CUETSentimentSillies@DravidianLangTech-EACL2024: Transformer-based Approach for Sentiment Analysis in Tamil and Tulu Code-Mixed Texts

Zannatul Fardaush Tripty, Md. Arian Al Nafis, Antu Chowdhury, Jawad Hossain, Shawly Ahsan, Avishek Das and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u1804030, u1804111, u1804112, u1704039, u1704057}@student.cuet.ac.bd,
{ avishek,moshiul_240}@cuet.ac.bd

Abstract

Sentiment analysis (SA) on social media reviews has become a challenging research agenda in recent years due to the exponential growth of textual content. Although several effective solutions are available for SA in high-resourced languages, it is considered a critical problem for low-resourced languages. This work introduces an automatic system for analyzing sentiment in Tamil and Tulu code-mixed languages. Several ML (DT, RF, MNB), DL (CNN, BiLSTM, CNN+BiLSTM), and transformer-based models (Indic-BERT, XLM-RoBERTa, m-BERT) are investigated for SA tasks using Tamil and Tulu code-mixed textual data. Experimental outcomes reveal that the transformer-based models XLM-R and m-BERT surpassed others in performance for Tamil and Tulu, respectively. The proposed XLM-R and m-BERT models attained macro F1-scores of 0.258 (Tamil) and 0.468 (Tulu) on test datasets, securing the 2^{nd} and 5^{th} positions, respectively, in the shared task.

1 Introduction

Social media has changed how people network and socialize, especially the younger generation, and multilingual user interfaces allow people to express their emotions in their native languages (Ahmad and Singla, 2021; Patra et al., 2018; Tarihoran and Sumirat, 2022). Sentiment analysis (SA) may help firms assess their brand's image and sentiment and make informed customer relationship management and marketing choices. It analyzes social media postings to detect user attitudes (Chakravarthi et al., 2020c). Code-mixed texts greatly concern sentiment analysis. Many multilingual societies use code-mixed texts, combining words, morphemes, and phrases from two or more languages (Chakravarthi et al., 2023). This behavior is problematic for SA systems, mainly when they utilize non-native scripts like Roman letters to represent languages written in other scripts (Hegde and Shashirekha, 2022). Coded language texts need specialized sentiment analysis due to language mixing and context-dependent emotions. Scholars are improving security awareness methods to govern virtual communication's growth (Chakravarthi et al., 2021; Hegde et al., 2023). The goal is to create a system that can classify code-mixed sentiment polarity in Tamil-English and Tulu-English code-mixed texts into four predetermined categories: positive, negative, mixed feeling, and neutral/unknown state. The main contributions of this study are:

- Developed numerous ML and DL methods and fine-tuned transformers to classify textual sentiment into four categories () for Tamil and Tulu code-mixed datasets.
- Investigated the effectiveness of the developed models for Tamil and Tulu subtasks, where XLM-RoBERTa exceeded other models for Tamil and m-BERT exceeded other models for the Tulu language.

2 Related Work

Researchers studying several SA techniques tend to focus on high-resource languages such as English and Spanish. However, SA is also being studied in code-mixed, low-resource languages. Shetty (2023) trained various ML models for SA of Tamil and Tulu code-mixed texts. The proposed method yielded F1 scores of 0.14 and 0.204 in Tamil and Tulu, respectively. To detect abusive comments in code-mixed Tamil text, Bharathi and Varsha (2022) employed BERT, m-BERT, and XLNet models. They obtained a weighted F1 score of 0.96 for Tamil-English code-mixed text and a weighted F1 score of 0.59 for Tamil text. Babu and Eswari (2021) improved sentiment analysis using Paraphrase XLM-R on Dravidian code-mixed YouTube comments. They trained the model using Tamil,

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 234–239

Malayalam, and Kannada code-mixed language datasets and achieved F1 scores of 71.1, 75.3, and 62.5, respectively. Chakravarthi et al. (2020a) created a gold standard Tamil-English code-switched, sentiment-annotated corpus containing 15,744 comment posts from YouTube.

An m-BERT-based model utilized by Zhu and Dong (2020) for SA where self-attention was employed to assign a weight to the output of the BiLSTM. The proposed model achieved weighted average F1 scores of 0.73 and 0.64 in Malayalam and Tamil, respectively. Rakshitha et al. (2021) proposed a model that used Twitter APIs to collect consumer reviews. TextBlob rated these reviews and classified them as favorable, negative, or neutral using a text classification algorithm. Ehsan et al. (2023) developed BiLSTM network-based models for sentiment analysis of code-mixed Tamil and Tulu. ELMo embedding was trained on larger unannotated code-mixed text corpora. The proposed model achieved macro F1-scores of 0.2877 and 0.5133 on Tamil and Tulu code-mixed datasets, respectively.

3 Task and Dataset Descriptions

The goal of this task is sentiment analysis in Tamil and Tulu, explicitly focusing on determining sentiment polarity in social media comments. This task aimed to develop two systems that can individually identify sentiment polarity from a given set of texts in Tamil or Tulu. To achieve this, we utilized the corpora provided by the shared task organizers[1] for sentiment analysis in Tamil (Chakravarthi et al., 2020b) and Tulu (Hegde et al., 2022). The task required classifying texts into four predefined classes, positive, negative, mixed feeling, and neutral/unknown state, for Tamil and Tulu code-mixed texts.

Table 1 summarizes the Tamil dataset. The combined training and development sets for Tamil exhibited the highest number of samples for the positive class (22,327 texts). Subsequently, the unknown state category comprised 6,239 texts, while negative had 4,751 texts, and mixed feelings had 4,559 texts, each containing fewer instances than the positive class. The Tulu dataset was divided into three subsets: training, development, and testing, containing 6,945, 500, and 501 samples, respectively (Table 2). The dataset demonstrated an uneven distribution among classes, with the positive class having the most samples with 3,831 texts, neutral with 2,118 texts, negative with 796 texts, and mixed feelings with 1,201 texts having fewer samples. Text lengths in the dataset varied from one word to 261 words, with an average length of 7 words.

Classes	Train+Dev	Test	Total words
Positive	22327	73	208365
Positive (after augmentation)	22327	73	187294
Unknown state	6239	137	69311
Unknown state (after augmentation)	17135	137	177181
Negative	4751	338	51459
Negative (after augmentation)	14040	338	188356
Mixed feelings	4458	101	64844
Mixed feelings (after augmentation)	13461	101	133810

Table 1: Tamil dataset statistics before and after augmentation

Classes	Train	Dev	Test	Total words
Positive	3352	231	248	22298
Negative	698	55	43	4658
Neutral	1854	124	140	12738
Mixed feelings	1041	90	70	7033
Total	**6945**	**500**	**501**	**46727**

Table 2: Tulu Dataset Statistics

4 Methodology

This section summarized the methods and techniques applied for sentiment analysis in Tamil and Tulu. Figure 1 outlines the employed techniques for SA in Tamil and Tulu.

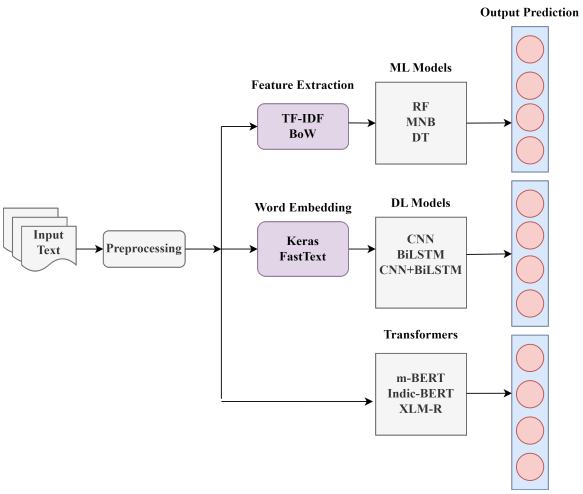

Figure 1: Abstract outlines of textual SA in Tamil and Tulu

[1] https://sites.google.com/view/dravidianlangtech-2024/home

4.1 Data Augmentation

The Tamil dataset 1 before augmentation exhibited an imbalance, specifically in the unknown state, negative, and mixed feelings classes, with fewer samples compared to the positive class. We merged the training and development sets to rectify this to minimize the class distribution gap. Additionally, we applied back translation using Google Translator for data augmentation. Google Translator was selected for its widespread availability and proven effectiveness in generating diverse language variations. The back translation process involved iteratively translating sentences from Tamil to another language and then back to Tamil, introducing nuanced variations. After augmentation, there were notable shifts in class distribution: the unknown state class increased to 17,135 texts, the negative class reached 14,040 texts, and the mixed feelings class grew to 13,461 texts. This combined strategy of merging datasets and the back translation method aimed to broaden the dataset's scope and ensure a more representative distribution of sentiment classes, specifically addressing data scarcity in the unknown-state, harmful, and mixed feelings categories. This precise approach enhanced the dataset's robustness and reliability for subsequent analysis and model development.

4.2 Preprocessing

The dataset obtained from YouTube comments underwent preprocessing to ensure that it was clear of irrelevant information. This process involved the elimination of emojis, punctuation, spaces, URLs, and numerical texts. English letters are transformed to lowercase. To enhance linguistic relevance, common stopwords were manually eliminated based on a curated list obtained from a Tamil stopwords repository on GitHub[2]. Similarly, for the Tulu language, English stop words were excluded. We also identified and removed Tamil and Tulu's ten most frequently occurring words.

4.3 Training

The initial step involved extracting features using various feature extraction techniques and applying different ML, DL, and transformer-based approaches.

ML Baselines: TF-IDF (Nayel, 2020) values were used as features for training ML models based on unigram features. Additionally, bag-of-words (BoW) representations are also utilized for feature extraction. Traditional ML-based methods, including RF, DT, and MNB, were employed for sentiment analysis. In the DT model, the regularization parameter was set to 2. RF was implemented with 100 estimators (n_estimator 100) to enhance its predictive performance.

DL Baselines: Three DL models, CNN, BiLSTM, and CNN+BiLSTM, along with FastText (Joulin et al., 2016) and Keras embeddings, were employed for sentiment analysis. In the CNN model, the process began with an embedding layer, followed by three convolutional layers featuring 64, 32, and 16 filters. MaxPooling layers were added after convolution layers for feature reduction. In the BiLSTM model, the embedding layer was followed by two BiLSTM layers with 32 and 16 units, respectively, capturing information bi-directionally. The resulting sequences were flattened, and a dense layer with softmax activation was added for classification. In the CNN+BiLSTM model, the embedding layer was followed by a convolutional layer with 128 filters and a kernel size of 5. A BiLSTM with 32 units and a dropout rate (0.2) is added after the convolution layer.

Transformers: Three transformer-based models, XLM-RoBERTA (Conneau et al., 2019), IndicBERT (Kakwani et al., 2020), and m-BERT (Devlin et al., 2018), were utilized for SA in Tamil and Tulu. This work used the same hyperparameters for Tamil and Tulu subtasks training. Specifically, during the training of all transformers, we used the Adafactor optimizer with a consistent learning rate of 2e-5 over 10 epochs, incorporating a warm-up ratio of 0.1 for a smoother initialization. To improve stability, gradient accumulation steps were doubled to 2. A weight decay of 0.01 was applied to regularize the training process. Fine-tuned hyperparameter values allowed us to do extensive training and optimization of the model parameters. The choice of a batch size of 16 facilitated efficient processing and updating of the model weights during each iteration.

5 Results and Analysis

Table 3 displays the results of various employed approaches for the SA task on the Tamil test set, with the XLM-RoBERTA model leading among transformers with a macro F1 score (0.258). The RF model with BoW surpassed other ML models, achieving the highest macro F1 score (0.248). No-

[2]https://gist.github.com/arulrajnet

tably, the CNN+BiLSTM model exhibited superior performance compared to other DL models.

Classifier	P	R	F
DT (TF-IDF)	0.247	0.251	0.237
RF (TF-IDF)	0.270	0.26	0.24
MNB (TF-IDF)	0.324	0.27	0.213
DT (BoW)	0.280	0.297	0.248
RF (BoW)	0.228	0.252	0.056
MNB (BoW)	0.282	0.248	0.185
CNN (Keras)	0.235	0.232	0.214
BiLSTM (Keras)	0.262	0.258	0.253
C+B (Keras)	0.250	0.257	0.241
CNN (FastText)	0.220	0.230	0.137
BiLSTM (FastText)	0.239	0.240	0.147
C+B (FastText)	0.234	0.236	0.148
m-BERT	0.275	0.269	0.255
XLM-RoBERTa	**0.288**	**0.27**	**0.258**
Indic-BERT	0.276	0.265	0.252

Table 3: Performance of various models on the Tamil test set where P, R, and F denote precision, recall, and macro F1-score, respectively, and C+B represents the CNN+BiLSTM model

For the Tulu test set, as shown in Table 4, the m-BERT model excelled among transformer models, attaining the highest macro F1 score of 0.468. Among ML models, the RF model with BoW stood out with the highest macro F1 score of 0.449, while within DL models, BiLSTM with Fasttext emerged as the top performer with macro F1 of 0.394.

5.1 Error Analysis

The best-performed models (XLM-RoBERTa for Tamil texts, and m-BERT for Tulu texts) are further investigated to understand better insights regarding the performance using quantitative and qualitative analysis.

Quantitative Analysis: The confusion matrix is used for error analysis for both Tamil (Figure 2) and Tulu (Figure 3) datasets.

In Tamil, we found that the model did well with TPR of 33.13% and 28.46% negative and unknown state, respectively. However, the positive class had a lower TPR of 20.54%, meaning the model struggled to identify positive sentiments. The confusion matrix for Tulu revealed a True Positive Rate (TPR) of 90.70% for the positive class. Conversely, the mixed feeling class exhibited the lowest TPR of 10%. Notably, the model misidentified 35 mixed-feeling class text samples as neutral, indicating difficulty distinguishing between texts conveying

Classifier	P	R	F
DT (TF-IDF)	0.442	0.449	0.443
RF (TF-IDF)	0.465	0.434	0.424
MNB (TF-IDF)	**0.565**	0.360	0.334
DT (BoW)	0.420	0.431	0.436
RF (BoW)	0.518	0.459	0.449
MNB (BoW)	0.514	0.428	0.427
CNN (Keras)	0.370	0.405	0.383
BiLSTM (Keras)	0.380	0.373	0.357
C+B (Keras)	0.379	0.374	0.367
CNN (Fasttext)	0.379	0.374	0.367
BiLSTM (Fasttext)	0.444	0.394	0.394
C+B (Fasttext)	0.379	0.374	0.367
m-BERT	0.512	**0.468**	**0.468**
XLM-RoBERTa	0.454	0.405	0.387
indic-BERT	0.307	0.399	0.344

Table 4: Performance of various models on the Tulu test set where P, R, and F denote precision, recall, and macro F1-score, respectively, and C+B represents the CNN+BiLSTM model

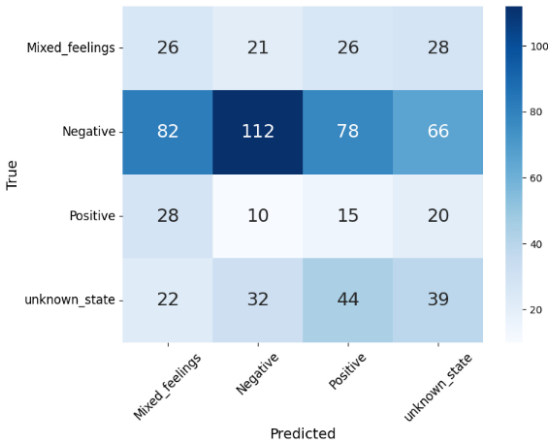

Figure 2: Confusion matrix of XLM-RoBERTA for Tamil test set

mixed feelings and those with neutral sentiments. This challenge arose due to the nuanced similarity in meaning between texts with mixed feelings and those that are neutral, leading to frequent misclassifications, primarily for the neutral class.

5.2 Qualitative Analysis:

Figure 4 illustrates some predicted outcomes by the best-performed model (XLM-RoBERTa) for Tamil SA task. It is revealed that the proposed model demonstrated accurate predictions for sample 2 while other samples were misclassified. It exhibited challenges in correctly categorizing text samples 1,3,4. Especially for texts of *mixed-feelings* and

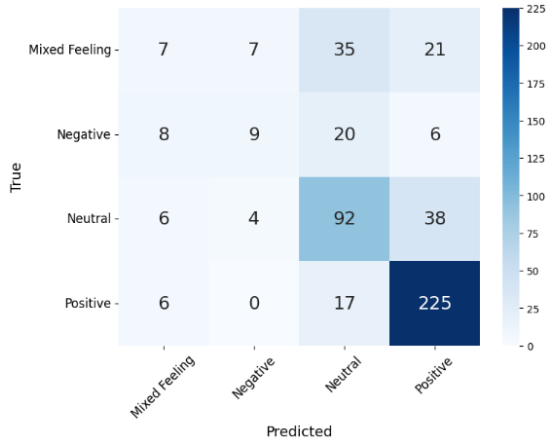

Figure 3: Confusion matrix of m-BERT for Tulu test set

unknown state categories. Because texts of these two categories were semantically analogous.

Text Sample	Actual	Predicted
Sample1: திரு நங்கைகளின் அட்டகாசம் அதிகமாகிக்கொண்டே போகிறது ரயில் நிலையம், பஸ்ஸில்..உடனே தமிழக அரசு நடவடிக்கை எடுக்க வேண்டும் .. (The violence of Mr. Nankais is increasing in railway station, bus.Tamil Nadu government should take action immediately.)	Positive	Negative
Sample 2: இவர்களை போலீஸ் கொண்டு போய் சூ லாடம் கட்டவேண்டும் (The police should take them and arrest them)	Negative	Negative
Sample 3:கடவுளின் சாபத்தில் வாழும் உங்களுக்கு இந்த அநியாயம் இருக்க கூடாது (You who live under God's curse should not have this injustice)	Unknown State	Mixed Feeling
Sample 4 : இவர்கள் அராஜகம் மிக அதிகம். போலீஸ்துறை கடுமையான நடவடிக்கை எடுக்க வேண்டும். (They are very anarchic. Police are strict Action should be taken.)	Mixed Feeling	Unknown State

Figure 4: A few examples of predicted output by the proposed (XLM-RoBERTa) model for Tamil SA task. The corresponding english texts were translated using 'Google Translator')

Figure 5 presents a few predicted samples by the best-performed model (m-BERT) for Tulu SA task. It is shown that the proposed model depicted correct predictions for text samples 1 and 3. However, it failed to classify *neutral* and *mixed-feelings* properly as their texts were semantically similar. Therefore, the proposed model classified *neutral* texts as *mixed feelings* and *mixed feelings* as *neutral*.

Limitations

Despite these developments, several obstacles, such as the need for pre-trained transformer models, limit our capacity to use current language representations effectively. Although a single pre-trained model has been identified and made available for Tamil, there needs to be a similar resource for Tulu

Text Sample	Actual	Predicted
Sample 1 : ಮಾತಾ ಬಿತ್ತಿಲ್\u200cಡ್\u200cದ್ ಉಂದು ಭಾರಿ ಚೋಕಾತ್ಂಡ್	Positive	Positive
Sample 2: ಈ ಪ್ರೋಣ್ಣ್ ಪನ್ನಿನಗುಲ್ ಯೇ ರ್ಗಾ ಇಜ್ಜೆರ	Negative	Neutral
Sample 3: ಅರೆಗ್ ಎಂಡ ದಾನೆ ಪತೊಂದಿನ ಆತು	Neutral	Neutral
Sample 4: ಕುಸಾಸ್ ಇತ್ತುಂಡು. ಪ್ರಾಕ್ಟಿಸ್ ಕಮ್ಮಿ. ಒಟ್ಟುಗ್ ಓಕೆ.	Mixed Feeling	Neutral

Figure 5: Some examples of predicted output by the m-BERT model for Tulu SA task

presents a more significant obstacle, restricting the extent and effectiveness of language processing endeavors. Lower accuracy is also a result of the model's inability to learn patterns for minority classes due to the small sample size and notable class imbalance in the datasets. The lack of resources for data augmentation as back translation, particularly in Tulu's case, makes addressing this imbalance even more difficult.

6 Conclusion

This paper investigated different ML, DL, and transformer-based models for categorizing textual sentiment in Tamil and Tulu. The experimental results revealed that the XLM-RoBERTa model obtained the highest macro F1 score (0.258) for the Tamil textual comments, whereas the m-BERT model scored the best (0.468) for the Tulu texts. For future research, collaborative efforts are necessary to create resources, especially for low-resource languages like Tamil and Tulu, and add more diverse data to the existing corpora. Additionally, improving the performance of sentiment analysis models for these languages will depend heavily on developments in pre-training methods and the availability of more thorough language representations.

References

Gazi Imtiyaz Ahmad and Jimmy Singla. 2021. Sentiment analysis of code-mixed social media text (sacmsmt) in indian-languages. In *2021 International Conference on Computing Sciences (ICCS)*, pages 25–33. IEEE.

Yandrapati Prakash Babu and Rajagopal Eswari. 2021. Sentiment analysis on dravidian code-mixed youtube comments using paraphrase xlm-roberta model. *Working Notes of FIRE*.

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.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023. 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 P McCrae. 2020a. Corpus creation for sentiment analysis in code-mixed tamil-english text. *arXiv preprint arXiv:2006.00206*.

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, 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 *Proceedings of the 12th Annual Meeting of the 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. *arXiv preprint arXiv:2111.09811*.

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.

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.

Toqeer Ehsan, Amina Tehseen, Kengatharaiyer Sarveswaran, and Amjad Ali. 2023. Sentiment analysis of code-mixed tamil and tulu by training contextualized elmo representations. *RANLP'2023*, page 152.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. 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, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, SK Lavanya, Durairaj Thenmozhi, 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*, pages 64–71.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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*.

Hamada A Nayel. 2020. Nayel at semeval-2020 task 12: Tf/idf-based approach for automatic offensive language detection in arabic tweets. *arXiv preprint arXiv:2007.13339*.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

Kakuthota Rakshitha, H M Ramalingam, M Pavithra, H D Advi, and Maithri Hegde. 2021. Sentimental analysis of indian regional languages on social media. *Global Transitions Proceedings*, 2(2):414–420.

Poorvi Shetty. 2023. Poorvi@ dravidianlangtech: Sentiment analysis on code-mixed tulu and tamil corpus. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 124–132.

Naf'an Tarihoran and Iin Ratna Sumirat. 2022. The impact of social media on the use of code mixing by generation z. *International Journal of Interactive Mobile Technologies (iJIM)*, 16(7):54–69.

Yueying Zhu and Kunjie Dong. 2020. Yun111@ dravidian-codemix-fire2020: Sentiment analysis of dravidian code mixed text. In *FIRE (Working Notes)*, pages 628–634.