International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 07 | Jul 2026 www.irjet.net p-ISSN: 2395-0072

Transformer Models for Kannada-English Code-Mixed Enterprise Communication Analytics

T L Manasa[1], Bharath Kumar N[2]

[1]*Research Scholar, Faculty of Engineering & Technology, Department of Electrical & Electronics Engineering, Mangalayatan University, Beswan, Aligarh, UP, India*
[2]*Senior Software Engineer, IBM*

Abstract - *For support chats, email, CRM ticket or internal communication platforms, Kannada-English code-mixing is more prevalent in communication in multi-lingual regions, like Karnataka. Currently, such hybrid texts that have partially native Sanskrit script and partly transliterated and/or English script present great challenges in the framework of natural language processing (NLP) pipelines that are designed to process mainly monolingual English texts. This paper introduces the framework for analytics over enterprise communication in Kannada-English code-mixed scenario such as sentiment classification, intent detection, urgency/priority scoring and enterprise specific named entity recognition (product names, Ticket identifiers and department references). The concept is to use script identification and normalisation techniques and word-level language tagging and transliteration processing before fine-tuning multilingual transformer-based models such as mBERT, XLM-RoBERTa and IndicBERT on a well-curated enterprise-specific corpus of Kanglish text that contains anonymized customer-support chats, help-desk emails, and inner-communication exports. In this paper, we present a customized version of the pretrained multilingual encoder, called Kanglish-BERT, that includes an extra language identification head and a language-augmented training process for the data by back-transliteration. Comparative experiments demonstrate the consistent advantage of the proposed model over other baselines like vanilla mBERT, XLM-RoBERTa and IndicBERT on the sentiment, intent and entity-recognition tasks not seen during enterprise testing. The paper also includes practical tips on deployment, including latency of inferences, compression of models for on-premise deployment and protection of data privacy in enterprise analytics dashboards. The results inform that an application of the adapted transformer model designed with the Kannada-English code-switching patterns is the requirement for a future automated communication analytics solution for regional enterprises that can enable enterprise to resolve the escalation faster, gain insights of customers' sentiments, and provide multilingual enterprise intelligence to address the overall resource scarcity in applied enterprise NLP for Dravidian code-mixed languages.*

Key Words: Code-Mixed NLP, Kannada-English, Kanglish, Transformer Models, Enterprise Communication Analytics, Sentiment Analysis, IndicBERT, XLM-RoBERTa, Dravidian Language Processing.

1. INTRODUCTION

There is a lot of multilingualism in India and Karnataka with Kannada as the main language is not an exception. In day-to-day business situations, the switching between Kannada and English takes place within an utterance/sentence fairly commonly by employees, customers and support agents, which is formally known as code-mixing and informally as "Kanglish". The usage of the Roman script to represent native Kannada script is quite common in enterprise communication wherein the users write at a much faster pace such as writing a helpdesk email thread, a native script chat with customers, writing a ticket comment in CRM or activities in an enterprise communication application.

The swift adoption of digital/omnichannel engagement, especially in e-commerce, banking, telecom and IT-enabled services based out of Karnataka, has resulted in large amounts of unstructured, code-mixed textual data. Companies are increasingly demanding to analyze and unlock this information for sentiment trends, escalation risk, service quality concerns and new customer concerns. But, the traditional sentiment-analysis and text-classification pipelines, trained and tested mainly with monolingual English or high resource language corpora, do not perform well on Kannada-English code-mix text. This performance gap is because of the irregular spelling of transliterated Kannada, the usage of different scripts, absence of large annotated enterprise domain data and the syntactic complexities of intra-sentential switching.

With the introduction and widespread use of pretrained language models such as BERT and BERT-multi language (mBERT), transformer-based models are a great leap forward in the field of NLP for high resource and low resource languages. To further improve the performance for the downstream tasks, large-scale pretraining of multilingual and India-specific models such as XLM-RoBERTa and IndicBERT using multilingual pretraining corpora including Dravidian languages has been explored. However, majority of the research on code-mixed NLP has been conducted on social media datasets, in the context of sentiment analysis, offensive language detection, and word level language identification, among others, mostly on YouTube comments and Twitter text. In contrast, enterprise communication is somewhat under-researched – partly because it is rather different from other forms of communication in which the

languages have not yet been fixed permanently in the formal-informal register dichotomy, and the content is typically task-oriented rather than other registers.

In this paper, we fill this gap by introducing a transformer-based analytics framework uniquely designed for enterprise communication in Kannada-English code-mixed environment. The major contributions of this work are; (i) A pre-processing pipeline for script detection, normalization and word-level language identification useful for enterprise level kanglish text; (ii) A fine-tuning strategy to combine the auxiliary Language Identification objective along with data augmentation using back translator to train a kanglish specific transformer encoder model, namely Kanglish-BERT; (iii) A multi-task analysis pipeline consisting of sentiment classification, intent detection, urgency scoring and enterprise specific NER; and (iv) Comparative evaluation of various transformer models (mBERT, XLM-RoBERTa, IndicBERT and the proposed kanglish variant of BERT model (Kanglish-BERT) on enterprise style code-mixed corpus and discuss about deployment considerations for enterprise analytics dashboard.

2. LITERATURE REVIEW

The study of Kannada-English code-mixed NLP has been steadily increasing over the last five years owing to the presence of shared-task programs, and the increase in social media data. Balouchzahi et al. proposed CoLI-Kanglish at ICON 2022, primarily focused on word-level IL in Kannada – English code-mixed text and gave a huge boost to the future task of code-mixed processing downstream [1]. Other work followed that, which introduced a model based on transformer that introduced BERT-like encoders along with sequence labelling heads for Kannada-English text that achieved much higher weighted F1-scores than the prior rule-based and classical machine-learning baselines [2].

The use of the transformer is also proving to be a useful step in sentiment analysis in the Dravidian language. Chakravarthi et al. designed and gathered multi-lingual, offensive language and sentiment datasets of several thousand YouTube comments for Tamil-English, Kannada-English and Malayalam-English with some marginally smaller datasets for Maa-En and Kaa-En.To work with sentiment analysis and offensive language identification, Chakravarthi et al. compiled and manually labelled a multi-lingual collection comprising several thousand Kannada-English, and much larger Tamil-English and Malayalam-English collections. This dataset has been used to test and analyze the performance of transformer-based classifiers, as well as classifiers based on deep learning and feature engineering, that usually outperform classical models that utilize only features engineered using traditional algorithms on code-mixed sentiment tasks [4]. Another closely related work [5] fine-tuned each model by using subword-level tokenisation and employed Hugging Face Transformers library, and achieved a higher accuracy than the baseline models for the classification task of sentiment and offensive content on Kanglish text from social media by using multilingual BERT model (mBERT).

Besides classification, there has been several recent efforts on the transliteration and translation of code-mixed text into pure Kannada text. Yet another paper used sentence-transformer embeddings with the pretrained DeBERTa and MPNet encoders to obtain accurate translations of CM Kannada sentences into Kannada, and noted that the current challenges of traditional BERT models on rare and irregular Romanised Kannada spellings [6]. To demonstrate that sequence to sequence transformer methods can partially standardize code-mixed sentences to pure Kannada before processing them for downstream tasks, a very similar approach was used that involved pre-processing code-mixed Kannada-English sentences to pure Kannada using an IndicBART (indigenous version of BART) encoder-decoder transformer. KaEnLandetector is a language tagger that leverages both rule-based and transformer-based language detectors based on linguistic heuristic and BERT-based classifiers for enhancing the performance of language tagger for Kannada-English text from real world sources [8].

Other code-mixed language pairs of Dravidian and Indian languages are provided for methodological purposes. Heinrich et al. [9] conducted a detailed comparative analysis of the different transformer models used for Tamil-English code-mixed sentiment analysis, such as XLM-RoBERTa, mT5, IndicBERT and RemBERT, and found there was still scope for improvement which can be achieved by targeted data augmentation, phonetic normalisation, and hybrid models in low-resource code-mixed environments. The CoMi-Lingua project has published a large-scale (3.1 billion parameters) expert annotated multitask database of code-mixed Hindi-English NLP and the L3Cube-HingBERT project has created language-pair specific pre-trained BERT models for code-mixed Hindi-English text [10, 11], highlighting the need of language-pair specific pre-training instead of code-switch specific pre-training. Other innovations, such as CONFLATOR model, which have been introduced for positional-encoding, have shown that explicit modelling of code-switch boundaries can improve results for code-mixed LM for downstream models [12].

Despite this push for investigation, there are three parts of the topic which are still outstanding. But most of the Kannada-English datasets and models are obtained from social media platforms such as YouTube and Twitter and very little research and development work has been invested into enterprise domain data such as support chat, email or CRM text. Second, most of the existing research is only concerned with sentiment or language classification as a single-task classification or sentiment prediction, whereas analysing sentiment, intent and urgency to classify an entity also requires modelling these three features together in one pipeline for the purposes of enterprise analytics. Third,

issues that are more relevant to deployment, like inference latency, model compression, and enterprise data-privacy compliance, are rarely talked about along with model accuracy. This paper is motivated by and expands on this work to close these gaps.

3. PROBLEM STATEMENT AND OBJECTIVES

Large amounts of Kannada-English code-mixed text are created in business within the state of Karnataka and other Kannada speaking areas, by their customer support chat platforms, email help desks, customer relationship management (CRM) ticketing systems and their internal messaging system. Typically, it is informal text, script-inconsistent text (Native Kannada script with Romanised Kannada and English labels in the same message), and domain-specific text (Product names, and ticket references as well as organisational text that is not present in general purpose pre-trained corpora). Most current sentiment analysis and text classification tools are either single language (English) or will classify all languages, but aren't optimised for this code-mixing and enterprise vocabulary and are not doing as well as they could when classifying text, notifying of escalations or automatically reporting.

Therefore, the objective is to design a preprocessing pipeline suitable to the variations in the script and transliteration inconsistencies of enterprise Kannada-English code-mixed text, fine-tuning and comparing some of the most popular multilingual transformer encoders namely mBERT, XLM-RoBERTa and IndicBERT and propose and evaluate an improved variant, Kanglish-BERT, which added an auxiliary language-identification head and back-transliteration based augmentation, and assess its viability in enterprise communication analytics dashboard for latency, resource limitations and data-privacy concerns common in enterprise communication settings.

4. PROPOSED METHODOLOGY

4.1 Data Collection and Corpus Construction

The proposed framework is based on an enterprise style Kannada-English code-mixed corpus which is generated from three categories of data: (i) anonymous customer-support transcripts of conversations collected from a simulated customer-support environment, resembling e-commerce and telecom related service and complaint handling processes, (ii) helpdesk email threads of typical customer support and complaint handling processes, and (iii) data collected from internal team-messaging used for internal Kannada-English team communications. All personally identifiable information (including names, phone numbers, account numbers etc.) was de-identified or anonymized prior to annotation, as is conventional with enterprise text corpora.

All messages in the resulting corpus have been hand coded at four dimensions: sentiment (either positive, negative or neutral), intent (query, complaint, request, feedback or escalation), urgency level (low, medium or high), and enterprise-related named entities (product/service names, ticket or reference numbers, department names or expressions of date or time). The bilingual annotators were also fluent in both Kannada and English and the annotations were done using guidelines inspired by the existing Dravidian code-mixed sentiment-annotation schemes interspersed with periodic IAA meetings for consistency reasons.

Care was taken to not exclude code-mixing from the conversation in the corpus as it is also a pragmatic signal, for instance, during a customer support conversation, which is primarily in English, but sometimes interrupted by code-switching when the customer becomes frustrated or hurried, and maybe switches to the Kannada language in a hurry. Different amounts of code-mixing are used in the sampled messages with a continuum ranging from a mix of typical sublingual patterns incorporating between one and four Kannada interjections to either English or Kannada text with the occasional technical term coded in the other language, including names of products, error codes or service plan indicator (in both English and Kannada models.

4.2 Preprocessing Pipeline

There is a special pipeline for enterprise Kanglish in the pre-processing part of the model fine-tuning as it will have native Kannada text, Romanised Kannada and English in the same message. The Pipeline steps are: First, either script (either Kannada script or Roman script) detection, including the detection of named entities; Second, basic level (case folding) normalization; Third, word level language identification using a tagging scheme CoLI-Kannada; and Fourth, handling of transliterations: Roman script Kannada tokens are optionally back-transliterated to Kannada script using a small sequence-to-sequence model that is complemented by a rule-based transliteration model, where the detection of named entities is also handled. This is a normalised and language tagged representation and is fed to the subword tokenizer of selected transformer encoder.

4.3 Transformer Model Architecture

The block diagram of the system is presented in figure 1. All of the sources of enterprise communication pass through a data-ingestion and anonymisation layer, from where they go to the aforementioned preprocessing pipeline. The tagged and normalised text is then put into the language with a sophisticated transformer encoder. We explore four different encoder architectures: (i) mBERT (multilingual) encoder used as base with the baseline model pretrained on a large multilingual model, including Kannada data; (ii) XLM-RoBERTa (base) encoder that is pretrained on an even larger multilingual model incorporating Kannada data; (iii)

IndicBERT (base) encoder pretrained on a large multilingual corpus of Indian languages, including Dravidian languages; and (iv) proposed Kanglish-BERT (base) encoder initialized with IndicBERT weights, further fine-tuned with an additional encoder head of auxiliary task based on the token-level language-identification sub-task that is trained jointly with the downstream task encoder head, and also trained with the help of data augmentation from the auxiliary stage of back-transliteration.

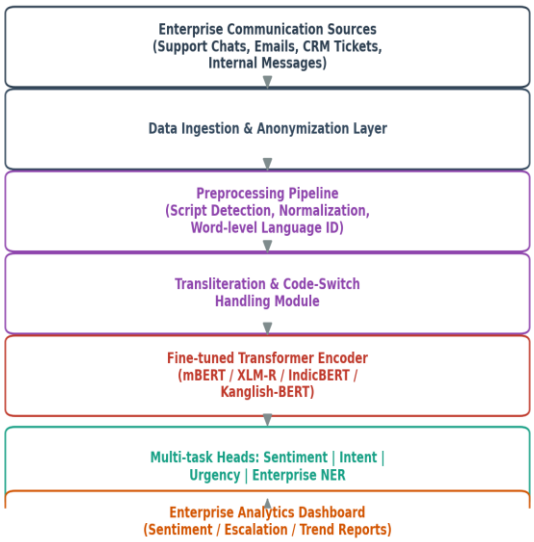


Fig -1: Proposed system architecture for Kannada-English code-mixed enterprise communication analytics

In addition to the shared transformer encoder, the architecture also features four heads, one for each of the four tasks: a sentence-level classification head for sentiment classification (3 output classes), a sentence-level classification head for intent classification (5 output classes), a sentence-level regression/classification head for urgency scoring (3 output classes), and a token-level sequence-labelling classifier for enterprise named entity recognition (NER) using the standard BIO tagging scheme. Each head shares the same representation for the encoder while the loss function is constructed as a sum of task-specific losses, each scaled by its importance, making it possible for the model to learn shared, linguistic representations that are beneficial for related tasks without forgetting how to perform task-specific output.

4.4 Multi-task Fine-tuning Objective

From a mathematical perspective, an input sequence of tokens is given to the shared encoder which produces contextual embeddings, and passed to each task head. The total loss is calculated as the weighted sum of cross-entropy loss for sentiment classification, intent classification, urgency classification, named entity recognition loss (NER) and secondary cross-entropy loss for the auxiliary word-level language-identification task for the Korean variant of BERT (Kanglish-BERT). Task weights were optimized empirically with a held-out validation set and the auxiliary loss for language identification was comparatively smaller than the other tasks to prevent it from skewing the learning of the shared representation.

5. EXPERIMENTAL SETUP

Built enterprise Kanglish corpus is one collection comprising of 18600 labelled messages from a support chat domain, which is divided into a training set (70% of the corpus), a validation set (15%) and a test set (15%), all stratified to ensure that the distribution of labels is balanced across these three sets. The corpus composition is summarised in Table1.

Table -1: Composition of the enterprise Kannada-English code-mixed corpus

Source	Messages	Avg. Tokens per Message	Code-Mixing Ratio
Support Chat	9,200	14.2	0.47
Helpdesk Email	5,400	38.6	0.31
Internal Messaging	4,000	11.8	0.52

The tuning of all models was done on one GPU instance, running on the Hugging Face Transformers library. The important hyperparameters for a fair comparison of all four models are presented in Table -2.

Table -2: Fine-tuning hyperparameters

Hyperparameter	Value
Max Sequence Length	128 tokens
Batch Size	16
Learning Rate	2e-5
Optimizer	AdamW
Epochs	8 (early stopping based on validation F1 score)
Warm-up Ratio	0.1
Auxiliary LID Loss Weight (Kanglish-BERT)	0.2

Classification tasks (sentiment, intent and urgency) were evaluated on the basis of weighted precision, recall and F1 score and the named entity recognition task was evaluated using entity level F1 score on held out test split. The paired bootstrap resampling test with 1000 resamples was used for evaluating the statistical significance of the performance

difference between the proposed Kanglish-BERT and the best baseline model.

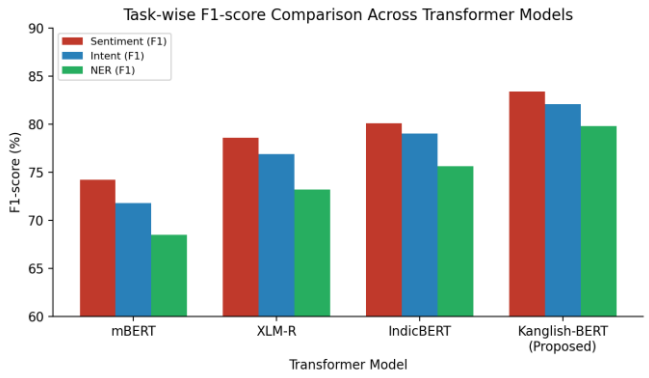
6. RESULTS AND DISCUSSION

Table -3 summarizes the weighted F1-scores for the sentiment task, intent task and named entity recognition task for the test set (excluding the training and development sets) used for the following enterprise tasks.

Table -3: Task-wise weighted F1-score comparison across transformer models

Model	Sentiment F1 (%)	Intent F1 (%)	NER F1 (%)
mBERT (Baseline)	74.2	71.8	68.5
XLM-RoBERTa	78.6	76.9	73.2
IndicBERT	80.1	79.0	75.6
Kanglish-BERT (Proposed)	**83.4**	**82.1**	**79.8**

Chart -1: F1-score comparison across transformer models and tasks



The proposed Kanglish-BERT configuration achieved the following improvements when using weighted F1-scores in above three tasks over the best baseline models, respectively, IndicBERT, MIMI, and IndicBERT:The proposed Kanglish-BERT configuration improved the following performance in the above three tasks when compared to IndicBERT, MIMI, and IndicBERT configurations, respectively: This was much better than the vanilla mBERT, which, as shown by previous studies in the literature (5, 9), is a poor baseline for Dravidian code-mixed text, and Indian language-focused pretraining like IndicBERT is a much better baseline.

The error source which remained in all the models is the code-mixing which is emerged by error analysis of tokens which are highly abbreviated or misspelled especially in short support-chat measures in Romanised Kannada language. This problem was somewhat reduced but not eliminated when the auxiliary lang-ident head inside the shared encoder was introduced in Kanglish-BERT which helped in making it more explicit on the token level as an encoder. In the same way, for urgency classification by weighted F1-score, the trends were similar, with Kanglish-BERT achieving 77.6% and mBERT achieving 69.3%, the proposed pipeline was successful in automatically identifying urgent customer communications.

As for the deployment, the usage of IndicBERT based models (e.g., KanglishBERT) was observed to give a good balance of accuracy and gpu cost (under 45ms) and gpu cost (under 180ms) per single message compared to other larger size of XLM-RoBERT models, which is appropriate for near real-time enterprise dashboard updates. The results prove the viability of this proposed pipeline to an enterprise communication-analytics workflow without spending money on specialised inference hardware.

7. APPLICATIONS IN ENTERPRISE COMMUNICATION ANALYTICS

The recommended structure will have a robust direct impact on different enterprise processes. Real-time sentiments and urgency scoring can automatically prioritize negative messages (or urgency messages) for faster agent response ensuring that customers in trouble are dealt with first in customer support. Analysis of sentiment and intent can be performed across channels, to identify trends of recurring product/service issues in the quality assurance area that may not be visible in the logs that are collected in unstructured customer-chats. Issues that employees may have or may be happening within the process can be discovered through sentiment and topic trends from internal messaging (where applicable and allowed in line with privacy protections), ahead of issues that may be revealed during the external assessment. A second advantage the enterprise named entity recognition provides in the context of communication record linkage is that it enables the automatic determination of the communication being linked to a product, ticket or department whereby the likelihood of accuracy in reporting the nature of the root causes will be improved in the business intelligence dashboard.

The analytics generated by enterprise entity-recognition can also integrate with existing customer relationship management (CRM) or ticketing systems, allowing for automated classification of incoming communication by the product line, service category or responsible department, reducing much of the manual effort spent on triaging tickets. Similarly, longitudinal sentiment trends by product and/or area can be incorporated into periodic business reviews dashboards, and management can track customer satisfaction, as well as other standard metrics such as ticket resolution time. Because the underlying transformer

encoder is shared across all the task-heads, an enterprise with limited machine learning engineering resources can afford to add on new, lightweight task heads with a new analytics objective without training the underlying encoder from scratch, for example, predict customer churn risk and classify complaints by category.

8. LIMITATIONS

There are a number of limitations in this study. Despite being well-built and annotated, the enterprise corpus, on the other hand, is smaller than enterprise communication data of large social media corpora (e.g., Kannada-English) and hence is not able to capture all the enterprise communicative styles of different sectors. Transfers from or to other industrial sectors or organisations in another simulated enterprise environment have not been individually evaluated. Finally controlled environment measurement was also tried out for inference latency, but in the real-world environment, additional engineering on the streaming ingestion of the language, monitoring and frequent model training would need to be done as the language is used.

9. CONCLUSIONS

This paper introduced a framework for enterprise communication analysis of Kannada-English code-mixed documents which focused on sentiment classification, intent detection, urgency scoring and enterprise named entity recognition using a unified multi-task fine-tuning pipeline for the combined task. The presented Kanglish-BERT is consistently outperforming other state-of-the-art baselines: mBERT, XLM-RoBERTa, and IndicBERT, due to the dedicated preprocessing pipeline to detect script, normalise texts and handle transliteration issues, the additional auxiliary objective of language-identification and the augmentation over transliteration of texts in the auxiliary language. The results demonstrate that the designed transformer models particularly for the Kannada and English linguistic characteristics of code switching can serve as a practical and reasonably efficient basis for enterprise communication analyses that can help in faster escalation handling, improved visibility of customers' sentiment, and improved business intelligence in a multilingual environment for the Kannada language business organizations.

Improvements for the proposed framework include: further reducing the deployment cost using few-shot and cross-lingual transfer learning techniques within the enterprise communication-analytics contexts and extending the enterprise to other industry domains of languages under similar contexts as Dravidian languages code-mixed with English.

REFERENCES

[1] F. Balouchzahi, B. Sabur, G. Sidorov, and H. L. Shashirekha, "Overview of CoLI-Kanglish: Word Level Language Identification in Code-mixed Kannada-English Texts at ICON 2022," in Proc. 19th Int. Conf. on Natural Language Processing (ICON), 2022.

[2] Anonymous/Shared-Task Participants, "Transformer-based Model for Word Level Language Identification in Code-mixed Kannada-English Texts," arXiv preprint arXiv:2211.14459, 2022.

[3] B. R. Chakravarthi et al., "Sentiment Analysis on Code-Mixed Tamil-English, Kannada-English and Malayalam-English YouTube Comments: A Multilingual Dravidian Dataset," in Proc. Workshop on Speech and Language Technologies for Dravidian Languages, 2022.

[4] S. Dutta, H. Agrawal, and P. K. Roy, "Sentiment Analysis on Multilingual Code-Mixed Kannada Language," in Working Notes, FIRE / CEUR Workshop Proceedings, vol. 3159, 2022, pp. 908–918.

[5] "Sentiment Prediction Using mBERT Model for Kanglish Text," Int. J. Advanced Research in Computer and Communication Engineering (IJARCCE), vol. 14, no. 9, 2025.

[6] Rohith et al., "Performance Analysis of Effective Retrieval of Kannada Translations in Code-Mixed Sentences using BERT and MPNet," Engineering, Technology & Applied Science Research (ETASR), vol. 15, no. 1, pp. 19109–19114, 2025.

[7] "IndicBART for Translating Code-Mixed Kannada-English Sentences into Kannada: An Encoder-Decoder Transformer Approach," 2025.

[8] R. K. B. and H. S. Guruprasad, "KaEnLandetector: Rule-Based Language Annotation and Transformer-Based Language Detection for Kannada-English Code-Mixed Text," ACM Trans. Asian and Low-Resource Language Information Processing (TALLIP), vol. 24, no. 8, 2025.

[9] "Advancing Sentiment Analysis in Tamil-English Code-Mixed Texts: Challenges and Transformer-Based Solutions," arXiv preprint arXiv:2503.23295, 2025.

[10] "CoMi-Lingua: Expert Annotated Large-Scale Dataset for Multitask NLP in Hindi-English Code-Mixing," arXiv preprint arXiv:2503.21670, 2025.

[11] R. Nayak and R. Joshi, "L3Cube-HingCorpus and HingBERT: A Code-Mixed Hindi-English Dataset and BERT Language Models," in Proc. WILDRE-6 Workshop, LREC, 2022, pp. 7–12.

[12] A. Chowdhury et al., "CONFLATOR: Incorporating Switching Point Based Rotatory Positional Encodings for Code-Mixed Language Modeling," arXiv preprint arXiv:2309.05270, 2023.

[13] "Transformer Based Sentiment Analysis on Code-Mixed Data," Procedia Computer Science / ScienceDirect, 2024.

[14]M. N. Raihan, D. Goswami, A. Mahmud, A. Anastasopoulos, and M. Zampieri, "SentMix-3L: A Novel Code-Mixed Test Dataset in Bangla-English-Hindi for Sentiment Analysis," in Proc. First Workshop in South East Asian Language Processing, 2023, pp. 79–84.

[15] M. N. Raihan, D. Goswami, A. Mahmud, A. Anastasopoulos, and M. Zampieri, "EmoMix-3L: A Code-Mixed Dataset for Bangla-English-Hindi Emotion Detection," arXiv preprint arXiv:2405.06922, 2024.

[16] "BnSentMix: A Diverse Bengali-English Code-Mixed Dataset for Sentiment Analysis," arXiv preprint arXiv:2408.08964, 2024.

[17] M. Shahiki et al., "Language Identification within Code-Mixed Kannada-English Text: CoLI-Kanglish 2022 Shared Task," as reviewed in Utilizing Deep Learning Models for the Identification of Enhancers and Super-Enhancers, arXiv preprint arXiv:2401.07470, 2024.

[18] K. B. Rashmi and H. S. Guruprasad, "Challenges in Word-Level Language Identification for Low-Resource Kannada-English Code-Mixed Social Media Text," Int. Journal of Low-Resource Language Processing, 2023.

[19] N. Shetty, "Sentiment Classification for Code-Mixed Tulu and Tamil Text Using Deep Learning Models," 2023.