Overview of Second Shared Task on Sentiment Analysis in Code-mixed Tamil and Tulu

Lavanya Sambath Kumar[1] **, Asha Hegde**[2] **, Bharathi Raja Chakravarthi**[3]**,**
Hosahalli Lakshmaiah Shashirekha[2], **Rajeshwari Natarajan**[4],
Sajeetha Thavareesan[5], **Ratnasingam Sakuntharaj**[5], **Durairaj Thenmozhi**[6],
Prasanna Kumar Kumaresan[7], **Charumathi Rajakumar**[8]
[1] Anna University, Tamil Nadu, India, [2] Mangalore University, Mangalore, India
[3] School of Computer Science, University of Galway, Ireland
[4] SASTRA University, Tamil Nadu, India, [5] Eastern University, Sri Lanka
[6] Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
[7] Data Science Institute, University of Galway, Ireland
[8] The American College, Madurai, Tamil Nadu, India
bharathiraja.akr@gmail.com

Abstract

Sentiment Analysis (SA) in Dravidian code-mixed text is a hot research area right now. In this regard, the "Second Shared Task on SA in Code-mixed Tamil and Tulu" at DravidianLangTech (EACL-2024) is organized. Two tasks namely SA in Tamil-English and Tulu-English code-mixed data, make up this shared assignment. In total, 64 teams registered for the shared task, out of which 19 and 17 systems were received for Tamil and Tulu, respectively. The performance of the systems submitted by the participants was evaluated based on the macro F1-score. The best method obtained macro F1-scores of 0.260 and 0.584 for code-mixed Tamil and Tulu texts, respectively.

1 Introduction

Sentiment Analysis (SA) employs machine learning and artificial intelligence techniques to extract and classify the sentiments and emotions about a person (Raj et al., 2024), service, event or a topic that individuals may be hiding behind a text. SA finds its applications in a wide range of fields, such as e-commerce, healthcare, banking, politics, and others. SA is challenging as sentiments vary based on context, emojis and usage of irony and sarcasm in casual conversations (Hegde et al., 2023b). Digital revolution paved the way for the native speakers of Dravidian languages to express their opinion in social media platforms which has lead to a great deal of attention for SA in Dravidian languages (Chakravarthi et al., 2020c). It is challenging due to language complexity and availability of low resources. Code-mixing adds much more complexity in analysing the sentiments as people have the tendency of using non-native scripts in place various historically used scripts (Chakravarthi et al., 2021b; Hegde and Shashirekha, 2022). In such cases, SA models trained on monolingual data may not suit for code-mixed data because of complex linguistic patterns (Chakravarthi et al., 2022, 2021a).

Tamil- holds the distinction of being one of India's oldest surviving classical languages (Vasantharajan et al., 2022; Hegde et al., 2022b). Recognized as a scheduled language under the Indian constitution, it serves as the official language in Tamil Nadu and Puducherry. Beyond India, Tamil is acknowledged as a national language in both Singapore and Sri Lanka. With a significant presence, it is spoken by a sizable minority in additional south Indian states, including Kerala, Karnataka, Andhra Pradesh, Telangana, and the Union Territory of Andaman and Nicobar Islands (Bharathi et al., 2023). The archaeological evidence of the first Tamil script, dating back to 580 BCE, was discovered on pottery in the Keezhadi, Sivagangai, and Madurai districts of Tamil Nadu by the Tamil Nadu State Department of Archaeology and the Archaeological Survey of India (Sivanantham and Seran, 2019).

Tulu - is a prominent Dravidian language, spoken by around 2.5 million people, primarily in the Dakshina Kannada and Udupi districts of Karnataka, along with certain areas in Kasaragod, Kerala. The language preserves key features of ancient Dravidian languages, offering a glimpse into linguistic traditions while also introducing unique innovations not observed in other Dravidian languages (Padmanabha Kekunnaya). Tulu stands as a testament to the linguistic diversity and rich cultural tapestry present in the Indian subcontinent (Antony et al., 2012; Hegde et al., 2022c).

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 62–70
March 22, 2024 ©2024 Association for Computational Linguistics

Team name	Run name	Precision	Recall	Macro F1-score	Rank
MUCS (B et al., 2024)	Run2	0.291	0.279	0.260	1
CUETSentimentSillies (Tripty et al., 2024)	Run1	0.288	0.270	0.258	2
CUET_Binary_Hackers (Eusha et al., 2024)	Run3	0.279	0.268	0.227	3
CEN-Amrita	Run1	0.250	0.259	0.220	4
Transformers (Singhal and Bedi, 2024)	Run1	0.245	0.279	0.212	5
KEC_DL_KSK	Run3	0.278	0.263	0.197	6
Habesha	Run2	0.299	0.253	0.171	7
KEC_AI_CODE_MAKER (Shanmugavadivel et al., 2024a)	Run1	0.284	0.258	0.170	8
bytesizedllm	Run1	0.277	0.245	0.157	9
kubapok	Run2	0.144	0.131	0.122	10
wordwizards (Balaji et al., 2024)	Run1	0.213	0.243	0.074	11
Fango	Run1	0.075	0.162	0.060	12
InnovationEngineers (Shanmugavadivel et al., 2024b)	Run1	0.102	0.165	0.035	13

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

2 Task Description

Tamil and Tulu are low-resource Dravidian languages, where limited resources are available specifically for SA. The primary objective of the proposed shared task[1] is to find the sentiment polarity in gold standard code-mixed Tamil-English (Chakravarthi et al., 2020b) and Tulu-Kannada-English (Hegde et al., 2022a) datasets. These code-mixed datasets consist of posts and comments gathered from YouTube comments. Class imbalance issues are also present in this dataset, which represents a real-world situation (Hegde et al., 2023a). The secondary objective is to support studies that will shed light on the expression of sentiment in code-mixed scenarios. This task involves classifying polarity of YouTube comments into four categories: positive, negative, neutral, or mixed emotions.

3 Dataset

Due to the widespread expansion of digital content on social media platforms like YouTube, Twitter, and Instagram, there has been a substantial increase in SA of social media text in even low-resource languages, including Kannada, Tamil, Telugu, and Malayalam (Hande et al., 2020; Chakravarthi et al., 2020a). Despite the development of numerous SA models, the evolving nature of user-generated content, which is becoming more diverse and creative, underscores the need for more efficient tools (Hegde et al., 2023c; Rachana et al., 2023). Focusing on YouTube comments for SA, two gold standard corpora: i) Tamil-English and ii) Tulu-Kannada-English are made available to the research community through this shared task. The corpora, acting as a crucial repository, offer substantial support for researchers and practitioners engaged in SA within multilingual contexts. Specifically, they enable the development and evaluation of models adept at processing code-mixed data present in Tamil and Tulu texts derived from YouTube comments. The statistics of these corpora are given in Table 2.

4 Related work

Recent advancements in SA on social media have witnessed notable progress, particularly in the realm of the internet (Wankhade et al., 2022). These studies have extended their focus to encompass not only high-resource languages but also low-resource languages, reflecting a growing recognition of the importance of linguistic diversity in understanding and analyzing user sentiments across various social media platforms (Chakravarthi et al., 2022; Priyadharshini et al., 2021). This inclusive approach enhances the applicability and effectiveness of sentiment analysis techniques in capturing the nuances of expression in a wide array of languages. With this view, Chakravarthi et al. (2020c) and B et al. (2022) organized shared tasks in code-mixed Dravidian languages to promote SA under low-resource scenarios.

SR et al. (2022) presented kernel-based extreme learning for SA in code-mixed Dravidian languages (Tamil, Kannada, and Malayalam). Their focus was

[1] `https://codalab.lisn.upsaclay.fr/competitions/16088`

Label	Train Set		Development Set		Test set	
	Tamil	Tulu	Tamil	Tulu	Tamil	Tulu
Positive	20,070	3,118	2,257	369	73	248
Neutral	5,628	1,719	611	202	137	140
Mixed Feeling	4,020	974	480	120	101	70
Negative	4,271	646	438	90	338	43

Table 2: Class-wise distribution of code-mixed Tamil and Tulu texts

more on handling data imbalance issues and extracting more relevant features employing feature selection techniques. A Deep Learning (DL) technique for SA on code-mixed Malayalam text using Hierarchical Attention Network (HAN) model was proposed by Pillai and Arun (2024). ALBERT[2] tokenization was executed and key features specifically Term Frequency-Inverse Document Frequency (TF-IDF) based and n-grams features were extracted in the feature extraction step. Feature fusion was applied to the retrieved features using HAN and Shannon entropy. Finally, sentiment in the comments was categorized into positive or negative class. They concluded that the Feature fusion+HAN technique achieved better results for Malayalam code-mixed data. SA on Tamil code-mixed data using a variety of cutting-edge learning and hybrid deep learning algorithms was implemented by Shanmugavadivel et al. (2022). Various pre-processing procedures, such as emojis removal, repeated characters removal, punctuation, symbols, and number removal were employed to clean the data set. TF-IDF technique was used for feature extraction. The authors made a claim stating that the creation of hybrid DL models by merging Convolutional Neural Network (CNN) + Long Short Term Memory (LSTM), LSTM+CNN, CNN+Bidirectional LSTM (BiLSTM), and BiLSTM+CNN makes their study effort novel. It was found that the hybrid DL model CNN+BiLSTM was effective at SA on data with mixed Tamil and English codes.

Late-off transformer models have gained tremendous attention from the scholarly community, specifically for SA in Dravidian languages (Elankath and Ramamirtham, 2023). The existing research on SA in code-mixed Dravidian languages has identified key trends and emphasized the need for further exploration (Saini and Roy, 2023). While some studies have utilized good-quality datasets and models, the literature underscores the substantial gaps that persist in understanding hidden sentiments within these languages (Hande et al., 2022). Despite the progress made, a significant research opportunity exists to delve deeper into the complexities of sentiment analysis in code-mixed Dravidian languages and contribute to the advancement of this field.

5 Methodology

We received a total of 19 submissions for Tamil and 17 submissions for Tulu. The systems were evaluated based on macro average F1 scores and rank lists were prepared. Table 1 and Table 3 show the rank lists of code-mixed Tamil and Tulu texts respectively. We briefly describe below the methodologies used by the top five teams.

- MUCS (B et al., 2024): The authors implemented SVM and an ensemble of three Machine Learning (ML) classifiers (Support Vector Model (SVM), Random Forest (RF), and k Nearest Neighbors (kNN)) for SA in code-mixed Tamil and Tulu text. They also employed Gridsearch algorithm to get the optimal hyperparameters of the classifiers. Their proposed model obtained macro F1 scores of 0.260 and 0.584 for code-mixed Tamil and Tulu texts securing 1st and 2nd ranks in the shared task.

- CUET_Binary_Hackers (Eusha et al., 2024): The authors fine-tuned indicbert[3] (Kakwani et al., 2020) and indic-sentence-bert-nli[4] (Deode et al., 2023) models for Tamil language and bert-base-multilingual-cased[5] (Devlin et al., 2018) (mBERT) model for Tulu language. In addition, they ensembled ML classifiers with majority voting for SA in Tamil

[2] https://huggingface.co/docs/transformers/en/model_doc/albert

[3] https://huggingface.co/ai4bharat/indic-bert

[4] https://huggingface.co/l3cube-pune/indic-sentence-bert-nli

[5] https://huggingface.co/bert-base-multilingual-cased

Team name	Run name	Precision	Recall	Macro F1-score	Rank
CUET_Binary_Hackers (Eusha et al., 2024)	Run1	0.590	0.580	0.584	1
kubapok	Run1	0.617	0.57	0.584	1
MUCS (B et al., 2024)	Run2	0.548	0.554	0.550	2
Habesha	Run1	0.502	0.531	0.504	3
CEN-Amrita	Run1	0.488	0.489	0.477	4
CUETSentimentSillies (Tripty et al., 2024)	Run1	0.512	0.468	0.468	5
KEC_DL_KSK	Run2	0.485	0.446	0.443	6
Fango	Run1	0.316	0.404	0.344	7
wordwizards_tulu (Balaji et al., 2024)	Run1	0.296	0.270	0.251	8
Transformers (Singhal and Bedi, 2024)	Run1	0.222	0.251	0.221	9

Table 3: Rank list based on macro average F1 score for code-mixed Tulu text

and Tulu languages. Their proposed ensemble model obtained a macro F1 score of 0.227 securing 3rd rank in the shared task for Tamil language. Further, their fine-tuned mBERT model obtained a macro F1 score of 0.584 securing 1st rank in the shared task for Tulu language.

- CUETSentimentSillies (Tripty et al., 2024): The authors have resampled the Tamil dataset before pre-processing and fine-tuned xlm-roberta-base-language-detection[6] - a pre-trained model to fins SA in code-mixed Tamil text (Conneau et al., 2019). Whereas, for Tulu, they fine-tuned bert-base-multilingual-uncased-sentiment[7] model for SA in code-mixed Tulu text. Their proposed models obtained macro F1 scores of 0.258 and 0.468 securing 2nd and 5th ranks in the shared task for code-mixed Tamil and Tulu texts respectively.

- CEN-Amrita: The team employs a combination of CNN in feature extraction and BiLSTM layers to capture contextual information for SA in Tamil and Tulu languages. Their proposed model obtained macro F1 scores of 0.220 and 0.477 for Tamil and Tulu languages respectively securing 4th rank in the shared task for both the languages.

- kubapok: The authors fine-tuned five BERT variants: twhin-bert-large[8] (Zhang et al., 2023), muril-base-cased[9] (MuRIL) (Khanuja et al., 2021), mDeBERTa V3 base[10] (He et al., 2021), xlm-roberta-large[11] (Conneau et al., 2019), and xlm-roberta-large for SA in code-mixed Tamil and Tulu texts. They averaged all the probabilities of the five models while taking considering the prediction. Their proposed methodology achieved macro F1 scores of 0.122 and 0.584 securing 10th and 1st ranks in the shared task for code-mixed Tamil and Tulu texts respectively.

- Transformers (Singhal and Bedi, 2024): The authors have carried out minority undersampling for both the datasets as the provided Tamil and Tulu datasets are imbalanced. They separately fine-tuned XLM Roberta (Conneau et al., 2019) for Tamil and Tulu languages. Their proposed models obtained the macro F1 scores of 0.212 and 0.221 securing 5th and 9th ranks in the shared task for code-mixed Tamil and Tulu texts respectively.

- Habesha: The team employs two distinct models: i) Run1 - a model based on transformers for embedding, coupled with DL techniques for the purpose of classification and ii) Run 2 - a fine-tuned DistilBERT[12] (Sanh et al., 2019) model for SA in Tamil and Tulu languages. Run 2 obtained macro F1 score of 0.171 for code-mixed Tamil text. Further, Run 1 obtained a macro F1 score of 0.504 for code-

[6] https://huggingface.co/papluca/xlm-roberta-base-language-detection

[7] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[8] https://huggingface.co/Twitter/twhin-bert-large

[9] https://huggingface.co/google/muril-base-cased

[10] https://huggingface.co/microsoft/mdeberta-v3-base

[11] https://huggingface.co/FacebookAI/xlm-roberta-large

[12] https://huggingface.co/docs/transformers/en/model_doc/distilbert

mixed Tulu text. This team secured 7th and 5th ranks in the shared task for Tamil and Tulu languages.

6 Evaluation

The sentiment class distribution is imbalanced in datasets, particularly in the Tamil code-mixed dataset, where the majority of comments are categorized as Positive sentiment (20,070). Addressing this imbalance is crucial for developing a robust SA model that can effectively capture nuanced patterns across different sentiment classes. Similarly, the Tulu code-mixed dataset has class imbalance with Positive (3,118) and Neutral (1,719) being the majority classes. In addressing class imbalance, the utilization of the macro F1 score serves as a robust metric for ranking systems. This approach is beneficial when evaluating models trained on imbalanced datasets, as it offers a balanced assessment of performance across all classes, irrespective of their distribution. Unlike accuracy, which can be misleading in imbalanced scenarios where a model may excel by predominantly predicting the majority class, the macro F1 score assigns equal importance to each class. This makes it a reliable metric for evaluating model performance in situations characterized by class imbalances, as it takes into account both precision and recall for each class. The computation of the macro F1 score involves averaging the F1 scores of individual classes in a multi-class classification problem, providing a comprehensive and fair evaluation of the model's effectiveness. To facilitate this calculation, we leveraged the classification report tool from Scikit-learn[13], which provided comprehensive metrics and insights for evaluating the performance of the systems.

7 Results and Discussion

There are 64 participants in the SA shared task, which focused on two languages: code-mixed Tamil and Tulu text. Among them, 19 for Tamil and 17 teams for Tulu actively engaged in the challenge, submitting their systems for Tamil and Tulu language tracks. The resulting rank lists, depicted in Tables 1 and 3 for Tamil and Tulu languages respectively, outline the performance of these systems. Notably, a majority of the submissions were adept at handling SA for both languages concurrently. This section unveils the top-ranked outcomes for both languages, emphasizing macro F1 scores as the evaluation metric. The rankings signify the systems' proficiency on the dataset, with higher positions denoting superior macro F1 scores across all classes, thereby providing a comprehensive assessment of their performance.

Teams that participated in the SA shared task have frequently employed transformer models such as XLM-RoBERTa, mBERT, DeBERTa-Large[14], MuRIL, and IndicBERT[15], despite these models not being originally pre-trained on code-mixed text. The application of these linguistic representations has been diverse, with teams exploring a variety of ML models, including SVM, kNN, Multi Layer Perceptron, Linear Support Vector Classifier, and RF. Furthermore, DL models such as CNN and BiLSTM have been extensively experimented with. To effectively address the challenges associated with processing code-mixed text, these models are combined with TF-IDF of word and character n-grams, showcasing a holistic approach to SA on multilingual and code-mixed data. One team (MUCS (B et al., 2024)) in the SA shared task has distinguished itself by employing a grid-search algorithm to meticulously determine optimal hyperparameter values for their ML algorithms. Going beyond individual models, this team has implemented a sophisticated ensemble approach, utilizing majority voting across three distinct ML classifiers. This ensemble strategy, combined with the fine-tuned hyperparameter values, underscores the team's commitment to maximizing performance and robustly addressing the challenges of SA in code-mixed text.

Participants in the SA shared task faced challenges when working with code-mixed text, particularly due to the inclusion of non-native scripts in the corpus. In response, they addressed this issue by acquiring pre-trained models from libraries and fine-tuning them to better suit their corpora. Notably, the limited availability of resources for Tulu code-mixed text, in contrast to Tamil, prompted participants to take proactive measures, including resource generation and training pre-trained models from scratch. Acknowledging the data imbalance within the dataset, participants strategically employed the undersampling technique to effectively mitigate the imbalance and enhance the per-

[13] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[14] https://huggingface.co/microsoft/deberta-large

[15] https://huggingface.co/ai4bharat/IndicBERTv2-MLM-only

formance of their SA models. Despite earnest efforts, both BiLSTM models and traditional ML algorithms fell short of delivering satisfactory results when compared to the performance of transformer-based models. Notably, among the diverse models tested, mBERT and other transformer-based architectures demonstrated the most promising and superior performance. Further, Gridsearch algorithm was found to be more beneficial in obtaining better performance.

The participated teams in the competition have highlighted a concerning trend of persistently low F1 scores, indicating a notable challenge in achieving robust SA. The likely reason for this issue is the significant data imbalance present in the provided datasets, a factor that has not been adequately addressed in the system descriptions provided by participating teams. Moreover, a predominant reliance on feature extraction and classifier construction methods by most participants, instead of incorporating feature selection techniques, may contribute to the suboptimal performance.

8 Conclusion

In presenting the results of the SA shared task on code-mixed Tamil and Tulu text, the dataset used was comprised of code-mixed instances sourced from social media, notably YouTube comments. A prevalent strategy among the participants involved the application of fine-tuning techniques to pre-train multilingual language models to address the SA challenge. By adopting this approach, participants were able to capitalize on the pre-existing knowledge embedded in the multilingual models while tailoring them to the context of given code-mixed Dravidian languages. This method proved effective in leveraging the strengths of pre-trained models for improved SA performance in the specific domain of social media discourse.

The top-performing systems in the shared tasks demonstrated success through the incorporation of Gridsearch algorithm, voting classifiers, and fine-tuned pre-trained models. Despite their achievements, the results underscore the existing potential for further enhancement in SA across Tamil and Tulu languages. The increased participation and improved performance of the systems signal a growing interest in the field of Dravidian Natural Language Processing (NLP) and indicate a positive trend toward advancing research within this domain. The ongoing efforts and achievements in the shared tasks suggest a promising trajectory for continued development and refinement of SA techniques for these languages.

Acknowledgement

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223, 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.

Prathvi B, Manavi K K, Subrahmanya, Asha Hegde, Kavya G, and H L Shashirekha. 2024. MUCS@DravidianLangTech-2024: A Grid Search Approach to Explore Sentiment Analysis in Code-mixed Tamil and Tulu . In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Premjith B, Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B, Soman Kp, Dhanalakshmi V, Sreelakshmi K, Arunaggiri Pandian, and Prasanna Kumaresan. 2022. Findings of the Shared Task on Multimodal Sentiment Analysis and Troll Meme Classification in Dravidian Languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 254–260, Dublin, Ireland. Association for Computational Linguistics.

Shreedevi Seluka Balaji, Akshatha Anbalagan, Niranjana A, Priyadharshini T, and Durairaj Thenmozhi. 2024. WordWizards@DravidianLangTech 2024: Sentiment Analysis in Tamil and Tulu using Sentence Embedding. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya Natarajan, Rajeswari Natarajan, S Suhasini, and Swetha Valli. 2023. Overview of the Second Shared Task on Tpeech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 31–37.

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

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P 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)*.

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, and Charangan Vasantharajan. 2021a. Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text. In *CEUR Workshop Proceedings*.

BR Chakravarthi, PK Kumaresan, R Sakuntharaj, AK Madasamy, S Thavareesan, S Chinnaudayar Navaneethakrishnan, and T Mandl. 2021b. Overview of the HASOC-DravidianCodeMix shared task on offensive language detection in Tamil and Malayalam. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation*. CEUR.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3Cube-IndicSBERT: A Simple Approach for Learning Cross-lingual Sentence Representations using Multilingual BERT. *arXiv preprint arXiv:2304.11434*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Syam Mohan Elankath and Sunitha Ramamirtham. 2023. Sentiment Analysis of Malayalam Tweets using Bidirectional Encoder Representations from Transformers: A Study. *Indonesian Journal of Electrical Engineering and Computer Science*, 29(3):1817–1826.

Asrarul Hoque Eusha, Salman Farsi, Ariful Islam, Avishek Das, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech-EACL 2024: Sentiment Analysis using Transformer-Based Models in Code-Mixed and Transliterated Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Adeep Hande, Siddhanth U Hegde, and Bharathi Raja Chakravarthi. 2022. Multi-task learning in under-resourced Dravidian Languages. *Journal of Data, Information and Management*, 4(2):137–165.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for Sentiment Analysis and Offensive Language Letection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

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, 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, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya Sambath Kumar, Durairaj Thenmozhi, Martha Karunakar, Shreya Sriram, and Sarah Aymen. 2023a. Findings of the Shared Task on Sentiment Analysis in Tamil and Tulu Code-Mixed Text. In *Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, SK Lavanya, Durairaj Thenmozhi, Martha

Karunakar, Shreya Shreeram, and Sarah Aymen. 2023b. Findings of the Shared task on Sentiment Analysis in Tamil and Tulu Code-mixed Text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71.

Asha Hegde, G Kavya, Sharal Coelho, Pooja Lamani, and Hosahalli Lakshmaiah Shashirekha. 2023c. 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*, pages 275–281.

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.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages.

K Padmanabha Kekunnaya. A comparative study of Tulu dialects.

Aditya R Pillai and Biri Arun. 2024. A Feature Fusion and Detection Approach using Deep Learning for Sentimental Analysis and Offensive Text Detection from Code-mix Malayalam Language. *Biomedical Signal Processing and Control*, 89:105763.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2021. Overview of the DravidianCodeMix 2021 Shared Task on Sentiment Detection in Tamil, Malayalam, and Kannada. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 4–6.

K Rachana, M Prajnashree, Asha Hegde, and HL Shashirekha. 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*, pages 258–265.

Vivek Suresh Raj, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya S.K, Frank Glavin, and Bharathi Raja Chakravarthi. 2024. ConBERT-RL: A Policy-driven Deep Reinforcement Learning Based Approach for Detecting Homophobia and Transphobia in Low-resource Languages. *Natural Language Processing Journal*, 6:100040.

Jatinderkumar R Saini and Saikat Roy. 2023. Preparation of Rich Lists of Research Gaps in the Specific Sentiment Analysis Tasks of Code-mixed Indian Languages. *SN Computer Science*, 5(1):117.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *NeurIPS EMC*$^2Workshop$.

Kogilavani Shanmugavadivel, Sowbharanika Janani J S, Navbila K, and Malliga Subramanian. 2024a. Code Maker@DravidianLangTech-EACL 2024: Sentiment Analysis in Code-Mixed Tamil using Machine Learning Techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An Analysis of Machine Learning Models for Sentiment Analysis of Tamil Code-mixed Data. *Computer Speech Language*, 76:101407.

Kogilavani Shanmugavadivel, Malliga Subramanian, Palanimurugan V, and Pavul chinnappan D. 2024b. InnovationEngineers@DravidianLangTech-EACL 2024: Sentimental Analysis of YouTube Comments in Tamil by using Machine Learning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Kriti Singhal and Jatin Bedi. 2024. Transformers@DravidianLangTech-EACL2024: Sentiment Analysis of Code-Mixed Tamil Using RoBERTa. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

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.

Zannatul Fardaush Tripty, Md. Arian Al Nafis, Antu Chowdhury, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUETSentimentSillies@DravidianLangTech-EACL2024: Transformer-based Approach for Sentiment Analysis in Tamil and Tulu Code-Mixed Texts . In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Charangan Vasantharajan, Ruba Priyadharshini, Prasanna Kumar Kumarasen, Rahul Ponnusamy, Sathiyaraj Thangasamy, Sean Benhur, Thenmozhi Durairaj, Kanchana Sivanraju, Anbukkarasi Sampath, and Bharathi Raja Chakravarthi. 2022. TamilEmo: Fine-grained Emotion Detection Dataset for Tamil. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 35–50. Springer.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A Survey on Sentiment Analysis Methods, Applications, and Challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. TwHIN-BERT: A Socially-enriched Pre-trained Language Model for Multilingual Tweet Representations at Twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5597–5607.