contributing the error analysis and conclusion.

2. Related Works

A survey of Language Identification of Code-Mixed Text based on Techniques, Data Availability and Challenges had been done by [7]. The findings revealed that excellent performance had been shown by a multichannel CNN incorporated with BILSTM and CRF. Considering Non-neural network techniques, SVM and CRF are recommended to be applied. Transformer based technique can also be considered one of the most robust techniques for code-mixed Language Identification [8] due to its remarkable performance in equivalent tasks. BERT, a transformer model, along with its variants - CamemBERT, DistilBERT had been used to implement a word level language identification of Malayalam-English code-mixed data on a dataset collected from social media platforms [9]. A word-level language identification model for code-mixed Indonesian, Javanese, and English tweets has been implemented using various approaches, like fine-tuning BERT, BLSTM-based, and CRF [10]. BERT's ability to understand each word's context from the given text sequence is evident from the results obtained by the fine-tuned IndoBERTweet model [11]. A framework for language identification has been proposed that makes use of a dynamic switching mechanism for effective language classification of both words that are borrowed or embedded from other languages as well as words that are valid in multiple languages [12]. Identification of languages on Twitter has been implemented by exploiting a transfer learning approach and fine-tuning BERT models by [13]. It involves Hindi-English-Urdu codemixed text for language pre-training and Hindi-English codemixed text for subsequent word-level language classification [14]. It is evident from the results that the representations pre-trained over codemixed data produce better results than their monolingual counterpart. The use of a Transformer based model for word-level language identification in code-mixed Kannada English texts has been proposed by [15]. An empirical analysis of Dravidian language identification in social media text using machine learning and deep learning approaches with k-fold cross validation has been implemented [16]. The empirical analysis focused on various Machine Learning and Deep Learning models based on performance measures like accuracy, precision, recall, and F1-score. It was found that the language agnostic model outperformed all other models considering the task of language detection in Dravidian languages. Language Identification from code-mixed text with English and one of the three South Dravidian languages: Kannada, Malayalam, and Tamil was a part of Dravidian Language Identification (DLI) shared task organized at VarDial 2021 workshop. [17] had used a Naive Bayes based classifier with adaptive language models to obtain a competitive performance in the shared task.

3. Data set

The dataset used by the proposed system was provided by the shared task organizers and was available for five languages, namely, Tamil, Tulu, Telugu, Kannada, and Malayalam. Separate datasets were provided for training, validation, and testing the model. The proposed task was to classify the language associated with the words from the test dataset.

The distribution of data in all the datasets is tabulated in the Table 1. The number of instances in the training dataset was 30910, 13514, 25995, 6280, and 29524 for Kannada, Tamil, Malayalam, Telugu, and

Table 1
Data Distribution

Language	Training Dataset	Evaluation Dataset	Test Dataset
Kannada	30910	2016	2075
Tamil	13514	1984	2006
Malayalam	25995	2008	1997
Telugu	6280	515	494
Tulu	29524	3006	3283

Tulu, respectively. The evaluation dataset had 2016 instances for Kannada, 1984 instances for Tamil, 2008 instances for Malayalam, 515 instances for Telugu, and 3006 instances for Tulu. There were 2075, 2006, 1997, 494, and 3283 instances in the test dataset for the languages Kannada, Tamil, Malayalam, Telugu, and Tulu, respectively.

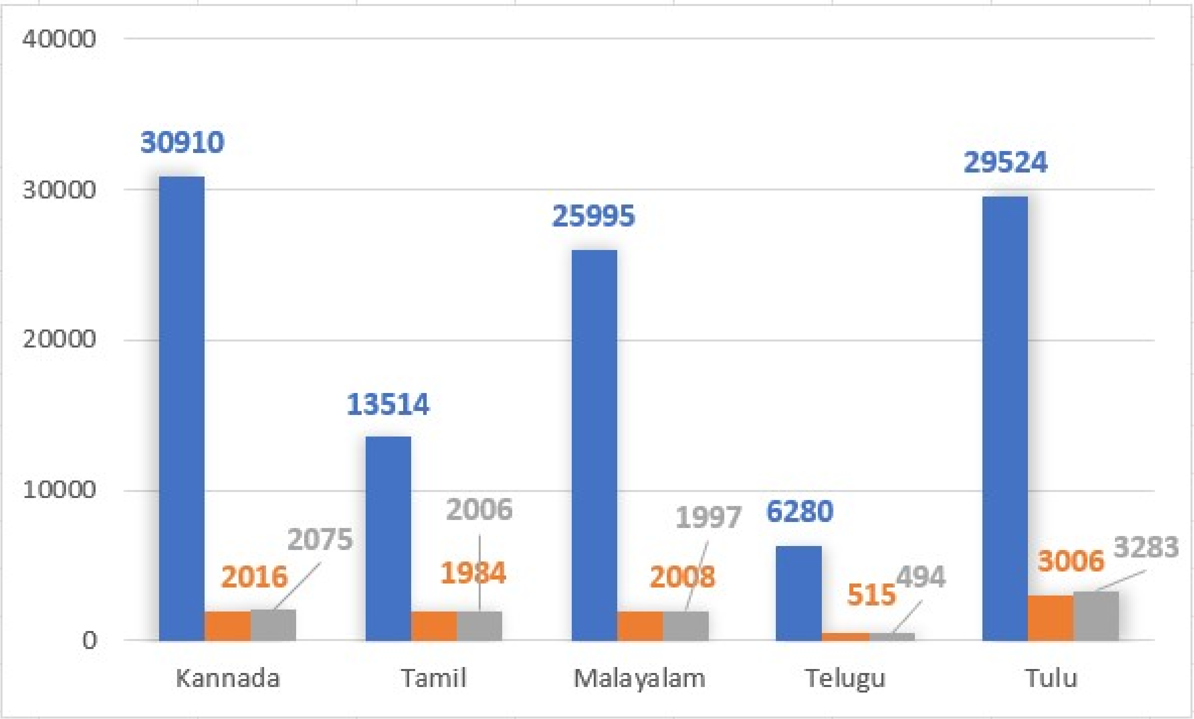

Figure 1: Data distribution

The Figure 1 gives a visual distribution of all the given datasets.

4. System Description

The architecture of the proposed system uses a language-agnostic model to identify the language associated with the words in the test dataset. The languages that are considered for identification include five Dravidian languages: Kannada, Tamil, Malayalam, Telugu, and Tulu. The figure 2 illustrates the components of the proposed system. The training dataset is used to train the model, optimizing it for maximum accuracy. The performance of the proposed model is evaluated using the evaluation dataset. The trained model is then used for predicting the language associated with instances of the test dataset, which uses different metrics, such as accuracy, precision, recall, and F1 scores, to assess its performance.

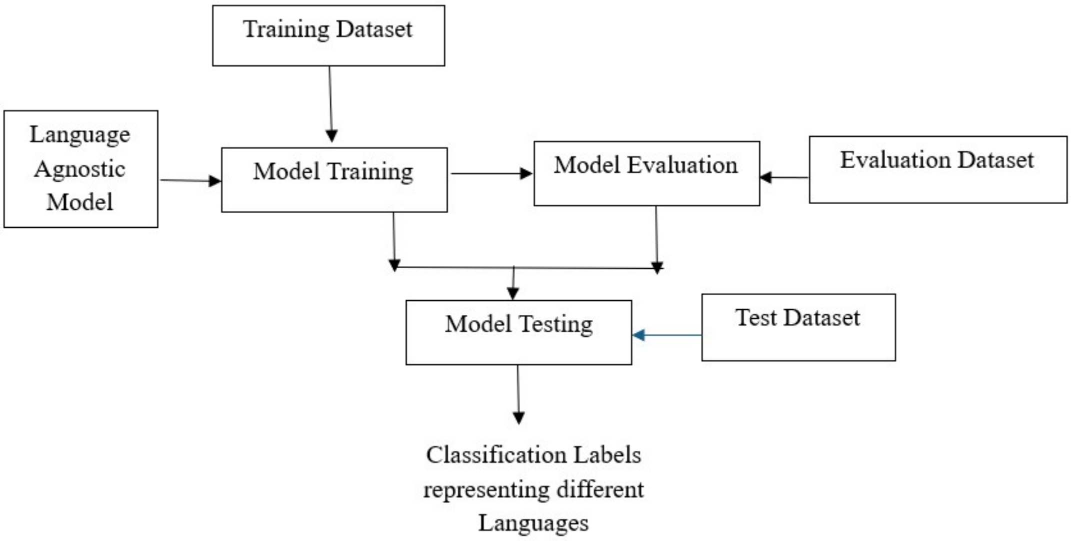

Figure 2: System Architecture

4.1. Methodology

To accomplish the language detection, from the available transformer models, a language-agnostic model was chosen. The model was trained for 10 epochs by setting the parameter representing the number of labels in the respective dataset.

Language agnostic BERT Sentence Embedding [18] is a multilingual model for cross-lingual sentence embedding in 109 languages. Pre-training can be accomplished by combining masked language modelling (MLM) with translation language modelling (TLM). This model performs well in multilingual sentence embedding and multilingual text retrieval. To make the process of training more efficient, a dual-encoder architecture has been used, which is considered to be an effective approach for learning cross-lingual embeddings. The BERT transformer model forms the base of the encoder architecture, which has 12 transformer blocks, 12 attention heads, and 768 per-position hidden units. All languages share the various encoder parameters. Tokenization in Language agnostic model plays a crucial role in preparing the text for input into the model, allowing it to process and understand the semantic meaning of sentences across multiple languages. The input text is tokenized into smaller units using the WordPiece tokenizer. This involves breaking down words into sub word units. Each token is assigned a unique token ID, which corresponds to its index in the tokenizer's vocabulary. Special tokens are also added to the tokenized input to mark the beginning and end of sentences, as well as to denote padding or unknown tokens. Along with token IDs, attention masks are also generated to indicate which tokens are actual words and which ones are padding tokens. This helps the model focus only on the relevant parts of the input during processing. From the last transformer block, normalized [CLS] token representations are extracted as the sentence embeddings. A shared transformer network is used to encode the source and target text, and the translation ranking task helps to get similar representations for the source and target text. Mapping similar words from different languages to a common representation is part of the parameter sharing capacity of the encoders by altering the hyperparameters associated with the model, it is being trained. The model is trained with the objective of feature prediction. The number of labels is set to 4 or 5 based on the number of labels provided in the dataset while tuning the model for language detection, which is trained for 10 epochs. The model has been implemented with Adam optimizer and a batch size of 32. The process behind this method is represented by Figure 3. Language embeddings represent entire languages as fixed-size vectors in an embedding space. These embeddings can capture various linguistic properties of languages, such as vocabulary, syntax, and semantics. In the context of language detection tasks, language embeddings contribute by capturing the unique linguistic

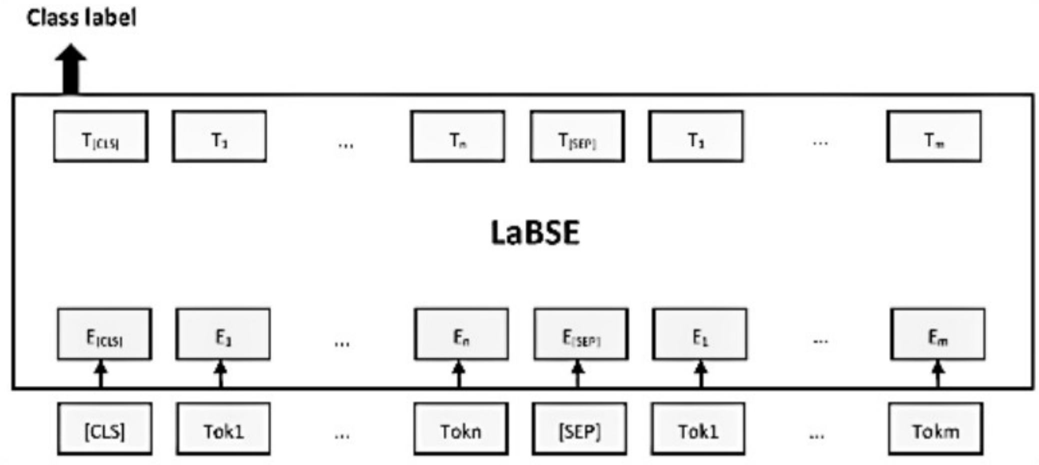

Figure 3: Language Agnostic BERT Model

Table 2
Performance score

Language	Weighted Precision	Weighted Recall	Weighted F1	Macro Precision	Macro Recall	Macro F1	Accuracy
Kannada	0.9686	0.9681	0.9683	0.8863	0.9147	0.8995	0.9681
Tamil	0.9249	0.9249	0.9242	0.7696	0.7353	0.7434	0.9249
Malayalam	0.8825	0.8843	0.8818	0.8624	0.8028	0.8271	0.8843
Telugu	0.9689	0.9676	0.9681	0.9446	0.9590	0.9515	0.9676
Tulu	0.9009	0.9028	0.9011	0.8425	0.8063	0.8224	0.9028

characteristics of different languages in a continuous vector space.

5. Results and Discussion

To assess the accuracy of the proposed model, the parameter macro-F1 score has been used by the organizers. Macro-F1 is a very balanced metric to use because it gives equal importance to how well we predict each of the labels, even if the sample sizes for them are unequal. It's the unweighted average of the F1 scores for each label in the dataset. And the F1 score itself is the harmonic mean of how many correct predictions we made for a class label and how many actual instances of that class labels were found correct. This works especially well for the proposed system because the dataset is clearly unbalanced across the five languages.

The performance scores of the proposed models are represented in the table 2. The proposed system achieved a macro F1 Score of 0.8995 for Kannada, 0.7434 for Tamil, 0.8271 for Malayalam, 0.9515 for Telugu and 0.8224 for Tulu. These scores show that our models were pretty accurate across the board. In the CoLI-Dravidian @ FIRE 2025 shared task, this helped us to be ranked 1st for Tamil, Malayalam, and Telugu, 2nd for Tulu and 7th for Kannada which shows that the proposed system for language detection predicted up really well compared to others.

6. Error Analysis

The Macro F1 score obtained for the task using the proposed language-agnostic model shows that more false positive and false negative classifications have occurred. One reason for this could be the data imbalance nature of the dataset. The confusion matrix of the proposed system considering the different languages is represented in figures 4, 5, 6, and 7. The consistent performance of the model suggests

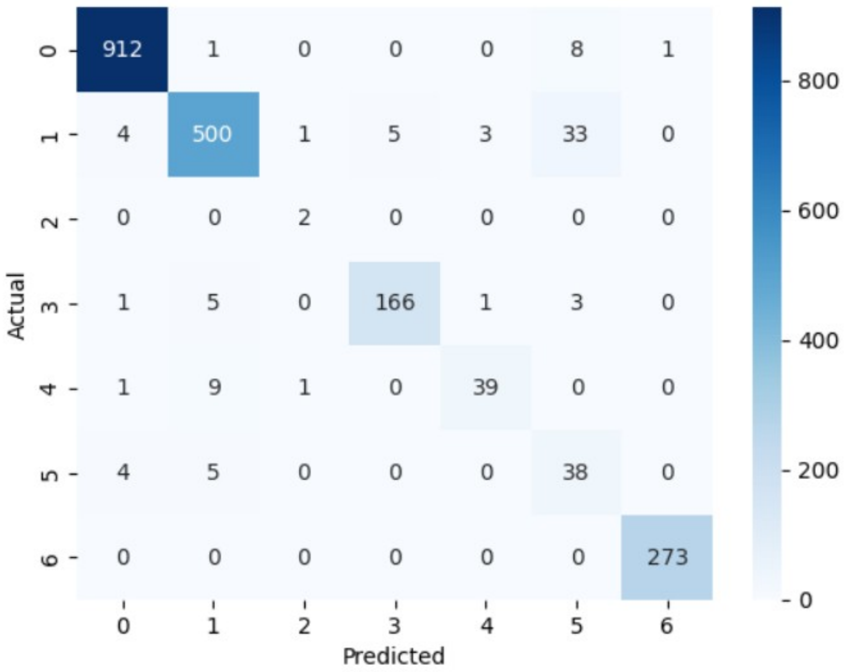

Figure 4: Classification Report - Kannada

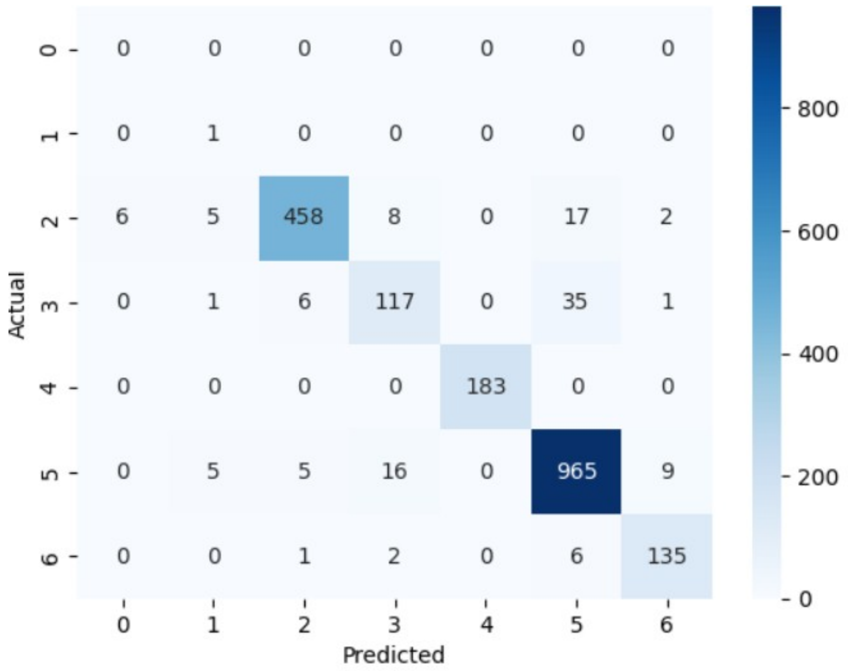

Figure 5: Classification Report - Tamil

that it generalizes beyond memorization, effectively classifying across languages. The proposed model utilizes a language-agnostic approach, which is effective for low-resource languages. This consistency across languages shows that the proposed model is equally effective for every class label, not just the ones with more data.

The occasional mismatches or deeply etymologically ambiguous words that led to errors are areas that can still be improved on. A few of the misclassifications are represented in Table 3.

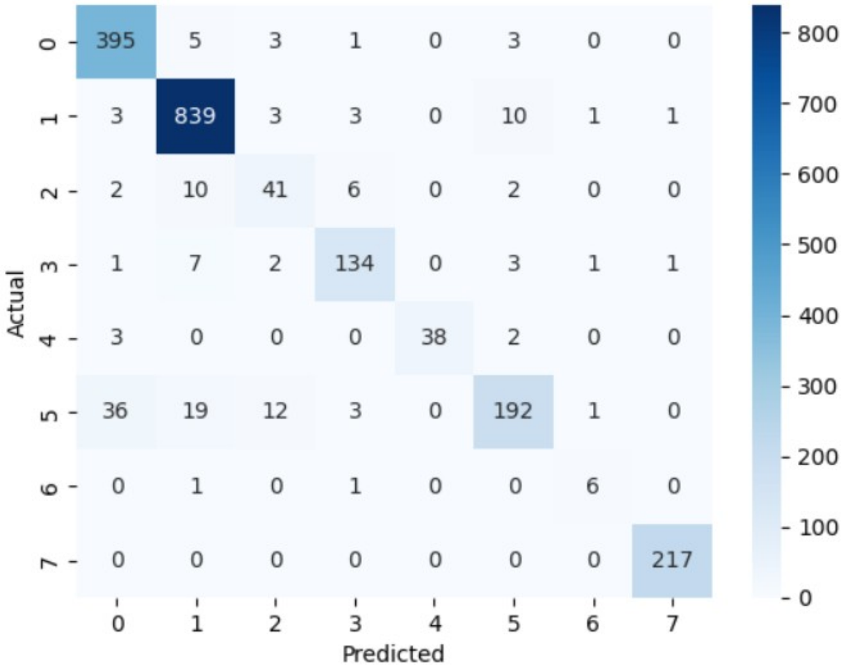

Figure 6: Classification Report - Malayalam

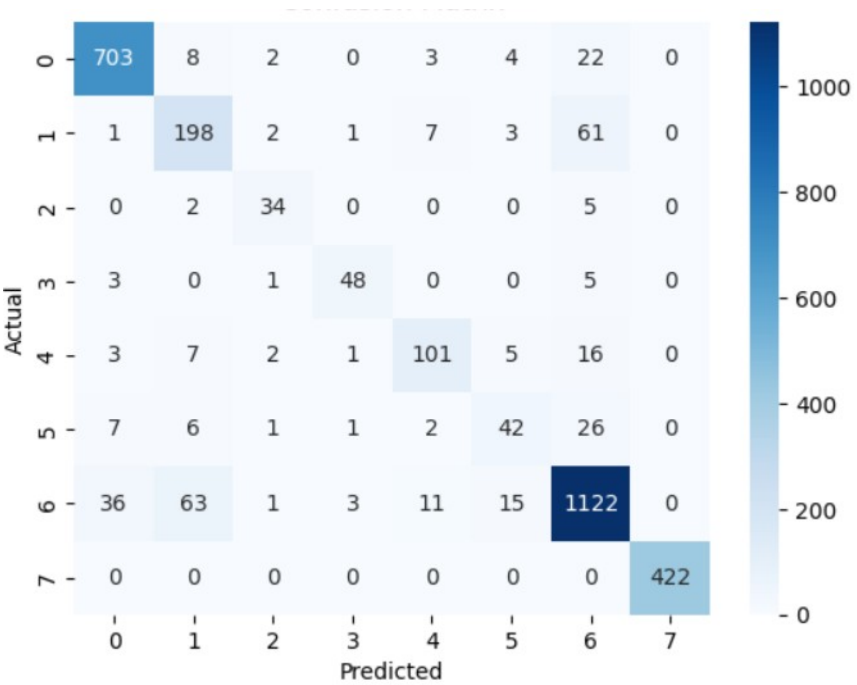

Figure 7: Classification Report - Tulu

7. Conclusion

Especially online, for digital content and personalization, language detection has become an increasingly important process. It enables platforms to deliver more customized content to users, moderate information better, regulate misinformation, and block hate speech. This process becomes even more crucial when dealing with similar, niche languages—like those of Dravidian origin—because it allows for better tagging, search engine results, sarcasm detection, and even sentiment analysis. Recognizing its criticality, FIRE 2025 introduced a shared task focused specifically on Dravidian language detec-

Table 3
Error Analysis

S.No.	Word	Actual Label	Predicted Label
01.	Torsidira	Kannada	Other
02.	Madugowda	Name	Kannada
03.	Sambavam	Tamil	Name
04.	Bunk	English	Tamil
05.	Chetta	Malayalam	Name
06.	Negatuve	English	Malayalam
07.	Randi	Telugu	Name
08.	Karo	Other	Telugu
09.	Satva	Tulu	Kannada
10.	Waasteee	English	Tulu

tion. For this task, given under the banner CoLI-Dravidian @ FIRE 2025 the proposed system uses a language-agnostic model to effectively classify languages. The achieved performance metrics are consistently high across languages, indicating strong and reliable performance. However, by adopting more customizable approaches—such as time-based anomaly detection and deeper model refinement—the boundaries can be pushed further and develop even more accurate and adaptable deep learning models.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] S. B. Steever, Introduction to the dravidian languages, in: The Dravidian languages, Routledge, 2019, pp. 1–44.

[2] R. Egger, Text representations and word embeddings: Vectorizing textual data, in: Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications, Springer, 2022, pp. 335–361.

[3] N. Sulaiman, F. Hamzah, Evaluation of transfer learning and adaptability in large language models with the glue benchmark, Authorea Preprints (2024).

[4] A. Hegde, F. Balouchzahi, S. Coelho, S. HL, H. A. Nayel, S. Butt, Coli@ fire2023: Findings of word-level language identification in code-mixed tulu text, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023, pp. 25–26.

[5] L. Ran, Y. Li, G. Liang, Y. Zhang, Pseudo labeling methods for semi-supervised semantic segmentation: A review and future perspectives, IEEE Transactions on Circuits and Systems for Video Technology (2024).

[6] M. V. Koroteev, Bert: a review of applications in natural language processing and understanding, arXiv preprint arXiv:2103.11943 (2021).

[7] A. F. Hidayatullah, A. Qazi, D. T. C. Lai, R. A. Apong, A systematic review on language identification of code-mixed text: techniques, data availability, challenges, and framework development, IEEE access 10 (2022) 122812–122831.

[8] F. Balouchzahi, S. Butt, A. Hegde, N. Ashraf, H. Shashirekha, G. Sidorov, A. Gelbukh, Overview of coli-kanglish: Word level language identification in code-mixed kannada-english texts at icon 2022, in: Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, 2022, pp. 38–45.

[9] S. Thara, P. Poornachandran, Transformer based language identification for malayalam-english code-mixed text, IEEE Access 9 (2021) 118837–118850.

[10] A. Hegde, F. Balouchzahi, S. Coelho, H. Shashirekha, H. A. Nayel, S. Butt, Overview of coli-tunglish: Word-level language identification in code-mixed tulu text at fire 2023., in: FIRE (Working Notes), 2023, pp. 179–190.

[11] A. F. Hidayatullah, R. A. Apong, D. T. Lai, A. Qazi, Corpus creation and language identification for code-mixed indonesian-javanese-english tweets, PeerJ Computer Science 9 (2023) e1312.

[12] N. Sarma, R. S. Singh, D. Goswami, Switchnet: Learning to switch for word-level language identification in code-mixed social media text, Natural Language Engineering 28 (2022) 337–359.

[13] M. Z. Ansari, M. S. Beg, T. Ahmad, M. J. Khan, G. Wasim, Language identification of hindi-english tweets using code-mixed bert, in: 2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), IEEE, 2021, pp. 248–252.

[14] H. L. Shashirekha, F. Balouchzahi, M. D. Anusha, 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).

[15] A. L. Tonja, M. G. Yigezu, O. Kolesnikova, M. S. Tash, G. Sidorov, A. Gelbuk, Transformer-based model for word level language identification in code-mixed kannada-english texts, arXiv preprint arXiv:2211.14459 (2022).

[16] G. Shimi, C. Mahibha, D. Thenmozhi, An empirical analysis of language detection in dravidian languages, Indian Journal of Science and Technology 17 (2024) 1515–1526.

[17] T. Jauhiainen, T. Ranasinghe, M. Zampieri, Comparing approaches to dravidian language identification, arXiv preprint arXiv:2103.05552 (2021).

[18] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2020).