# Overview of CoLI-Kanglish: Word Level Language Identification in Code-mixed Kannada-English Texts at ICON 2022

**F. Balouchzahi**[1,*] **S. Butt**[1], **A. Hegde**[2], **N. Ashraf**[3],
**H.L. Shashirekha**[2], **G. Sidorov**[1], and **A. Gelbukh**[1]
[1]Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico
[2]Department of Computer Science, Mangalore University, Mangalore, India
[3]Dana-Farber Cancer Institute, Harvard Medical School, United States
Corresponding: [*]fbalouchzahi2021@cic.ipn.mx

## Abstract

The task of Language Identification (LI) in text processing refers to automatically identifying the languages used in a text document. LI task is usually been studied at the document level and often in high-resource languages while giving less importance to low-resource languages. However, with the recent advancement in technologies, in a multilingual country like India, many low-resource language users post their comments using English and one or more language(s) in the form of code-mixed texts. Combination of Kannada and English is one such code-mixed text of mixing Kannada and English languages at various levels. To address the word level LI in code-mixed text, in CoLI-Kanglish shared task, we have focused on open-sourcing a Kannada-English code-mixed dataset for word level LI of Kannada, English and mixed-language words written in Roman script. The task includes classifying each word in the given text into one of six predefined categories, namely: Kannada (kn), English (en), Kannada-English (kn-en), Name (name), Location (location), and Other (other). Among the models submitted by all the participants, the best performing model obtained averaged-weighted and averaged-macro F1 scores of 0.86 and 0.62 respectively.

## 1 Introduction

South Asia is the most linguistically diverse region in the world that embodies more than 650 different languages[1] and India is a multilingual country having a rich heritage of languages in South Asia. Kannada is one of the Dravidian[2] languages as well as scheduled languages of India and the official and administrative language of Karnataka state with more than 40 million native Kannada speakers. A significant number of people in this region are comfortable using English in addition to their native/local/regional language for the day-to-day communication. These multilingual speakers preferably use multiple scripts and/or languages to post their comments/ideas/opinions on social media platforms, making code-mixing a default language on social media.

Code-mixing can be carried out at the paragraph, sentence or word level and even at sub-word level (Chakravarthi et al., 2020; Hegde et al., 2022a). People usually mix their native and/or local language with English and prefer to write the content mostly in Roman script rather than using the native script as most of the keyboard layouts of computers and keypads of smartphones have Roman alphabets by default (Balouchzahi et al.).

People who write Kannada find difficult to use Kannada script while posting comments/reviews on social media mainly because of the difficulty in keying the consonant conjuncts (*ottaksharas*) and consonants with the secondary forms of vowels (*gunitaskaras*) (Kittel, 1903), using Roman keyboards/keypads and hence prefer to use Roman script on social media (Balouchzahi et al., 2021b). The situation remains the same for most of the Indian languages as they have their own script.

Social media platforms have given their users the freedom of writing text very casually without following the grammar or syntax of any language. This has resulted in a huge volume of user-generated content which includes incomplete words and/or sentences, catchy phrases, user-defined short forms for words (e.g., 'gn8' for 'good night'), different slangs (e.g., meme, Gmeet), abbreviations ('OMG' for 'Oh my God'), recurrent characters ('soooooo sad' for 'so sad'), etc. The presence of these informal words in any text makes it difficult to understand the content (Shashirekha et al., 2022). Further, a code-mixed scenario where words of one language are transcribed with words of other languages as prefixes or suffixes creates a

[1]https://www.deccanherald.com/content/652273/intl-meet-south-asian-languages.html
[2]https://en.wikipedia.org/wiki/Kannada



*Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 38 - 45

lot of problems to analyze the text, particularly due to conflicting phonetics.

The increasing number of social media users is increasing the user-generated content which makes it difficult to handle this text manually (Scotton, 1982). This demands the tools and techniques that can process the user-generated text automatically for various applications.

The preliminary step in handling code-mixed text for many of the Natural Language Processing (NLP) tasks like Machine Translation (Patel and Parikh, 2020), Parts-Of-Speech tagging (Dowlagar and Mamidi, 2021), Sentiment Analysis (Bansal et al., 2020; Balouchzahi et al., 2021c; Balouchzahi and Shashirekha, 2021), Emotion Analysis (Hegde et al., 2022b), Hate Speech and Offensive Language Identification (Balouchzahi et al., 2021a; Hegde et al., 2021), Hope Speech Detection (Gowda et al., 2022), Identification of Native Language (Nayel and Shashirekha, 2018, 2017), etc., is identifying the language of each word/phrase/sentence.

Several research works in LI tasks have been carried out focusing on high-resource languages like French-English, Spanish-English, and German-English. However, very little attention is given to the low-resource Indian languages. Furthermore, code-mixing is quite common in a multilingual country like India where many people are bilingual and English is considered as the official language along with the local/administrative language. Hence, in India, code-mixing is mostly observed between any Indian language and English in social media text (Balouchzahi and Shashirekha, 2020).

The rapidly increasing code-mixed content on social media in Indian languages in general and Kannada-English in particular requires efficient methods to perform LI at word level.

## 2 Literature Review

Recent decades have witnessed the immense interest of researchers in code-mixed text specifically for low-resource and under-resource languages and few LI works have also been carried out as a part of handling such code-mixed text. Word level LI is modeled as a typical supervised learning problem and various Machine Learning (ML) and Deep Learning (DL) algorithms are experimented for the same. Some of the relevant works are described below:

(Chaitanya et al., 2018) developed learning models for word level LI of Hindi-English code-mixed data using feature vectors generated by the Continuous Bag of Words (CBOW) and Skipgram models. They experimented with various ML models including: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Gaussian Naive Bayes (GNB), k-Nearest Neighbor (kNN), and Adaptive Boosting (Adaboost), on the dataset consisting of 7,210 words selected from the corpus prepared by (Jamatia and Das, 2016). Among all the models, SVM classifiers obtained the highest accuracies of 67.33% and 67.34% using CBOW and Skipgram models respectively. A word level LI in code-mixed Telugu-English text proposed by (Gundapu and Mamidi, 2020), tokenized 1,987 Telugu-English code-mixed sentences obtained from the International Conference on Natural Language Processing (ICON) 2015 shared task dataset[3] and manually tagged the tokenized words with Parts-Of-Speech (POS) and LI tags. By using previous, current and next words and their POS tags, length of the word, and character n-grams in the range n = (1, 3) as features, they trained Conditional Random Field (CRF) classifier to perform word level LI and obtained an accuracy of 91.28%.

(Mandal and Singh, 2018) proposed a multichannel Neural Network (NN) model for LI of code-mixed Hindi-English and Bengali-English text using contextual information. They selected 6,000 instances from the dataset developed by (Patra et al., 2018) and Mandal et al. (2018) for Hindi-English and Bengali-English respectively and implemented multichannel neural associations by combining Convolutional Neural Network (CNN) and Long short-term memory (LSTM) models coupled with BiLSTM-CRF for word level LI. Their proposed models obtained accuracies of 93.32% and 93.28% for Hindi-English and Bengali-English data respectively. (Thara and Poornachandran, 2021) created a dataset for word level LI in code-mixed English-Malayalam text and implemented transformer-based models for LI. The authors extracted 50K code-mixed English-Malayalam comments from YouTube and tokenized them to obtain 7,75,430 words. These words are then annotated with the language to which they belong to using an unsupervised approach. Transformer-based multilingual Bidirectional Encoder Representations from Transformers (mBERT), Cross-lingual

[3] https://ltrc.iiit.ac.in/icon2015/

Language Model for Robustly Optimized BERT (XLM-RoBERTa), CamemBERT, Distilled version of BERT (DistilBERT), and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) models, are fine-tuned to perform LI. Among all the models, fine-tuned ELECTRA model performed best with an F1 score of 0.9933.

Word and Character embedding-based learning models presented by (Veena et al., 2018) for LI of code-mixed Hindi-English text are experimented on ICON 2016 shared task dataset (Jamatia and Das, 2016) that consists of 772, 1,096, and 763 sentences from Facebook, Twitter, and WhatsApp respectively. By collecting additional code-mixed Hindi-English text from other resources, word and character ngrams are used to train three skip-gram models with n = 1, 3, 5 which in turn is used to train the SVM models for LI. Compared to various SVM models trained on word-based and character-based embeddings, SVM classifier trained on character-based 5-gram embeddings obtained better accuracy.

Even though few research works are carried out in low-resource Indian languages like Kannada, Tamil, Telugu, etc., no works have been reported on word level LI in code-mixed Kannada-English text. This provides scope for research at word level LI in code-mixed Kannada-English text.

## 3 Task Description

The task of automatically identifying languages used in a given text is called LI and it is a preprocessing step for many applications. LI at the word level can be viewed as a sequence labeling problem where each and every word in a sentence is tagged with one of the languages in the predefined set of languages. Despite a lot of work being done in LI, the problem of LI in the code-mixed scenario is still a long way from being illuminated (Mandal and Singh, 2018).

To address word level LI in code-mixed Kannada-English texts, these texts are extracted from Kannada video comments in YouTube to construct CoLI-Kenglish (Shashirekha et al., 2022) dataset.

## 4 Dataset

Comments for Kannada videos in YouTube are scraped using the youtube-comment-downloader[4] and are used to build CoLI-Kenglish dataset (Shashirekha et al., 2022). The scraped texts contain around 1,00,000 comments from 373 Kannada YouTube videos. Preprocessing involves the removal of duplicate comments and comments written only in Kannada script. After preprocessing, the total number of comments amounts to 72,815. The nature of comments are generally in one of the following forms:

- Only in Kannada
- Only in English
- Combination of Kannada and English
- Other languages e.g., Hindi, Telugu and Tamil

A random sample of around 10% of the text is annotated by two native Kannada speakers to generate CoLI-Kenglish dataset and the rest of raw text is released as additional Kannada-English code-mixed resource.

The annotated CoLI-Kenglish dataset contains 19,432 unique words extracted from nearly 7,000 sentences that are categorized into 6 classes, namely: 'Kannada', 'English', 'Mixed-language', 'Name', 'Location' and 'Other'. While 'Kannada' and 'English' classes represent Kannada and English words respectively, 'Mixed-language' class represents words created using a combination of Kannada and English in any order. 'Name' class represents the names of persons and 'Location' class the names of locations or places. Any other words are represented as an 'Other' class. The words described by 'Mixed-language' pose a real challenge to LI task as these words are framed by various combinations of English/Kannada words and Kannada/English affixes (prefixes and suffices). The beauty as well as the complexity of these mixed-language words lies in the word pattern created by an individual posting comments on social media. Description of the class labels and their samples along with the English translation are presented in Table 1 and the statistics of CoLI-Kenglish dataset in terms of Train and Test set are shown in Table 2. The statistics of the CoLI-Kenglish dataset illustrates that the dataset is imbalanced.

[4] https://github.com/egbertbouman/youtube-comment-downloader

| Category | Tag | Description | Samples |
|---|---|---|---|
| Kannada | kn | Kannada words written in Roman script | kopista (one who get angry soon), baruthe (will come), barbeku (must come) |
| English | en | Pure English words | small, need, take, important |
| Mixed-language | kn-en | Combination of Kannada and English words in Roman script | coolagiru (cool + agiru, be cool), leaderge (leader + ge, to a leader), homealli (home + alli, inside home) |
| Name | name | Words that indicate name of person (including Indian names) | Madhuswamy, Hemavati, Swamy |
| Location | location | Words that indicate locations | Karnataka, Tumkur, Bangalore |
| Other | other | Words not belonging to any of the above categories and words of other languages | Znjdjfjbj – not a word<br>kannada words in kannada script<br>hindi words in Devanagari script<br>hindi words in Roman script<br>tamil words in Tamil script |

Table 1: Description of the classes and their samples in CoLI-Kenglish dataset

| Tag | Train set | Test set |
|---|---|---|
| kn | 6,526 | 2,194 |
| en | 4,469 | 1,812 |
| kn-en | 1,379 | 93 |
| name | 708 | 354 |
| location | 102 | 31 |
| other | 1,663 | 100 |
| **Total** | 14,847 | 7,241 |

Table 2: Statistics of Train and Test set

## 5 Evaluation Metrics

In the case of an imbalanced dataset, categories with a larger number of samples affect the averaged-weighted scores and could be high always. Therefore, reporting only weighted scores could provide misleading information about models' performance. Hence, inspired by (Balouchzahi et al., 2022), code-mixed LI models for imbalanced CoLI-Kenglish dataset are evaluated using macro-averaged and weighted-averaged F1 scores.

## 6 Baselines

CoLI-ngrams - the best performing model proposed by (Shashirekha et al., 2022) employ a feature engineering module that generates a feature set of prefixes and suffixes of length 1, 2 and 3 along with char n-grams (n = 2, 3, 5) from words, and Byte-Pair Encoding (BPE) embeddings of sub-word n-grams (n = 1, 2, 3). The extracted features are vectorized using TfidfVectorizer[5] to train Linear SVM (LSVM), Multi-layer Perceptron (MLP), and Logistic Regression (LR) classifiers. These three models are used as baselines in this shared task. All the models are trained with default parameters.

[5]http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

## 7 Overview of the Submitted Approaches

Thirty different runs are submitted by eight different teams for the Kanglish 2022 shared task and eventually six teams submitted their working notes. Figure 1 refers to the different learning approaches used by the participants in this shared task to submit the runs. The findings indicate that, while 54% of the participants experimented different transformers, 27% used traditional ML models and the remaining used DL models. Figure 2 shows that about 46% of run submissions are made by employing pretrained Language Model (LM) or pretrained embeddings and 27% did not use any pretrained models for the task.

**Team Tiya1012** presented a transformer-based model by fine-tuning DistilBERT-based-cased model on the CoLI-Kenglish dataset and obtained 0.62 averaged-macro F1 score and ranked first in the competition.

**Team Abyssinia** experimented different LM models, namely: BERT, mBERT, XLM-R and RoBERTa from HuggingFace with a LSTM architecture. Among all the LM models, both mBERT and XLM-R with an averaged-macro F1 score of 0.61 outperformed the rest of the models and also ranked second in the shared task.

**Team PDNJK** also explored several transformer-based models for the task of LI in code-mixed Kannada-English words and their best performing

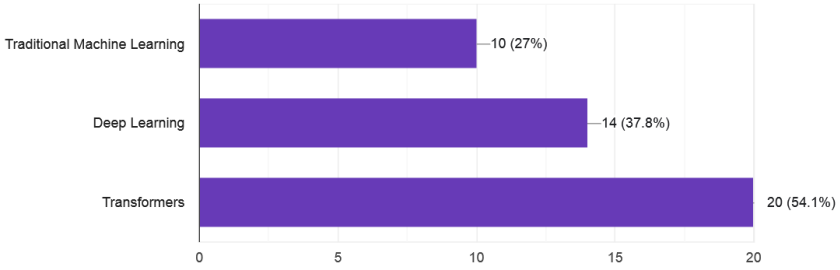


Figure 1: Learning approaches used by participants

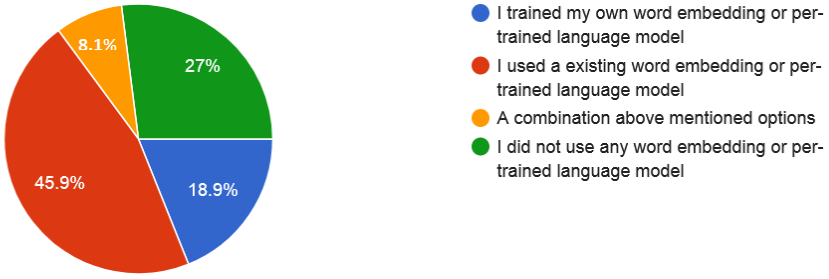


Figure 2: Pretrained models used by participants

model using BERT scored an averaged-macro F1 score of 0.57 and ranked fourth in the shared task.

**Team Habesha** trained character-level LSTM and BiLSTM models with attention that reads the text as a sequence of characters. The proposed BiLSTM model outperformed the LSTM model and obtained an averaged-macro F1 score of 0.61 and ranked second in the shared task.

**Team Lidoma** explored character n-grams to generate character TF-IDF to train traditional ML classifiers. Among all the classifiers they explored, the highest performance of an averaged-macro F1 score of 0.58 was reported with a simple kNN classifier. Similarly, Bag-of-Characters were turned into character vectors by **Team NLP_BFCAI**. They introduced a character representation called Bag-of-n-Characters model which has very similar structure to character n-grams and experimented several traditional ML algorithms. Eventually, the RF model on the proposed features obtained the highest averaged-macro F1 of 0.43.

## 8 Results and Discussion

The best results obtained for each team among all the predictions submitted by them are presented in Table 3 along with the results of the baseline models. Comparison of the results of the participating teams with that of the baseline models shows a slight improvement on F1-score for the first three best performing teams. The best averaged-macro F1 score of 0.62 shows the difficulty of the shared task. Further, our baselines utilizing n-grams generated from BPEmb sub-words, characters and affixes had a better performance of models that experimented only character n-grams.

Other findings indicate that, all teams who employed NN and transformer models outperformed the baselines and other traditional ML classifiers. In general, the higher weighted scores are the results of successful predictions for pure English and Kannada words and the difficulty on identifying mixed-language words and less frequent entities resulted in less performance for macro scores. Most of the teams relied on multilingual transformers or only character n-grams for solving the problem of LI in code-mixed text. This reveals that the participants have only a shallow understanding of code-mixed texts. No method was used by the participants that could directly target the issue of code-mixed texts except the multilingual transformers that partially handled the task.

## 9 Conclusion

The task of LI is a primary step for many NLP tasks that are usually overlooked for low-resource languages. However, the recent advancement in technologies caused a rapid increase in the volume of texts in low-resource languages. These texts on so-

| Rank | Team name | Weighted | | | Macro | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1-score | Precision | Recall | F1-score |
| 1 | Tiya1012 | 0.87 | 0.85 | 0.86 | 0.67 | 0.61 | 0.62 |
| 2 | Abyssinia | 0.85 | 0.84 | 0.84 | 0.62 | 0.62 | 0.61 |
| 2 | Habesha | 0.85 | 0.83 | 0.84 | 0.66 | 0.6 | 0.61 |
| - | LSVM-Baseline | 0.84 | 0.84 | 0.83 | 0.67 | 0.57 | 0.59 |
| 3 | Lidoma | 0.83 | 0.83 | 0.83 | 0.64 | 0.56 | 0.58 |
| 4 | PDNJK | 0.86 | 0.85 | 0.86 | 0.58 | 0.58 | 0.57 |
| - | MLP-Baseline | 0.84 | 0.81 | 0.82 | 0.60 | 0.60 | 0.57 |
| - | LR-Baseline | 0.84 | 0.84 | 0.83 | 0.69 | 0.53 | 0.56 |
| 5 | NLP_BFCAI | 0.73 | 0.73 | 0.72 | 0.52 | 0.41 | 0.43 |
| 6 | iREL | 0.68 | 0.62 | 0.64 | 0.38 | 0.45 | 0.39 |
| 7 | JUNLP | 0.69 | 0.67 | 0.67 | 0.33 | 0.34 | 0.3 |
| 8 | PresiUniv | 0.57 | 0.59 | 0.53 | 0.22 | 0.22 | 0.2 |

Table 3: Participating team's best run score in the shared task

cial media are often a mixture of low-resource language with English resulting in code-mixed texts. In the code-mixed scenario, a sentence alone can have multiple languages at word level. Hence, the aim of Kanglish 2022 shared task was to promote word level LI for Kannada-English code-mixed texts. Initially, thirteen teams registered for the task and eventually more than thirty different runs were submitted by eight different teams. The majority of teams explored different NN models including transformers for the task. A fine-tuned DistilBERT model outperformed the rest of the models with averaged-weighted and averaged-macro F1 scores of 0.86 and 0.62 respectively.

The observation of performances of different models in the shared task reveals the difficulty of the LI task in code-mixed text. These difficulties are mainly due to the nature of code-mixed texts that do not follow the rules of and grammar of any language. This task aims to attract the attention of researchers for word level LI of different language pairs in code-mixed text. In future work, we would like to include more mixed-language words into CoLI-Kenglish dataset and also extend the corpus to different Dravidian languages including Tamil, Malayalam, etc.

## Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

## References

Fazlourrahman Balouchzahi, Aparna B K, and H L Shashirekha. 2021a. MUCS@DravidianLangTech-EACL2021:COOLI-Code-Mixing Offensive Language Identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 323–329, Kyiv. Association for Computational Linguistics.

Fazlourrahman Balouchzahi, Aparna B K, and H L Shashirekha. 2021b. MUCS@LT-EDI-EACL2021:CoHope-Hope Speech Detection for Equality, Diversity, and Inclusion in Code-Mixed Texts. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 180–187, Kyiv. Association for Computational Linguistics.

Fazlourrahman Balouchzahi and H L Shashirekha. 2021. LA-SACo: A Study of Learning Approaches for Sentiments Analysis in Code-Mixing Texts. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 109–118, Kyiv. Association for Computational Linguistics.

Fazlourrahman Balouchzahi and HL Shashirekha. 2020. MUCS@ Dravidian-CodeMix-FIRE2020: SACO-Sentiments Analysis for CodeMix Text. In *FIRE (Working Notes)*, pages 495–502.

Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha, and Grigori Sidorov. 2021c. CoSaD-

Code-Mixed Sentiments Analysis for Dravidian Languages. In *CEUR Workshop Proceedings*, pages 887–898.

Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha, Grigori Sidorov, and Alexander Gelbukh. A Comparative Study of Syllables and Character Level N-grams for Dravidian Multi-script and Code-mixed Offensive Language Identification. In *Journal of Intelligent & Fuzzy Systems*, Preprint, pages 1–11. IOS Press.

Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. 2022. PolyHope: Dataset Creation for a Two-Level Hope Speech Detection Task from Tweets. In *arXiv preprint arXiv:2210.14136*.

Neetika Bansal, Vishal Goyal, and Simpel Rani. 2020. Experimenting Language Identification for Sentiment Analysis of English Punjabi Code Mixed Social Media Text. In *International Journal of E-Adoption (IJEA)*, pages 52–62. IGI Global.

Inumella Chaitanya, Indeevar Madapakula, Subham Kumar Gupta, and S Thara. 2018. Word Level Language Identification in Code-mixed Data using Word Embedding Methods for Indian Languages. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1137–1141. IEEE.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P McCrae. 2020. Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210.

Suman Dowlagar and Radhika Mamidi. 2021. A Pretrained Transformer and CNN Model with Joint Language ID and Part-of-Speech Tagging for Code-mixed Social-media Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 367–374.

Anusha Gowda, Fazlourrahman Balouchzahi, Hosahalli Shashirekha, and Grigori Sidorov. 2022. MUCIC@LT-EDI-ACL2022: Hope Speech Detection using Data Re-Sampling and 1D Conv-LSTM. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 161–166, Dublin, Ireland. Association for Computational Linguistics.

Sunil Gundapu and Radhika Mamidi. 2020. Word Level Language Identification in English Telugu Code Mixed Data. In *arXiv preprint arXiv:2010.04482*.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022a. Corpus Creation for Sentiment Analysis in Code-Mixed Tulu Text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40.

Asha Hegde, Mudoor Devadas Anusha, and Hosahalli Lakshmaiah Shashirekha. 2021. Ensemble based machine learning models for hate speech and offensive content identification. In *Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org*.

Asha Hegde, Sharal Coelho, and Hosahalli Shashirekha. 2022b. MUCS@DravidianLangTech@ACL2022: Ensemble of Logistic Regression Penalties to Identify Emotions in Tamil Text. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 145–150, Dublin, Ireland. Association for Computational Linguistics.

Anupam Jamatia and Amitava Das. 2016. Task Report: Tool Contest on POS Tagging for Code-mixed Indian Social Media (Facebook, Twitter, and WhatsApp) text@ ICON 2016. In *Proceedings of ICON*.

Ferdinand Kittel. 1903. A Grammar of the Kannaḍa Language in English: Comprising the Three Dialects of the Language (Ancient, Mediæval and Modern). Basel Mission Book and Tract Depository.

Soumil Mandal, Sainik Kumar Mahata, and Dipankar Das. 2018. Preparing Bengali-English Code-mixed Corpus for Sentiment Analysis of Indian Languages. In *arXiv preprint arXiv:1803.04000*.

Soumil Mandal and Anil Kumar Singh. 2018. Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 116–120. Association for Computational Linguistics.

Hamada A Nayel and HL Shashirekha. 2017. Mangalore-University@ INLI-FIRE-2017: Indian Native Language Identification using Support Vector Machines and Ensemble Approach. In *FIRE (Working Notes)*, pages 106–109.

Hamada A Nayel and HL Shashirekha. 2018. Mangalore University INLI@ FIRE2018: Artificial Neural Network and Ensemble based Models for INLI. In *FIRE (Working Notes)*, pages 110–118.

Devshree Patel and Ratnam Parikh. 2020. Language Identification and Translation of English and Gujarati Code-mixed Data. In *2020 International Conference on emerging trends in information technology and engineering (ic-ETITE)*, pages 1–4. IEEE.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment Analysis of Code-mixed Indian Languages: An Overview of Sail_Code-mixed Shared Task@ ICON-2017. In *arXiv preprint arXiv:1803.06745*.

Carol Myers Scotton. 1982. The Possibility of Code-Switching: Motivation for Maintaining Multilingualism. In *Anthropological linguistics*, pages 432–444.

Hosahalli Lakshmaiah Shashirekha, Balouchzahi Fazlourrahman, Mudoor Devadas Anusha, and Sidorov Grigori. 2022. CoLI-Machine Learning Approaches for Code-mixed Language Identification at Word Level in Kannada-English Texts. In *Special Issue of Acta Polytechnica*.

S Thara and Prabaharan Poornachandran. 2021. Transformer Based Language Identification for Malayalam-English Code-Mixed Text. In *IEEE Access*, pages 118837–118850. IEEE.

PV Veena, M Anand Kumar, and KP Soman. 2018. Character Embedding for Language Identification in Hindi-English Code-mixed Social Media Text. In *Computación y Sistemas*, pages 65–74. Instituto Politécnico Nacional, Centro de Investigación en Computación.