MUNLP@DravidianLangTech2023: Learning Approaches for Sentiment Analysis in Code-mixed Tamil and Tulu Text

Kavya G[a], **Asha Hegde**[b], **Sharal Coelho**[c],
Pooja Lamani[d], **Hosahalli Lakshmaiah Shashirekha**[e]
Department of Computer Science, Mangalore University, Mangalore, India
{[a]kavyamujk,[b]hegdekasha,[c]sharalmucs,[d]poojamucs}@gmail.com
[e]hlsrekha@mangaloreuniversity.ac.in

Abstract

Sentiment Analysis (SA) examines the subjective content of a statement, such as opinions, assessments, feelings, or attitudes towards a person, product or anything. The increase in the online users has also increased the SA content demanding the automated tools to analyze such content. Though several models are developed for SA in high-resource languages like English, Spanish, German, etc., under-resourced languages like Dravidian languages are less explored. Added to this is the complexity of code-mixed texts on social media. To address the challenges of SA in code-mixed under-resourced Dravidian language texts, in this paper, we team - MUNLP, describe the models submitted to "Sentiment Analysis in Tamil and Tulu - DravidianLangTech" shared task at Recent Advances in Natural Language Processing (RANLP)-2023. Three models: i) n-gramsSA - an n-grams based model in which Term Frequency-Inverse Document Frequency (TF-IDF) of word n-grams and characters sequences within the word boundary (char_wb) both in the range (1, 3), is used to train Linear Support Vector Classifier (LinearSVC), ii) EmbeddingsSA - a Linear SVC model trained with a concatenation of fastText and Byte Pair word embeddings, and iii) BERTSA - a Transfer Learning (TL) model with Tamil sentiment Bidirectional Encoder Representations from Transformers (BERT) are proposed for SA. Among the three models, BERTSA exhibited a macro F1 score of 0.26 for code-mixed Tamil texts securing 2nd place in the shared task and EmbeddingsSA exhibited a macro F1 score of 0.53 securing 2nd place for Tulu code-mixed texts.

1 Introduction

The dynamic nature of social media platforms like Twitter, Facebook, and YouTube, characterized by rapidly evolving user-generated content, underscores the importance of automated SA, opinion mining, hate speech detection and offensive language detection (Hegde et al., 2022b; Chinnaudayar Navaneethakrishnan et al., 2023). Analyzing the sentiments of opinions, reviews, comments, etc., about the photos, videos, songs, movies or anything, on social media platforms such as YouTube and Facebook, can offer valuable insights to the organizations and individuals to make informed decisions about the content (Bharathi and Agnusimmaculate Silvia, 2021; Bharathi and Varsha, 2022; Swaminathan et al., 2022; Subramanian et al., 2022; Chakravarthi et al., 2023a,b). By analyzing the opinions of the users', content providers can also tailor their strategies to better align the content with the preferences and expectations of users, fostering a stronger connection with the users with the intention of potentially enhancing their reputation. Such sentiment-driven insights enable effective engagement, enabling them to address concerns, capitalize on positive feedback, and adapt their approaches for the improved outcomes (Balouchzahi and Shashirekha, 2021; Chakravarthi et al., 2022a,b; Chakravarthi, 2023).

SA is the process of computationally identifying and categorizing opinions, emotions, and attitudes expressed in written or spoken language. It has emerged as a crucial field of study due to the significant marks it leaves on the online users. SA is witnessing the growing adoption on social media platforms like YouTube, where it is leveraged as a recommender system due to its significant impact on viewers. (Balouchzahi et al., 2021; Hegde and Shashirekha, 2022a)..

Technological limitations often lead users to express their sentiments and opinions in their native languages using the roman script, along with the inclusion of English words. This practice arises from the convenience of keying in roman letters compared to the more complex key combinations

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 275–281, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_040

required for native language scripts, particularly in the case of Indian languages (Balouchzahi et al., 2022a; Hegde and Lakshmaiah, 2022). This results in code-mixed text, where more than one languages are combined at sentence, word, or sub-word level within the same text. With the rise of social media platforms and user generated content like sentiments and emotions, code-mixing texts have become increasingly prevalent as users find it easier to communicate their thoughts mixing the words of different languages they are very much familiar.

The complexity of SA is significantly increased in code-mixed texts as they combine words from different languages at various linguistic-levels (Varsha et al., 2022; Balouchzahi et al., 2022b). Analyzing sentiments in such texts require the language models and algorithms that are capable of effectively handling code-mixed content, accurately identifying sentiment-bearing units, and deciphering the sentiments expressed in different languages. Overcoming these challenges is crucial for obtaining accurate and meaningful insights for code-mixed SA.

Indian languages in general and Dravidian languages in particular, are under-resourced languages and code-mixing adds further dimension, mainly due to the problems with collecting and annotating code-mixed data for various applications. "Sentiment Analysis in Tamil and Tulu" is a shared task in DravidianLangTech at RANLP 2023 with the aim of promoting SA of code-mixed texts in Tamil and Tulu (Hegde et al., 2023). We, team MUCS, submitted three distinct models: i) n-gramsSA - an n-grams based model in which Term Frequency-Inverse Document Frequency (TF-IDF) of word n-grams and char_wb in the range (1, 3) are used to train LinearSVC, ii) EmbeddingsSA - a Linear SVC model trained using concatenation of fastText and Byte Pair word embeddings, and iii) BERTSA - a model constructed by fine-tuning Tamil sentiment BERT (only for Tamil language), to address the challenges provided by the shared task.

The rest of paper is organized as follows: while Section 2 describes the recent literature on code-mixed text processing and SA, Section 3 focuses on the description of the models submitted to the shared task followed by the experiments and results in Section 4. Conclusion and future works are included in Section 5.

2 Related Work

SA aims to identify and classify sentiments expressed in text data into one of the predefined set of sentiments such as positive, negative, neutral, and mixed feelings. Machine Learning (ML), Deep Learning (DL), and Transfer Learning (TL), are the commonly used approaches for SA and few of the relevant works are described below:

Kumar et al. (2021) presented an ensemble of ML classifiers (Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF)) trained with TF-IDF of character n-grams in the range (1, 3). to classify code-mixed Kannada, Malayalam, and Tamil texts for SA. Their models exhibited weighted F1 scores of 0.63, 0.73, and 0.62 for Kannada, Malayalam, and Tamil code-mixed texts respectively. Babu et al. (2020) proposed two distinct Sentence BERT (SBERT) models, one which uses Manglish features as additional features during fine-tuning cross entropy loss and another that utilizes Class Balanced Loss (CBL) to handle data imbalance for SA in code-mixed Malayalam-English text. Out of these models SBERT with CBL outperformed the other model with macro F1 score of 0.71. Puranik et al. (2021) fine-tuned two pretrained models: Universal Language Model FIne-Tuning (ULMFiT) and multilingual BERT (mBERT) models for SA in code-mixed Dravidian languages (Kannada, Tamil, and Malayalam) and obtained macro F1 scores of 0.63, 0.65 and 0.70 for code-mixed Kannada, Tamil and Malayalam texts respectively.

Hegde and Shashirekha (2022b) describe Long Short Term Memory (LSTM) models trained using Dynamic Meta Embedding (DME) features to perform SA and homophobia detection as Task A (code-mixed Kannada and Malayalam texts) and Task B (code-mixed Tamil-English and English texts) respectively. Their proposed models exhibited macro F1 scores of 0.61, and 0.44 for code-mixed Malayalam and Kannada texts respectively in Task A and macro F1 scores of 0.58 and 0.74 for code-mixed Tamil-English and English texts respectively in Task B. Balouchzahi et al. (2021) proposed ensemble model (LR, SVM, and Multilayer Perceptron (MLP)) with majority voting for SA in code-mixed Kannada, Malayalam, and Tamil texts. Using TF-IDF of character n-grams in the range (1, 5) and syllables in the range (1, 6), to train the ensemble model, they obtained weighted average F1 scores of 0.628, 0.726, and 0.619 for Kannada,

Hyperparameters	Values
penalty	l2
C	1.0
class_weight	balanced
max_iter	max_iter
random_state	100
loss	squared_hinge

Table 1: Hyperparameters and their values used in LinearSVC algorithm

Malayalam, and Tamil code-mixed texts respectively. Three models: SACo-Ensemble, SACo-Keras, and SACo-ULMFiT, using ML, DL, and TL respectively are proposed by Balouchzahi and Shashirekha (2021) for SA in Tamil and Malayalam code-mixed texts. SACo-Ensemble is an ensemble model (MLP, eXtreme Gradient Boosting (XGB) and LR) trained with the count vectors of character sequences in range (2, 6), sub-words extracted from Byte Pair embeddings, syntactic bi-grams, and tri-grams, vectorized using CountVectorizer[1]. These features are also used to train SACo-Keras models. To build SACo-ULMFiT model, the authors pre-trained it with raw text (Dakshina dataset along with code-mixed Tamil and Malayalam texts) and they fine-tuned with their Train sets for SA. Among the three models, SACo-Ensemble models obtained weighted average F1 scores of 0.62 and 0.72 for code-mixed Tamil and Malayalam texts respectively.

From the above literature, it is clear that the researchers explored several models to perform SA in both high-resource and low-resource languages. However, for most of the low-resource languages, performances of the models are still less indicating the scope for developing models for SA in low-resource languages. Further, the code-mixed nature of the social media comments in low-resource languages intensifies the SA task.

3 Methodology

Three distinct models: n-gramsSA, EmbeddingsSA, and BERTSA, are proposed for SA of code-mixed Tamil and Tulu texts. The steps involved in building the proposed models are given below:

3.1 Preprocessing

Preprocessing step includes removing punctuation, numerical data, user mentions, hashtags as well as stopwords. English stopwords available at the Natural Language Tool Kit (NLTK)[2] library and Tamil[3] stopwords available at GitHub repository are used as references to remove the English and Tamil stopwords respectively. Emojis are transformed into English text using demoji[4] library.

Hyper Parameters	Values
Learning Rate	2e-5
Batch Size	16
Number of Epochs	2
Dropout Rate	0.1
Optimizer	Adam
Weight Decay	0.01
Warmup Steps	500
Maximum Sequence Length	128
Embedding size	768

Table 2: Hyperparameters and their values used in BERTSA

3.2 Model Description

The framework of n-gramsSA and EmbeddingsSA models is visualized in Figure 1. Both the models use LinearSVC classifer and the hyperparameters and their corresponding values used in LinearSVC classifier are shown in Table 1. The hyperparameters which are not mentioned in table are used with their default values. The description of the models follows:

n-gramsSA model - n-grams are sequential collections of lexical units viz, words/characters which capture the context by the sequential patterns present in the text/words facilitating a deeper comprehension of relationships between word/characters. Selecting the appropriate value for 'n' in n-grams involves the desired level of context. Larger 'n' values, such as 3 or more, capture longer sequences and dependencies between words, which is beneficial for text classification tasks like, SA, hate speech detection, and emotion analysis but also increases the complexity of the learning models (Balouchzahi and Shashirekha, 2021). Word and character n-grams in the range of (1, 3) are extracted and vectorized using TfidfVectorizer[5] to train the LinearSVC model for SA.

[1]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[2]https://pythonspot.com/nltk-stop-words/

[3]https://gist.github.com/arulrajnet/e82a5a331f78a5cc9b6d372df13a919c

[4]https://pypi.org/project/demoji/

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

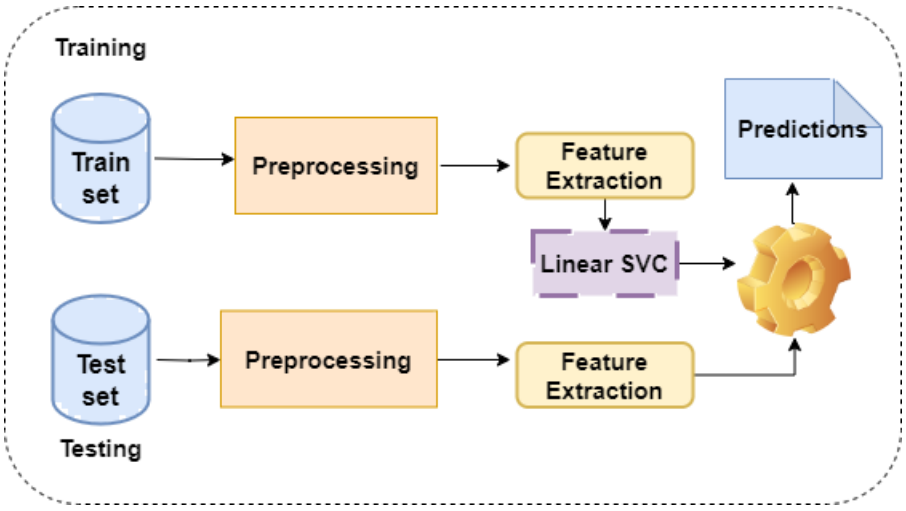
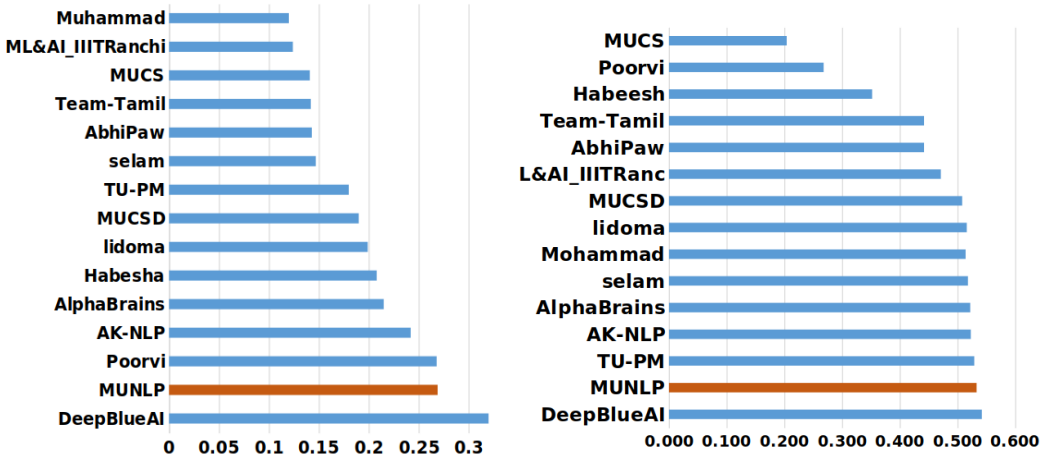


Figure 1: Framework of n-gramsSA and EmbeddigsSA models



(a) Tamil (b) Tulu

Figure 2: Comparison of macro F1 scores of the participating teams in the shared task

EmbeddingsSA model - distributed representation of words, also known as word embeddings, is a popular word representation technique, where each word is represented by a low-dimensional dense vector such that words having the same meaning will have a similar representation. Sub-word embeddings are dense vector representations of sub-word units - linguistic units that are smaller than complete words but larger than individual characters, that capture their semantic and syntactic properties. The advantage of sub-word embeddings is that it helps to capture the morphological structure even for rare or unseen words. Word embeddings extracted from fastText[6] and Byte Pair embeddings[7] are concatenated to train the LinearSVC model for SA.

[6]https://fasttext.cc/docs/en/crawl-vectors.html
[7]https://bpemb.h-its.org/

Label	Train Set		Development Set	
	Tulu	Tamil	Tulu	Tamil
Positive	3,118	20,070	369	2,257
Neutral	1,719	5,628	202	611
Mixed Feeling	974	4,020	120	480
Negative	646	4,271	90	438

Table 3: Classwise distribution of code-mixed Tamil and Tulu texts

BERTSA - BERT pretrained models have shown remarkable performance in capturing subtle information in text, making them one of the most trending approaches for text classification tasks (Sun et al., 2019). By leveraging large-scale pretraining on diverse text data, BERT models excel in understanding contextual relationships, handling syntactic and semantic information and capturing

Language	Model	Development set		Test set	
		With imbalanced data	With balanced data	With imbalanced Data	With balanced Data
Tamil	**n-gramsSA**	0.40	0.45	0.16	0.18
	EmbeddingsSA	0.40	0.46	0.15	0.18
	BERTSA	0.45	0.48	0.23	0.26
Tulu	**n-gramsSA**	0.30	0.33	0.40	0.45
	EmbeddingsSA	0.54	0.56	0.48	**0.53**

Table 4: Performance of the proposed models with imbalanced and balanced dataset

the fine-grained details present in the text.

Tamil sentiment BERT[8] is a variant of the BERT model specifically trained for SA in Tamil text catering to the specific linguistic and cultural intricacies of the language. This variant is fine-tuned to capture the subtle information of sentiments in Tamil text which helps to identify the emotions and opinions effectively. This model leverages the power of pretraining on a large amount of text data and fine-tuning on SA tasks to classify the sentiments in Tamil language. Hyperparameters and their values of Tamil sentiment BERT are shown in Table 2. The hyperparameters which are not mentioned in Table 2 are used with their default values.

4 Experimental results

The dataset provided by the shared task organizers includes code-mixed Tamil (Chakravarthi et al., 2020) and Tulu (Hegde et al., 2022a) texts and the distribution of the labels across the Train and Development sets for these two languages are shown in Table 3. The imbalance nature in both the datasets motivated to use LinearSVC with the hyperparameter - class weight = 'balanced'. The proposed models are evaluated on the unlabeled Test set provided by the organizers and the predictions are evaluated based on macro F1 score. The performance of the proposed models on the Development and Test sets with imbalanced and balanced data are shown in Table 4.

The results illustrate that, both n-gramsSA and EmbeddingsSA models exhibited similar performance with the same macro F1 score of 0.18. Further, BERTSA model outperformed the other models with a macro F1 score of 0.26 securing 2nd rank for code-mixed Tamil text. For Tulu, among n-gramsSA and EmbeddingsSA models, EmbeddingsSA model outperformed the other with macro F1 score of 0.53 on Test set securing 2nd rank in the shared task. However, the macro F1 score obtained by the proposed models are below average. The macro F1 scores are very low due to the imbalance in the Train set. Though the hyperparameter 'class_weight' set to 'balanced' resolves the data imbalance issue to some extent, the extreme data imbalance in the Train set leads to overfitting.

The misclassified comments in Tamil and Tulu Test sets along with the English translations, actual labels, model predictions (obtained from BERTSA and EmbeddingsSA models evaluated for Tamil and Tulu Test sets respectively), and probable reasons for misclassification are shown in Table 5. It can be observed that the wrong classifications may also be due to the incorrect annotations.

The comparison of macro F1 scores of all the participating teams for SA in both code-mixed Tamil and Tulu text is shown in Figure 2.

5 Conclusion

This paper describes the models submitted by our team - MUNLP, to the shared task "Sentiment Analysis in Tamil and Tulu" in DravidianLangTech at RANLP 2023, for SA in code-mixed Tamil and Tulu texts. Three distinct models: i) n-gramsSA - an n-grams based model trained with n-grams of words and char_wb both in the range (1, 3) are used to train LinearSVC, ii) EmbeddingsSA - a model built using combination of fastText and Byte Pair embeddings to train LinearSVC, and iii) BERTSA - a model constructed by fine-tuning Tamil sentiment BERT (only for Tamil language), to address the challenges of the shared task. Among the proposed models, BERTSA and EmbeddingsSA models obtained macro F1 scores of 0.26 and 0.53 for code-mixed Tamil and Tulu datasets respectively both securing 2nd rank.

The results indicate that the macro F1 scores are low. Suitable oversampling or text augmentation techniques will be explored to improve macro F1

[8]https://huggingface.co/vishnun/bert-base-cased-tamil-mix-sentiment

Language	Comments	English Translations	Actual Label	Predicted Label	Remarks
Tamil	முதலில் இவர்கள் மீது பாவம் வரும் இப்போ வெறுப்பு வருகிறது	First there is sin against them, now there is hatred	Mixed Feeling	unknown_ state	The content words 'முதலில்' and 'வருகிறது' are associated with the class 'unknown_state' rather than with the class 'Mixed Feeling' in the Train set. Further, as the other content words (இவர்கள், இப்போ, and வெறுப்பு) will not give any hint about 'Mixed Feeling' class, the comment is classified as 'unknown_state'.
	இதுபோன்ற சம்பவம் எங்கள் புதுமனை புகுவிழாவில் நடந்தது 😈	A similar incident happened at our New Year's Eve party	Negative	Positive	This comment is annotated as 'negative' because of the emoji. However, the content words (இதுபோன்ற, எங்கள், புதுமனை, புகுவிழாவில், and நடந்தது) speaks about positive sentiment because of which the comment is classified as 'positive'.
Tulu	Embe onji mage karmda	A scolding Tulu comment	Negative	Neutral	The wrong classification may be because of only four words in the comment. Moreover, only one word (karmda) belongs to negative polarity and the rest speaks about positive polarity.
	Kandani ...gode gode panpuna part...sakath	The part in which the husband is repeatedly telling is good	Positive	Mixed Feeling	The word 'sakath', which makes the entire comment as positive, has rarely appeared in the Train set. This has lead to the wrong classification of the comment.

Table 5: Samples of misclassification for code-mixed Tamil and Tulu text

scores of the proposed models. Efficient resampling techniques will be explored further to handle imbalanced classes with effective feature extraction.

References

Yandrapati Prakash Babu, Rajagopal Eswari, and K Nimmi. 2020. CIA_NITT@ Dravidian-CodeMix-FIRE2020: Malayalam-English Code mixed Sentiment Analysis Using Sentence BERT And Sentiment Features. In *FIRE (Working Notes)*, pages 566–573.

Fazlourrahman Balouchzahi, Sabur Butt, A Hegde, Noman Ashraf, HL Shashirekha, Grigori Sidorov, and Alexander Gelbukh. 2022a. Overview of CoLI-Kanglish: Word Level Language Identification in Code-mixed Kannada-English Texts at ICON 2022. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 38–45.

Fazlourrahman Balouchzahi and HL 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.

Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha, and Grigori Sidorov. 2021. CoSaD-Code-Mixed Sentiments Analysis for Dravidian Languages. In *CEUR Workshop Proceedings*, volume 3159, pages 887–898. CEUR-WS.

Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha, Grigori Sidorov, and Alexander Gelbukh. 2022b. 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.

B Bharathi and A Agnusimmaculate Silvia. 2021. SSNCSE_NLP@DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318, Kyiv. Association for Computational Linguistics.

B Bharathi and Josephine Varsha. 2022. SSNCSE NLP@TamilNLP-ACL2022: Transformer based approach for detection of abusive comment for Tamil language. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 158–164, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in Youtube comments. *International Journal of Data Science and Analytics*, pages 1–20.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022a. How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023a. Offensive language identification in Dravidian languages using MPNet and CNN. *International Journal of Information Management Data Insights*, 3(1):100151.

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

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023b. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Subalalitha Chinnaudayar Navaneethakrishnan, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi, Lavanya Sambath Kumar, and Rahul Ponnusamy. 2023. Findings of shared task on sentiment analysis and homophobia detection of Youtube comments in code-mixed Dravidian languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 18–21, New York, NY, USA. Association for Computing Machinery.

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, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, SUBALALITHA CN, Lavanya S K, Thenmozhi D, Martha Karunakar, Shreya Shreeram, and Sarah" Aymen. 2023. 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*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Asha Hegde, Sharal Coelho, Ahmad Elyas Dashti, and Hosahalli Shashirekha. 2022b. MUCS@ Text-LT-EDI@ ACL 2022: Detecting Sign of Depression from Social Media Text using Supervised Learning Approach. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 312–316.

Asha Hegde and Shashirekha Lakshmaiah. 2022. Mucs@ mixmt: Indictrans-Based Machine Translation for Hinglish Text. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1131–1135.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2022a. Learning Models for Emotion Analysis and Threatening Language Detection in Urdu Tweets.

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

Abhinav Kumar, Sunil Saumya, and Jyoti Prakash Singh. 2021. An Ensemble-based Model for Sentiment Analysis of Dravidian Code-mixed Social Media Posts. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR*.

Karthik Puranik et al. 2021. IIITT@ Dravidian-CodeMix-FIRE2021: Transliterate or Translate? Sentiment Analysis of Code-mixed Text in Dravidian Languages. In *arXiv preprint arXiv:2111.07906*.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in Tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for Aspect-based Sentiment Analysis via Constructing Auxiliary Sentence. In *arXiv preprint arXiv:1903.09588*.

Krithika Swaminathan, Bharathi B, Gayathri G L, and Hrishik Sampath. 2022. SSNCSE_NLP@LT-EDI-ACL2022: Homophobia/transphobia detection in multiple languages using SVM classifiers and BERT-based transformers. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 239–244, Dublin, Ireland. Association for Computational Linguistics.

Josephine Varsha, B Bharathi, and A Meenakshi. 2022. Sentiment Analysis and Homophobia Detection of YouTube Comments in Code-Mixed Dravidian Languages using Machine Learning and Transformer Models. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.