MUCS@DravidianLangTech-2024: A Grid Search Approach to Explore Sentiment Analysis in Code-mixed Tamil and Tulu

Prathvi B[a], **Manavi K K**[b], **Subrahmanya**[c],
Asha Hegde[d], **Kavya G**[e], **H L Shashirekha**[f]
Department of Computer Science, Mangalore University, Mangalore, Karnataka, India
{[a]bprathvi968, [b]kkmanavi, [c]subrahmanyapoojary789}@gmail.com,
{[d]hegdekasha, [e]kavyamujk}@gmail.com, [f]hlsrekha@mangaloreuniversity.ac.in

Abstract

Sentiment Analysis (SA) is a field of computational study that analyzes users' reviews, opinions, and emotions, towards any entity on online platforms. As user sentiments play a major role in decision making, there is an increasing demand for the tools that can effectively analyze the user-generated sentiments. The availability of user-generated code-mixed sentiments in low-resource languages like Tamil and Tulu further necessitates the growing need for efficient SA tools. To address SA in code-mixed Tamil and Tulu text, this paper describes the Machine Learning (ML) models submitted by our team - MUCS to "Sentiment Analysis in Tamil and Tulu - DravidianLangTech" - a shared task organized at European Chapter of the Association for Computational Linguistics (EACL) 2024. Two models: i) Linear Support Vector Classifier (LinearSVC) and ii) Ensemble of ML classifiers (k Nearest Neighbour (kNN), Stochastic Gradient Descent (SGD), Logistic Regression (LR), LinearSVC, and Random Forest Classifier (RFC)) with hard voting, are trained individually with the features obtained by the concatenation of TfidfVectorizer and CountVectorizer of word and character n-grams, for SA in code-mixed Tamil and Tulu texts. Gridsearch method is employed to get the best hyperparameter values for the proposed classifiers. Among the two models, the proposed Ensemble models achieved macro F1 scores of 0.260 and 0.550 for Tamil and Tulu languages respectively.

1 Introduction

SA is the process of examining opinions, emotions, and reviews to recognise the sentiments expressed by the users regarding a topic, movie, song, product, etc., available on online platforms (Chakravarthi et al., 2021). This user-generated content is used by businesses and individuals to gain knowledge and make well-informed decisions regarding their content (Mahadzir et al., 2021). The user sentiments are usually available in code-mixed language where words and/or sub-words belong to more than one language. Processing the code-mixed user-generated content to develop SA models poses a significant challenge (Hegde and Shashirekha, 2022). This is especially notable when addressing SA in low-resource languages such as Tulu, Tamil, Malayalam, and Telugu (Ka et al., 2023).

To address the challenges of detecting SA in user-generated code-mixed low-resource languages, in this paper, we - team MUCS, describe ML models submitted to the shared task "Sentiment Analysis in Tamil and Tulu - DravidianLangTech@EACL-2024" (S. K. et al., 2024). This shared task is modeled as a multi-class text classification problem with two distinct models: i) LinearSVC and ii) Ensemble of ML classifiers (kNN, SGD, LR, LinearSVC, and RFC) with hard voting, trained individually with the features obtained by the concatenation of TfidfVectorizer[1] and CountVectorizer[2] of word and character n-grams, for SA in code-mixed Tamil and Tulu texts. In addition, the Gridsearch method is used to find the ideal values for the hyperparameters of these classifiers.

The rest of the paper is organized as follows: while Section 2 describes the related works of SA, Section 3 focuses on the description of the models submitted to the shared task followed by the experiments and results in Section 4. The conclusion and future works are included in Section 5.

2 Related Work

Several ML models are experimented with various features for SA of user-generated content in code-mixed low-resource languages (Hegde et al., 2023a). Some of the relevant works are outlined

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

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

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 257–261
March 22, 2024 ©2024 Association for Computational Linguistics

below:
Ponnusamy et al. (2023) proposed ML models (LR, Multinomial Naive Bayes (MNB), and LinearSVC) trained with Term Frequency-Inverse Document Frequency (TF-IDF) of word unigrams for SA in Tamil and Tulu languages. Their proposed LR, MNB, and LinearSVC models obtained macro F1 scores of 0.43, 0.20, 0.41 and 0.51, 0.25, 0.49 for Tamil and Tulu languages respectively. Coelho et al. (2023) used ML models (LinearSVC, LR, and an Ensemble model (LR, Decision Tree (DT), and Support Vector Machine (SVM)) with hard voting), trained with TF-IDF of word unigrams achieving macro F1 scores of 0.189 and 0.508 for code-mixed Tamil and Tulu texts respectively. Ehsan et al. (2023) implemented Bidirectional Long Short-Term Memory (BiLSTM) networks for SA of code-mixed Tamil and Tulu text, utilizing contextualized Elmo representations and obtained macro F1 scores of 0.2877 and 0.5133 for Tamil and Tulu code-mixed datasets respectively. Hegde et al. (2023b) implemented three models: i) n-gramsSA (LinearSVC trained with TF-IDF, ii) EmbeddingsSA (LinearSVC trained with fastText and Byte Pair embeddings), and iii) BERTSA (a transformer classifier trained with Bidirectional Encoder Representations from Transformer (BERT) embeddings) and obtained macro F1 scores of 0.26 and 0.53 for Tamil using BERTSA model and for Tulu using EmbeddingsSA model respectively.

Puranik et al. (2021) fine-tuned: the Universal Language Model Fine-Tuning (ULMFiT) and multilingual BERT (mBERT) models, the two pre-trained models for SA in code-mixed Kannada, Tamil, and Malayalam and obtained macro F1 scores of 0.63, 0.65, and 0.70, respectively. Garain et al. (2020) presented the Support Vector Regression model (SVR) model with Grid Search approach, trained with TF-IDF of word unigrams and GloVe word vector features, for Hindi code-mixed sentences and obtained a macro F1 score of 0.662.

The related work reveals that the performances of SA models for code-mixed low-resource languages are still low, indicating the scope for developing models to improve the performance further.

3 Methodology

The proposed methodology for SA in code-mixed Tamil and Tulu texts include: Pre-processing, Feature Extraction (FE), and Classifier Construction. The framework of the proposed methodology is shown in Figure 1 and the steps are explained below:

Language	Sample Text	Label
Tamil	நம்ப நடே நாசாமா தான் போச்சி	Negative
	ennaya trailer Ku mudi Ellam nikkudhu... Vera	Positive
Tulu	Tulu panda enku masth ista i love tulu tulunadu	Positive
	Bega 2 nd part padle	Neutral

Table 1: Sample code-mixed Tamil and Tulu comments along with the corresponding labels

3.1 Pre-processing

During pre-processing, punctuation, digits, user mentions, and hashtags are removed to clean the text. English stopwords available at Natural Language Tool Kit (NLTK)[3] library and Tamil[4] stopwords from a GitHub repository are utilized as references for filtering out English and Tamil stopwords in Tamil dataset respectively and English stopwords from Tulu text. As the given dataset is code-mixed, English words will be present in the dataset. Additionally, emojis are converted to English text using the demoji library. The resulting pre-processed text is then used for FE.

3.2 Feature Extraction

FE involves extracting distinguishing characteristics from the given data and the performance of the classifiers depends on the quality of the features. n-grams refers to ’n’ consecutive lexical units where th lexical units are words or characters. These word/character n-grams capture the local context by following sequential patterns, facilitating a deeper understanding of relationships between words/characters (Bahdanau et al., 2014). Choosing the right value for ’n’ in n-grams is crucial for capturing contextual relationships between the words/characters and the selection of ’n’ depends on the desired level of context. While the higher ’n’ value provide more extensive context at the cost of increased computational complexity, lower ’n’ value focus on shorter and more immediate relationships (Nagao and Mori, 1994). In this work, word n-grams in the range (1, 3) are obtained.

As the given Tamil and Tulu dataset includes text in native script, they are romanized using libindic[5] library and character n-grams in the range (1, 5)

[3]https://pythonspot.com/nltk-stop-words/
[4]https://gist.github.com/arulrajnet/e82a5a331f78a5cc9b6d372df13a919c
[5]https://github.com/libindic/indic-trans

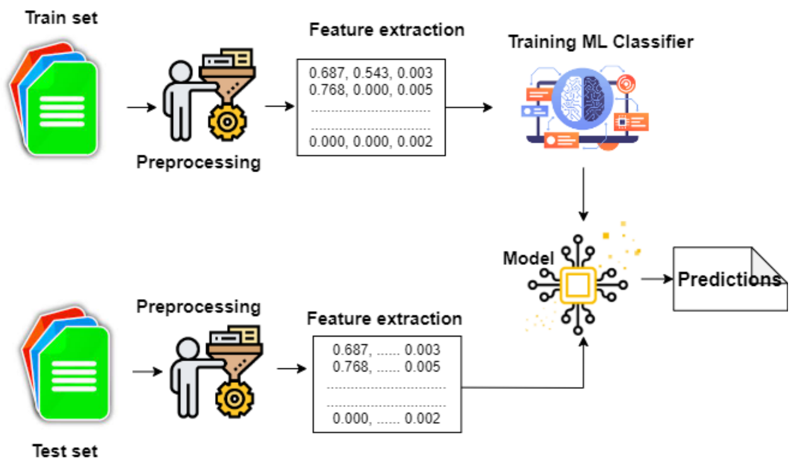

Figure 1: Framework of the proposed methodology

Classifier	Hyperparameters and values
LinearSVC	class_weight = balanced, C = 1
RFC	criterion = gini, max_depth = 8, max_features = log2, n_estimators = 200, class_weight = balanced
LR	C = 3, penalty = l2, class_weight = balanced,
kNN	n_neighbors = 7, p = 2, weights = distance
SGD	class_weight = balanced, loss = log, penalty = elasticnet, alpha = 4, l1_ratio = 0.1

Table 2: Hyperparameter values obtained from Gridsearch algorithm

	Labels	Tamil	Tulu
Train set	**Positive**	20,070	3,352
	Negative	4,271	698
	Unknown state	5,628	1,854
	Mixed Feeling	4,020	1,041
Dev set	**Positive**	2,257	231
	Negative	480	55
	Unknown state	611	124
	Mixed Feeling	438	90

Table 3: Class-wise distribution of Tamil and Tulu datasets

are obtained from the romanized Tamil and Tulu texts. The word and character n-grams are vectorized using TfidfVectorizer and CountVectorizer and the resulting vectors are concatenated to train the learning models. The sample code-mixed Tamil and Tulu comments along with their corresponding labels are shown in Table 1.

3.3 Classifier Construction

This work utilizes LinearSVC and an Ensemble of ML classifiers (RFC, LR, kNN, SGD), for SA in code-mixed Tamil and Tulu texts. A brief description of the classifiers is given below:

- LinearSVC - uses a linear kernel function, which calculates the dot product between data points in the feature space. This makes it particularly effective for high-dimensional datasets and situations where the relationship between features and classes is approximately linear (Hegde et al., 2023b).

- Ensemble - is a method of generating a new classifier using a pool of classifiers such that the strength of one classifier is used to overcome the weakness of other classifier, with the objective of obtaining a better classification performance (Hegde and Shashirekha, 2021). When compared to the performance of individual baseline classifier in the ensemble, this configuration of several classifiers will perform better. As the Ensemble model uses more than one classifier to predict class labels for an unlabeled sample, it is also called a voting classifier.

Optimal hyperparameter values are obtained by employing gridsearch[6] algorithm and the hyperparameters and their values used for the classifiers are shown in Table 2.

[6]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Language	Model	Precision	Recall	Macro F1-score
Tamil	**LinearSVC**	0.284	0.263	0.252
	Ensemble	0.291	0.279	**0.260**
Tulu	**LinearSVC**	0.546	0.546	0.546
	Ensemble	0.548	0.554	**0.550**

Table 4: Performance of the proposed models for code-mixed Tamil and Tulu texts

4 Experiments and Results

Code-mixed Tamil and Tulu SA datasets are provided by the organizers of the shared task and statistics of the datasets are shown in Table 3. Using these datasets, several experiments were conducted by employing various FE techniques and classifiers. Combination of features and classifiers which gave good performance on the Development (Dev) sets are used to train the proposed models. The proposed models are evaluated on the Test set and the predictions are assessed by the organizers based on macro F1-score for the final evaluation and ranking. Performance of the proposed models for both Tamil and Tulu datasets are shown in Table 4. Ensemble models outperformed the LinearSVC models obtaining macro F1 scores of 0.260 and 0.550 securing 1st and 2nd ranks in the shared task for Tamil and Tulu languages respectively. Though class_weight is set to 'balanced' for both the classifiers, the extreme data imbalance in the given datasets has lead to low macro F1 scores.

4.1 Error Analysis

The confusion matrix reveals the percentage of classification error obtained by the learning model. As the Ensemble models performed better than the LinearSVC model, confusion matrix is shown for Ensemble model. The confusion matrix for code-mixed Tamil texts is shown in Figure 2. The results reveal that the Ensemble model exhibits a relatively weak True Positive Rate (TPR) of 38.61% for the 'Mixed Feelings' class (though it is the highest rate among the TPRs obtained across all the classes) indicating lower performance of the proposed model. This may be due to extreme data imbalance in the training set. Additionally, the model faces difficulty in identifying 'Unknown state' class by exhibiting a notably low TPR of 13% for this class, as the learning model fails to distinguish between 'Unknown state' and 'Mixed Feelings' sentiments.

The confusion matrix for code-mixed Tulu texts is shown in Figure 3. The results reveal that the Ensemble model exhibits a good performance with a TPR of 79.44% for the 'Positive' class. However, the model fails to distinguish between 'Mixed Feelings' and 'Neutral' sentiments, as reflected in a lower TPR of 37.14% for the 'Mixed Feelings' class.

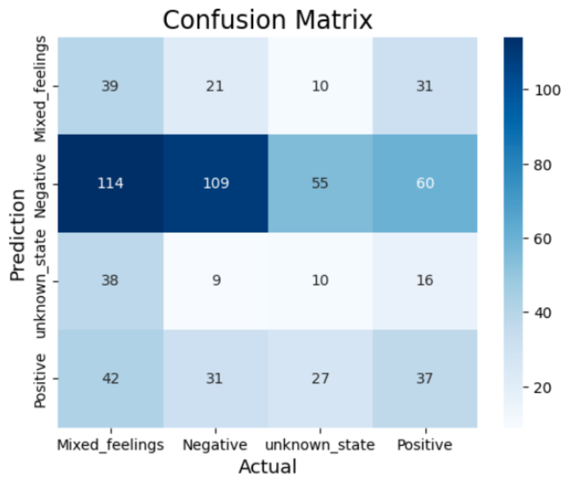

Figure 2: Confusion matrix of the proposed Ensemble model for code-mixed Tamil text

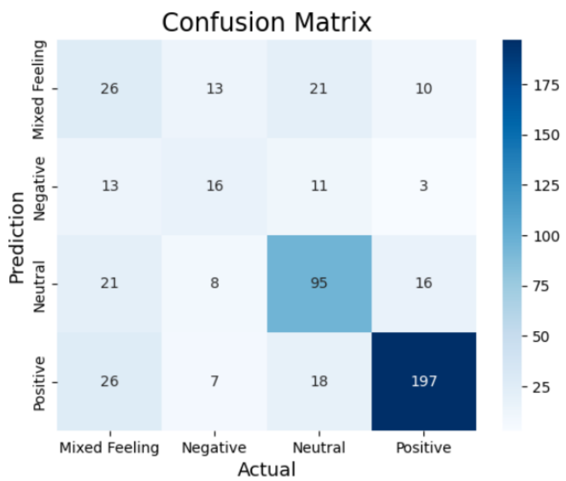

Figure 3: Confusion matrix of the proposed Ensemble model for code-mixed Tulu text

5 Conclusion and Future Work

This paper describes the models submitted by our team MUCS to "Sentiment Analysis in Tamil and Tulu - DravidianLangTech@EACL-2024" shared task. The proposed methodology consists of using LinearSVC and Ensemble of ML classifiers with hard voting, trained individually with the features obtained by the concatenation of TfidfVectorizer and CountVectorizer of word and character n-grams. Further, in order to get the optimal hyperparameter values for these classifiers, the Gridsearch method is used during training. The proposed Ensemble models exhibited macro F1 scores of 0.260 and 0.550 securing 1st and 2nd ranks in the shared task for Tamil and Tulu languages respectively. Suitable oversampling or text augmentation techniques will be explored further to improve the performance of the proposed models.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. In *arXiv preprint arXiv:1409.0473*.

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, G Kavya, 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*, pages 282–287.

Toqeer Ehsan, Amina Tehseen, Kengatharaiyer Sarveswaran, and Amjad Ali. 2023. Sentiment Analysis of Code-Mixed Tamil and Tulu by Training Contextualized ELMo Representations. In *RANLP'2023*, page 152.

Avishek Garain, Sainik Kumar Mahata, and Dipankar Das. 2020. JUNLP@ SemEval-2020 Task 9: Sentiment Analysis of Hindi-English Code Mixed Data using Grid Search Cross Validation.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, SK Lavanya, Durairaj Thenmozhi, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023a. 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. 2023b. 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. 2021. Urdu Fake News Detection Using Ensemble of Machine Learning Models. In *CEUR Workshop Proceedings*, pages 132–141.

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

Rachana Ka, Prajnashree Mb, Asha Hegdec, and HL Shashirekhad. 2023. MUCS@ DravidianLangTech2023: Sentiment Analysis in Code-mixed Tamil and Tulu Texts using fastText. In *RANLP'2023*, page 258.

Nurul Husna Mahadzir et al. 2021. Sentiment Analysis of Code-Mixed Text: A Review. In *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, volume 12, pages 2469–2478.

Makoto Nagao and Shinsuke Mori. 1994. A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Kishore Kumar Ponnusamy, Charmathi Rajkumar, Prasanna Kumar Kumaresan, Elizabeth Sherly, 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*, pages 211–216.

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

Lavanya S. K., Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, and Rajkumar Charmathi Kumaresan, Prasanna Kumar. 2024. Overview of Second Shared Task on 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.