Findings of the Shared Task on Sentiment Analysis in Tamil and Tulu Code-Mixed Text

Asha Hegde[1] , **Bharathi Raja Chakravarthi**[2], **Hosahalli Lakshmaiah Shashirekha**[1], **Rahul Ponnusamy**[2], **Subalalitha Chinnaudayar Navaneethakrishnan**[3], **Lavanya Sambath Kumar**[4] , **Durairaj Thenmozhi**[5], **Martha Karunakar**[5] , **Shreya Sriram**[5], **Sarah Aymen**[5]

[1]Department of Computer Science, Mangalore University, Mangalore, India
[2]National University of Ireland Galway, Ireland
[3]SRM Institute of Science And Technology, Tamil Nadu
[4]Madras Institute of Technology, Tamil Nadu, [5]SSN College of Engineering, Tamil Nadu
bharathiraja.akr@gmail.com

Abstract

In recent years, there has been a growing focus on Sentiment Analysis (SA) of code-mixed Dravidian languages. Despite this, there is currently lack of research on SA specifically tailored for code-mixed Dravidian languages, highlighting the need for further exploration and development in this domain. In this view, "Sentiment Analysis in Tamil and Tulu - DravidianLangTech" shared task at Recent Advances in Natural Language Processing (RANLP) - 2023 is organized. This shared task consists two language tracks: code-mixed Tamil and Tulu texts. Tulu text is first ever explored in public domain for SA in this shared task. Fifty seven research teams registered for the shared task and we received 27 systems each for code-mixed Tamil and Tulu texts. The performance of the systems (developed by participants) has been evaluated in terms of macro average F1 score. The top system for code-mixed Tamil and Tulu texts scored macro average F1 scores of 0.32, and 0.542 respectively. The high quality and substantial quantity of submissions demonstrate a significant interest and attention in the analysis of code-mixed Dravidian languages like Tamil and Tulu.

1 Introduction

Sentiment Analysis (SA) is a task to understand how people perceive and react to a particular topic, and it has gained significant attention in both academic and industrial settings over the past two decades. On social media, SA provides insights into the opinions and emotions expressed by individuals towards a brand or product online, or a video, movie, or an incident (Chakravarthi et al., 2022). The field of SA has evolved to encompass the collection and analysis of data from social media posts, allowing organizations to gain a comprehensive understanding of public sentiment towards their brand. By utilizing SA techniques, companies can gauge the overall perception of their brand, identify positive or negative sentiments, and make informed decisions about marketing strategies or customer relationship management. Unlike a simple examination of mentions or comments, SA delves deeper into the sentiments and evaluations conveyed in terms of posts shared by people on various social media platforms (Chakravarthi et al., 2020c).

One specific challenge in SA on social media is dealing with code-mixed texts. Code-mixing is a common occurrence in multilingual communities, where individuals blend words, morphemes, and phrases from two or more languages in their speech or writing (Kachru, 1978; Bali et al., 2014). This phenomenon poses a challenge for SA systems, particularly when they are written in non-native scripts, such as using roman characters to represent languages that traditionally use different scripts (Hegde and Shashirekha, 2022). The complexity arises from the presence of code-switching at various linguistic levels, including phonological, lexical, and syntactic aspects of the text. These intricate language patterns make it difficult for SA models trained on monolingual data to accurately interpret the sentiments expressed in code-mixed texts.

The demand for SA on social media texts, particularly those that are code-mixed, has been on the rise. Companies recognize the importance of understanding the sentiments and opinions expressed in diverse linguistic contexts, as it enables them to cater to a broader range of customers and tailor their marketing strategies accordingly. Researchers and practitioners continue to develop and refine SA techniques to effectively handle code-mixed texts and provide accurate SA results in multilingual settings Balouchzahi and Shashirekha (2020). To effectively analyze sentiment in code-mixed texts, specialized techniques and models need to be de-

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 64–71, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_009

veloped. These approaches should consider the unique linguistic characteristics of code-mixing, including the mixing of different languages, the potential shifts in sentiment across languages, and the context-dependent nature of sentiment interpretation Chakravarthi et al. (2020a). Researchers and practitioners are actively working on improving SA systems to handle code-mixed data, as the prevalence of code-mixing on social media and other online platforms continues to grow Chakravarthi et al. (2021).

This shared task introduces two gold standard corpora for SA of code-mixed text in Dravidian languages, namely Tamil-English and Tulu-Kannada-English. The corpora serve as a valuable resource for researchers and practitioners working on SA in multilingual contexts, allowing them to develop and evaluate models that can effectively handle code-mixed data in these specific language pairs. The objective of this task is to determine the sentiment polarity of code-mixed comments/posts in Tamil-English and Tulu-Kannada-English, sourced from social media. This dataset includes annotations of sentiment polarity at the comment/post level, aiming to identify whether the sentiment expressed is 'Positive', 'Negative', 'Neutral', or 'Mixed Feeling'. The training and development sets contain 33,989 and 3,786 sentences for Tamil, and 6,458 and 781 sentences for Tulu, respectively. Further details regarding the dataset annotation can be found in references Chakravarthi et al. (2020b) and Hegde et al. (2022a).

Tamil - is the first language considered as one of the longest surviving classical languages in India Subalalitha (2019); Srinivasan and Subalalitha (2019). It is a scheduled language under the Indian constitution and official language of the Indian states of Tamil Nadu and Puducherry. In addition, it is considered as one of the national languages of Singapore and SriLanka. Tamil is spoken by sizable minority in four more south Indian states, in addition to Kerala, Karnataka, Andhra Pradesh, Telangana, and the Union Territory of Andaman and Nicobar Islands. The Tamil Nadu State Department of Archaeology and Archaeological Survey of India has recorded first Tamil script in 580 BCE on pottery from Keezhadi, Sivagangai, and Madurai districts of Tamil Nadu, India Sivanantham and Seran (2019). Tamil script is known as Tamili or Tamil-Brahmi and it consists of 18 consonants, 12 vowels, and 216 compound letters followed by a special character Hewavitharana and Fernando (2002); Hegde et al. (2022b).

Tulu - is a prominent Dravidian language spoken by approximately 2.5 million individuals primarily in the Dakshina Kannada and Udupi districts of Karnataka, as well as some parts of Kasaragod in Kerala. It holds great significance as the mother tongue for its speakers, who have made significant contributions to Karnataka's cultural history and, in turn, to Indian culture as a whole. Tulu retains several features of ancient Dravidian languages while also introducing innovations not found in other Dravidian languages Padmanabha Kekunnaya. It utilizes its own script called Tigalari, derived from the Grantha script, which is no longer in use. The Tulu script consists of 52 letters, including 16 vowels and 36 consonants Antony et al. (2012); Hegde et al. (2022c).

2 Task Description

Sentiment analysis in Tamil and Tulu aims to determine the sentiment polarity of social media comments and posts that are written in both Tamil-English and Tulu-English. The following schema was used to annotate the data for sentiments.

- Positive state: The text contains an explicit or implicit signal that the speaker is in a positive state, such as content, appreciative, at ease, and forgiving.
- Negative state: The language contains a clue that indicates the speaker is in a negative state, such as sad, angry, nervous, or violent.
- Mixed Feeling: The sentence contains a clue, either explicitly or implicitly, indicating the speaker is having both good and bad feelings.
- Neutral/unknown state: The text does not contain any explicit or implicit clue of the speaker's feelings.

To aid in improved understanding, the text was translated into Tamil and Tulu and sent to the annotators. A minimum of three annotators contributed to each sentence's annotation.

Sentiment analysis in Tamil and Tulu task involves polarity categorization at the message level. Systems must categorize a YouTube comment into one of the four classes: positive, negative, neutral, or mixed emotions for Tamil-English dataset and positive, negative, unknown, or mixed emotions for Tulu-English dataset. Our datasets include code-switching at three different levels: tag, intra-sentential, and inter-sentential. The comments for the Tamil-English dataset were typed in Roman characters using either Tamil vocabulary and English grammar, or vice versa. Similar to this, comments from the Tulu-English dataset were typed in Roman characters using either Tulu or English grammar and lexicon. The following Tamil-English dataset samples show how this scripting pattern is used.

- Hotel vechu govrama vaalra thirunangaigalum irukaanga.ithu pondra theru porukki kalum irukaanga.-*There are trans women who live with respect by owning hotel. There are street vendors like this*. Tag switching with English words.
- 5rs kudutha pogamatamga 10 or 20 kuduka sollu thittuvaanga ! Idhukku Peru dha vazhipari. - *If we give 5 rupees they won't leave they will scold to give 10 or 20 rupees. This is what named as robbery*. Tamil words written in Roman script with no English switch.
- Hello eallarum apadi illa. nanum tirunagai than. nalaiku si exam elutha poren. tnpsc grp 2 eluthiruken. theriyama ellarum ore mathiri nu ninaikathinga - *Everybody is not like this.I am also a trans woman. Tomorrow going to write SI exam. Written TNPSC group 2 exam. Without knowing do not think like this.*. Inter-sentential switch
- True i didn't give single Paisa to this people. - *True I didn't give single paise to these people.* Intra-sentential switch between clauses.

The following Tulu-English dataset samples show how this scripting pattern is used.

- ayyo encha la onji love letter unda- *Oh is there a love letter like this*. -Tag switching with English words.
- Aye dayeg mandde kanatha patherunu - *Why he is talking like a mad*- Tulu words written in Roman script with no English switch.
- Super Arpit Anna picture tuvodunde ejji devere Mande haal and Kali review tude - *Super Arpith brother, no need to see the picture. Oh God by seeing the review itself mind got ruined*. Intra-sentential switch
- masth edde ithand. super brother. - *It was very nice. Super brother.* Inter-sentential switch between clauses.

3 Related work

A 3-parallel Long Short Term Memory (LSTM) architecture which takes random, Word2Vec and random character embeddings to categorize sentiment of Dravidian code-mixed YouTube comments was implemented by Mishra et al. (2021). They obtained 15^{th}, 12^{th}, and 12^{th} positions respectively for Tamil, Malayalam, and Kannada datasets. SR et al. (2022) experimented with Kernel based Extreme Learning Machines (ELM) for sentiment analysis. They concluded that Polynomial kernels works best in the ELM architecture for Code-Switched Dravidian Languages.
A deep learning-based framework using Bidirectional-LSTM (Bi-LSTM) was proposed by Roy and Kumar (2021) to extract the features from input sentences and classify the sentiment in code-mixed languages. Shanmugavadivel et al. (2022) proposed sentiment analysis and offensive language identification on multilingual code-mixed data. They worked on extraction of semantically meaningful information from code-mixed data using word embedding for sentiment classification. Yadav and Chakraborty (2021) proposed an unsupervised method for Sentiment Analysis of Code-mixed Data. They employed multilingual and cross-lingual embeddings of monolingual text to find the sentiment in code-mixed text. Clustering of Tamil text with and with-out considering class-wise information for sentiment analysis was proposed by Thavareesan and Mahesan (2021). They tested by varying number of centroids in k-means clustering and k-nearest neighbour classifier. They achieved better results using FastText embeddings when compared to Bag Of Words (BOW) embeddings.

4 Methodology

We received a total of 27 submissions for both Tamil and Tulu. The systems were evaluated based on macro average F1 scores and a rank list was prepared. Table 1 and Table 2 show the rank lists of code-mixed Tamil and Tulu texts respectively. We briefly describe below the methodologies used by the top five teams.

- DeepBlueAI (Luo and Wang, 2023): The authors fine-tuned XLM-RoBERTa, a pre-trained multilingual language model, as the base model and they also combined various language datasets in different proportions.

Team name	Macro F1 score	Rank
DeepBlueAI (Luo and Wang, 2023)	0.32	1
MUNLP (G et al., 2023)	0.269	2
Poorvi (Shetty, 2023)	0.268	3
AK-NLP	0.242	4
AlphaBrains (Ehsan et al., 2023)	0.215	5
Habesha (Yigezu et al., 2023)	0.208	6
lidoma (Tash et al., 2023)	0.199	7
MUCSD (Coelho et al., 2023)	0.19	8
TU-PM	0.18	9
selam (Kanta, 2023)	0.147	10
AbhiPaw	0.143	11
Team-Tamil (Ponnusamy et al., 2023)	0.142	12
MUCS (Kulal et al., 2023)	0.141	13
ML&AI_IIITRanchi (Kumari et al., 2023)	0.124	14
Muhammad	0.12	15

Table 1: Rank list based on macro average F1 score for code-mixed Tamil text

Team name	Macro F1 score	Rank
DeepBlueAI (Luo and Wang, 2023)	0.542	1
MUNLP (G et al., 2023)	0.533	2
TU-PM	0.529	3
AK-NLP	0.523	4
AlphaBrains (Ehsan et al., 2023)	0.522	5
selam (Kanta, 2023)	0.518	6
lidoma (Tash et al., 2023)	0.516	7
Muhammad	0.514	8
MUCSD (Coelho et al., 2023)	0.508	9
L&AI_IIITRanc (Kumari et al., 2023)	0.471	10
AbhiPaw	0.442	11
Team-Tamil (Ponnusamy et al., 2023)	0.442	12
Habeesha (Yigezu et al., 2023)	0.352	13
Poorvi (Shetty, 2023)	0.268	14
MUCS (Kulal et al., 2023)	0.204	15

Table 2: Rank list based on macro average F1 score for code-mixed Tulu text

They also used cross-validation to assess the performance and generalization of the fine-tuned model across multiple languages. They achieved the best macro F1 scores for both Tamil and Tulu code-mixed text securing the 1^{st} place in the shared task.

- MUNLP (G et al., 2023): This team used two distinct approaches for Tamil and Tulu code-mixed texts. They fine-tuned Tamil sentiment BERT with the oversampled train data for Tamil. Whereas, for Tulu, they combined FastText pre-trained word embeddings with Tulu byte pair embeddings to train LinearSVC model. Their system achieved Rank 2 for both Tamil and Tulu.
- Poorvi (Shetty, 2023): Authors trained ensemble of ML (LinearSVC, Random Forest, MultinomialNB, and Logistic Regression) classifiers considering TF-IDF of n-grams in the range n = (1, 2). Their model secured 3^{rd} and 14^{th} ranks for Tamil and Tulu respectively.
- AK-NLP: This team trained LSTM model fed with TF-IDF features for both code-mixed Tamil and Tulu texts. They achieved Rank 4 for both Tamil and Tulu.
- TU-PM: Authors fine-tuned XLM-RoBERTa-base for both code-mixed Tamil and Tulu. Their proposed models achieved 3^{rd} and 9^{th} ranks for Tulu and Tamil code-mixed texts.
- AlphaBrains (Ehsan et al., 2023): Using trainsliteration, this team augmented the existing Tamil and Tulu text before training their proposed models. They used TL by training character based contextualized ELMO representations for both Tamil and Tulu. Their proposed models secured 5^{th} rank for both Tamil and Tulu code-mixed texts.

5 Evaluation

The distribution of the sentiment classes are imbalanced in both the datasets. In the Tamil code-mixed dataset, we have a class imbalance with the majority of comments belonging to Positive (20,070) class. Similarly, the Tulu code-mixed dataset has class imbalance with Positive (3,118) and Neutral (1,719) being the majority classes.

To address class imbalance, we utilized the macro average F1 score for ranking the systems. Macro F1 scores are often used in evaluating models trained on imbalanced data because they provide a balanced assessment of model performance across all classes, regardless of class distribution. In imbalanced datasets, accuracy alone can be misleading, as a model may achieve high accuracy by simply predicting the majority class. By using the macro average F1 score, equal importance is given to each class, making it a suitable metric to evaluate model performance in scenarios where class imbalances exist. This score is computed by averaging the F1 scores of each class in the multi-class classification problem. To facilitate this calculation, we leveraged the classification report tool from Scikit-learn[1], which provided comprehensive metrics and insights for evaluating the performance of the systems.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

6 Results and Discussion

The sentiment analysis shared task focused on two languages: code-mixed Tamil and Tulu. A total of 57 participants registered for the shared task, out of which 27 teams submitted their systems for both Tamil and Tulu language tracks. Table 1 and Table 2 present the rank lists for Tamil and Tulu languages respectively, showcasing the performance of the submitted systems. It is worth noting that the majority of submissions were designed to handle SA for both languages, as indicated earlier. In this section, we present the top-ranked results for both the languages based on the macro average F1 scores. The rankings reflect the performance of the systems on the dataset, with the top positions indicating the highest macro average F1 scores across all classes.

Teams have commonly employed transformer-based models like XLM-RoBERTa, Tamil sentiment BERT, DeBERTa-Large, MuRIL, and IndicBERT, even though these models were not originally pre-trained on code-mixed text. In addition, some teams have also experimented with pre-trained word embeddings like FastText and BPEmb.

[1]Macro F1 score

Using these linguistic representations, a range of machine learning (ML) models including SVM, k-NN, MLP, OneVsRest, XGB, and GradientBoosting, as well as deep learning (DL) models such as CNN and LSTM, have been explored. These models are often combined with TF-IDF representations of word or character n-grams to effectively address the challenges of code-mixed text processing.

Participants encountered challenges in working with code-mixed text due to the inclusion of non-native script in our corpus. To address this, they acquired pre-trained models from libraries and fine-tuned them for our corpora. However, the availability of resources for Tulu code-mixed text is limited in comparison to Tamil. This scarcity prompted participants to take initiatives in resource generation and training pre-trained models from scratch. Given the data imbalance in the dataset, participants opted to employ resampling techniques like SMOTE to mitigate this imbalance effectively.

Despite efforts, both LSTM and traditional ML algorithms fell short in delivering satisfactory results when contrasted with transformer-based models. Notably, among the various models tested, XLM-RoBERTa and the transformer-based architecture demonstrated the most promising performance. Even though several systems and approaches had F1 scores that were below the average, we accepted those articles in an effort to promote the use of a variety of research techniques to address the issue in Dravidian languages. Most of the submissions in working notes provided comprehensive insights into class-specific precision, recall, and F1 scores. The primary evaluation metric utilized was the weighted F1 score, enabling a comprehensive assessment of model effectiveness.

7 Conclusion

We present the results of the SA shared task on code-mixed Tamil and Tulu text. The dataset utilized in the shared task consisted of code-mixed instances sourced from social media, particularly YouTube comments. To tackle the SA challenge, a majority of the participants employed fine-tuning techniques on pre-trained multilingual language models. This approach allowed them to leverage the existing knowledge captured by the pre-trained models while adapting them to the specific code-mixed Dravidian language context. The top-performing systems in the shared tasks employed additional resources such as multilingual sentiment analysis datasets and fine-tuned pre-trained models. However, despite their success, the results also suggest that there is still potential for improvement in sentiment analysis for all three Dravidian languages: Tamil, Malayalam, and Kannada. The increased number of participants and the improved performance of the systems indicate a growing interest in the field of Dravidian NLP and a positive trend towards advancing research in this domain.

Acknowledgments

The author Bharathi Raja Chakravarthi was supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2).

References

PJ Antony, Hemant B Raj, BS Sahana, Dimple Sonal Alvares, and Aishwarya Raj. 2012. Morphological Analyzer and Generator for Tulu Language: A Novel Approach. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pages 828–834.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "I am Borrowing Ya Mixing?" An Analysis of English-Hindi Code Mixing in Facebook. In *Proceedings of the first workshop on computational approaches to code switching*, pages 116–126.

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

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2020a. A Sentiment Analysis Dataset for Code-mixed Malayalam-English. In *arXiv preprint arXiv:2006.00210*.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. 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, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022. Dravidiancodemix: Sentiment Analysis and Offensive Language Identification Dataset for Dravidian Languages in Code-mixed Text. In *Language*

Resources and Evaluation, volume 56, pages 765–806. Springer.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020c. Overview of the Track on Sentiment Analysis for Dravidian Languages in Code-mixed Text. In *Forum for information retrieval evaluation*, pages 21–24.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, Elizabeth Sherly, John P McCrae, Adeep Hande, Rahul Ponnusamy, Shubhanker Banerjee, et al. 2021. Findings of the Sentiment analysis of Dravidian Languages in Code-mixed Text. In *arXiv preprint arXiv:2111.09811*.

Sharal Coelho, Asha Hegde, Pooja Lamani, Kavya G, and Hosahalli Lakshmaiah Shashirekha. 2023. MUCSD@DravidianLangTech2023: Predicting Sentiment in Social Media Text using Machine Learning Techniques. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Toqeer Ehsan, Amina Tehseen, Kengatharaiyer Sarveswaran, and Amjad Ali. 2023. AlphaBrains@DravidianLangTech: Sentiment Analysis of Code-Mixed Tamil and Tulu by Training Contextualized ELMo Word Representations. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Kavya G, Asha Hegde, Sharal Coelho, Pooja Lamani, and Hosahalli Lakshmaiah Shashirekha. 2023. MUNLP@DravidianLangTech2023: Learning Approaches for Sentiment Analysis in Code-mixed Tamil and Tulu Text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

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, Marseille, France. European Language Resources Association.

Asha Hegde, Shubhanker Banerjee, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Hosahalli Shashirekha, John Philip McCrae, et al. 2022b. Overview of the Shared Task on Machine Translation in Dravidian Languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 271–278.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2022. Leveraging Dynamic Meta Embedding for Sentiment Analysis and Detection of Homophobic/Transphobic Content in Code-mixed Dravidian Languages.

Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Anand Kumar Madasamy, and Bharathi Raja Chakravarthi. 2022c. A Study of Machine Translation Models for Kannada-Tulu. In *Congress on Intelligent Systems*, pages 145–161. Springer.

S Hewavitharana and HC Fernando. 2002. A Two Stage Classification Approach to Tamil Handwriting Recognition. In *Tamil Internet*, volume 2002, pages 118–124.

Braj B Kachru. 1978. Toward Structuring Code-mixing: An Indian Perspective. Walter de Gruyter, Berlin/New York Berlin, New York.

Selam Kanta. 2023. Sentiment Analysis of Code-Mixed Dravidian Texts Using SVM Classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Rachana Kulal, Prajnashree M, Asha Hegde, and Shashirekha H. L. 2023. MUCS@DravidianLangTech2023: Sentiment Analysis in Code-mixed Tamil and Tulu Texts using fastText. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Kirti Kumari, Shirish Shekhar Jha, Zarikunte Kunal Dayanand, and Praneesh Sharma. 2023. MLAI_IIITRanchi@DravidianLangTech: Fine-Tuning IndicBERT for Exploring Language-specific Features for Sentiment Classification in Code-Mixed Dravadian Language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Zhipeng Luo and Jiahui Wang. 2023. Deepblueai@dravidianlangtech. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Ankit Kumar Mishra, Sunil Saumya, and Abhinav Kumar. 2021. Sentiment Analysis of Dravidian-CodeMix Language. In *Forum for information retrieval evaluation*, pages 1011–1019.

K Padmanabha Kekunnaya. A comparative study of tulu dialects.

Charmathi Ponnusamy, Kishore Kumar andRajkumar, Elizabeth Kumaresan, Prasanna Kumar andSherly, and Ruba Priyadharshini. 2023. VEL@DravidianLangTech: Sentiment Analysis of Tamil and Tulu. In *Proceedings of the Third*

Workshop on Speech and Language Technologies for Dravidian Languages, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Pradeep Kumar Roy and Abhinav Kumar. 2021. Sentiment Analysis on Tamil Code-Mixed Text using Bi-LSTM. In *Forum for information retrieval evaluation*, pages 1044–1050.

K. Shanmugavadivel, V.E. Sathishkumar, and S. Raja. 2022. Deep Learning Based Sentiment Analysis and Offensive Language Identification on Multilingual Code-mixed Data. volume 12, page 21557. nature.

Poorvi Shetty. 2023. Sentiment Analysis on Code-Mixed Tulu and Tamil Corpus. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

R Sivanantham and M Seran. 2019. Keeladi: An Urban Settlement of Sangam Age on the Banks of River Vaigai. In *India: Department of Archaeology, Government of Tamil Nadu, Chennai*.

Mithun Kumar SR, Lov Kumar, and Aruna Malapati. 2022. Sentiment Analysis on Code-Switched Dravidian Languages with Kernel Based Extreme Learning Machines. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages.*, pages 184–190. Association for Computational Linguistics.

R Srinivasan and CN Subalalitha. 2019. Automated Named Entity Recognition from Tamil Documents. In *2019 IEEE 1st international conference on energy, systems and information processing (ICESIP)*, pages 1–5. IEEE.

CN Subalalitha. 2019. Information Extraction Framework for Kurunthogai. In *Sādhanā*, volume 44, page 156. Springer.

Moein Tash, Jesus Armenta-Segura, zahra ahani, Grigori Kolesnikova, Olga andSidorov, and Gelbukh Alexander. 2023. LIDOMA@DravidianLangTech: Convolutional Neural Networks for Studying Correlation Between Lexical Features and Sentiment Polarity in Tamil and Tulu Languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

S. Thavareesan and S. Mahesan. 2021. Sentiment Analysis in Tamil Texts using k-means and k-Nearest Neighbour. In *Proceedings of 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48– 53. IEEE.

S. Yadav and T. Chakraborty. 2021. Unsupervised Sentiment Analysis for Code-mixed Data. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 15941–15942.

Tadesse Yigezu, Mesay Gemeda andKebede, Olga Kolesnikova, Grigori Sidorov, and Gelbukh Alexander. 2023. Habesha@DravidianLangTech: Utilizing Deep and Transfer Learning Approaches for Sentiment Analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.