MUCSD@DravidianLangTech2023: Predicting Sentiment in Social Media Text using Machine Learning Techniques

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

Abstract

Social media users utilize online platforms to express their thoughts, sentiments, and views through posts/comments. Identifying such sentiments expressed in reviews or comments on a given concept/topic is known as Sentiment Analysis (SA). SA has considerable applications including customer service, social media monitoring, product reviews analysis, and so on. Content creators, bloggers, and researchers can evaluate public opinions, obtain feedback, and make informed choices by analyzing reviews and comments on social media platforms. Online users often express their sentiments using mixing words/scripts of more than one language leading to code-mixed texts. Analysing the code-mixed text to predict sentiments is challenging and lack of resources for code-mixed low-resource languages enhances the complexity further due to complexities of code-mixed texts.

To address the challenges of predicting sentiments in code-mixed low-resource languages, in this paper, we - team MUCSD, describe Machine Learning (ML) models submitted to ”Sentiment Analysis in Tamil and Tulu” shared task at DravidianLangTech@RANLP 2023. The proposed methodology makes use of ML models: i) Linear Support Vector Classifier (LinearSVC), ii) Logistic Regression (LR), and iii) Ensemble model (LR, Decision Tree (DT), and Support Vector Machine (SVM)) with hard voting, trained with Term Frequency-Inverse Document Frequency (TF-IDF) of word unigrams, to perform SA in Tamil and Tulu languages. Among these models, LinearSVC model performed better with macro F1-scores of 0.189 and 0.508 and obtained 8th and 9th rank for Tamil and Tulu code-mixed texts respectively.

1 Introduction

Social media usage is increasing continuously due to the ease of use to share reviews even for those who are not familiar about the technologies (Asghar et al., 2015; Bharathi and Agnusimmaculate Silvia, 2021; Bharathi and Varsha, 2022; Swaminathan et al., 2022; Anita and Subalalitha, 2019; Subalalitha, 2019; Sakuntharaj and Mahesan, 2016, 2017, 2021). This has led to the increased user-generated text of opinions or reviews, comments, and posts on social media like Twitter, Facebook, YouTube, etc (Yue et al., 2019). Understanding the users' comments on topics/events in social media allows for more informed decision-making. It helps the content creators like YouTubers to evaluate the emotional impact of their videos on viewers (Chakravarthi et al., 2022a,b; Chakravarthi, 2023). For example, while positive comments about a video indicate satisfaction, negative comments indicate the need for improvement (Asghar et al., 2015).

SA is the task of analysing the reviews, opinions or comments to identify their polarity. This analysis provides insights into public opinions, ideas, and statements, which can be valuable for bloggers, researchers, and even individuals (Hussein, 2018; Thavareesan and Mahesan, 2019, 2020a,b). Social media posts/comments such as YouTube comments, Facebook posts, and tweets, often contain slang, misspellings, contractions, etc. which may impair the ability of the learning models to recognize patterns in the social media content and produce precise predictions. As there is no restriction on the use of language in social media, users often use more than one language to write comments on social media platforms. They may even use more than one script to key in the comments leading to code-mixed data.

Majority of the SA works focus on high-resource languages like English and Spanish giving less importance for low-resource languages and code-mixed low-resource languages (Hegde and Shashirekha, 2021). To promote the SA research

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 282–287, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_041

work on low-resource languages, in this paper, we - team MUCSD, describe the ML models submitted to the shared task on ”Sentiment Analysis in Tamil and Tulu” at DravidianLangTech@RANLP 2023[1] (Hegde et al., 2023). The aim of the shared task is to classify the comments in code-mixed Tamil and Tulu languages, into one of the four categories: Positive, Neutral/Unknown state, Mixed Feelings, and Negative. The proposed methodology makes use of LinearSVC, LR, and Ensemble of classifiers (LR, DT, and SVM) with hard voting, trained with TF-IDF of word unigrams to predict the sentiment of the given text.

The rest of the paper is structured as follows: The related work is briefly described in Section 2. Section 3 describes the methodology, while Section 4 discusses experiments with the results. The study concludes with future work in Section 5.

2 Related Work

Researchers have developed several ML approaches to handle monolingual and code-mixed texts for SA. A brief description of few of the relevant works are given below:

Das and Chakraborty (2018) proposed LinearSVM using TF-IDF along with Next Word Negation (NWN) for sentiment classification of three different datasets (Movie Review Dataset, Product Review Dataset, and SMS Spam Dataset). They obtained accuracies of 89.91%, 88.86%, and 96.83% on IMDB review datasets, Amazon product review, and SMS spam datasets respectively. To perform SA in Code-Mixed Dravidian languages (Tamil and Malayalam) and English language, Hegde and Shashirekha (2022) used Long Short Term Memory (LSTM) model trained with Dynamic Meta Embedding (DME) and obtained F1-scores of 0.36, 0.74, and 0.37 for Tamil, English, and Malayalam languages respectively. Balouchzahi and Shashirekha (2020) proposed Hybrid Voting Classifier (HVC) using char and word n-grams in the range (1, 5) and (1, 3) respectively and word embeddings, as features to train Multi-Layer Perceptron (MLP) and Multinomial Naive Bayes (MNB) classifiers and sub-words embeddings to train Bidirectional Long Short Term (BiLSTM) classifier. With majority voting, their HVC model obtained a weighted F1-score of 0.62 and 0.68 on Tamil-English and Malayalam-English language pairs respectively.

To solve the Arabic sentiment classification problem, Elgeldawi et al. (2021) used various hyperparameter tuning techniques on six ML algorithms (LR, SVM, DT, Ridge Classifier, Random Forest (RF), and Naive Bayes (NB)) trained using TF-IDF of word unigrams. All algorithms were experimented with and without the hyperparameter tuning process and among all the algorithms, they obtained highest accuracy for SVM with an accuracy of 95.62% using Bayesian Optimization. Hande et al. (2020) proposed multitask learning for Offensive Language Detection (OLD) and SA and experimented on the dataset Kannada Code-Mixed Dataset (KanCMD) created by scraping the YouTube comments. ML algorithms (DT, LR, SVM, MNB, k-Nearest Neighbors (k-NN), and RF) are trained with TF-IDF of n-grams in the range (1, 3) for both OLD and SA tasks. The RF model outperformed other models with macro F1-scores of 0.59 and 0.66 for SA and OLD respectively. Poornima and Priya (2020) proposed ML classifiers (MNB, SVM, and LR) trained with word bigrams for SA of Malayalam Tweets and obtained an accuracy of 86.23% for LR classifier.

In spite of several approaches explored for SA, the performance of many of the approaches are still low for low-resource languages, indicating the need to explore models for SA in to improve the performance.

3 Methodology

The proposed methodology for SA in code-mixed Tamil and Tulu (Dravidian languages) includes three major steps: Pre-processing, Feature Extraction, and Classifier Construction. The framework of the proposed methodology is shown in Figure 1 and the steps are briefly explained below:

3.1 Pre-processing

The procedure of cleaning text data with the goal of enhancing the classifier’s performance is known as Pre-processing. In this procedure, punctuation marks, digits, stopwords, and extra spaces are removed and English text is converted to lowercase. English stopwords[2] list available in Natural Language Toolkit (NLTK) library and Tamil stopwords[3] list available in github are used as references to remove the stopwords. Sentiments are

[1] https://codalab.lisn.upsaclay.fr/competitions/11095

[2] https://www.nltk.org/nltk_data/

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

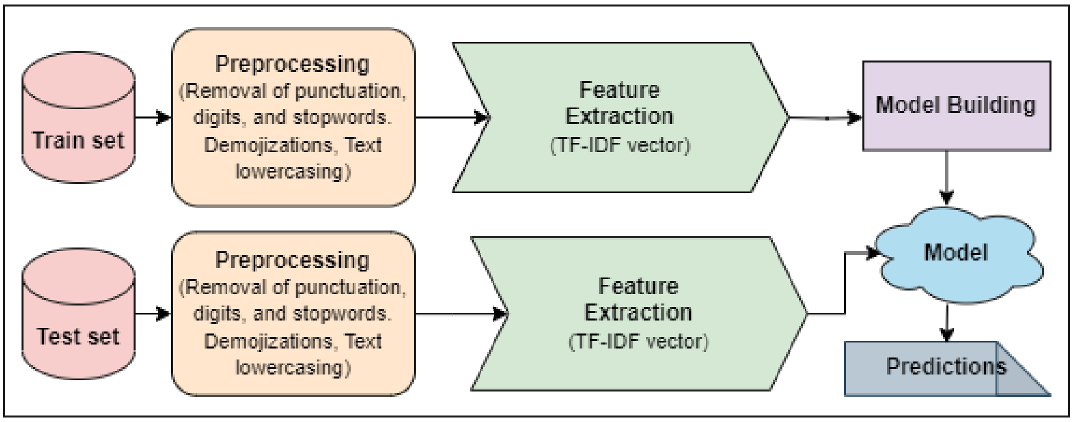
Train set
Preprocessing
(Removal of punctuation, digits, and stopwords. Demojizations, Text lowercasing)
Feature Extraction
(TF-IDF vector)
Model Building
Model
Test set
Preprocessing
(Removal of punctuation, digits, and stopwords. Demojizations, Text lowercasing)
Feature Extraction
(TF-IDF vector)
Predictions

Figure 1: The proposed framework of ML classifiers

Dataset:	Tamil-English			Tulu-English		
Classes	**Train set**	**Dev set**	**Test set**	**Train set**	**Dev set**	**Test set**
Positive	20070	2257	338	3118	369	344
Negative	4271	480	101	646	90	60
Neutral / Unknown state	5628	611	137	1719	202	197
Mixed feelings	4020	438	73	974	120	107
Total	33989	3786	649	6457	781	781

Table 1: Class-wise distribution of Tamil-English and Tulu-English dataset

	Tamil-English		Tulu-English	
Classifier	**Dev set**	**Test set**	**Dev set**	**Test set**
LR	0.444	0.117	0.516	0.442
LinearSVC	**0.464**	**0.189**	**0.555**	**0.508**
Ensemble model	0.437	0.103	0.509	0.461

Table 2: Macro F1-scores of the proposed models

sometimes expressed through emojis. So instead of removing the emojis, they are converted into text to capture the useful information for SA.

3.2 Feature Extraction

The process of extracting the features from the dataset is known as Feature Extraction. In the proposed work, TF-IDF of word unigrams are obtained using TfidfVectorizer[4] from the scikit-learn library to train the model. 12,515 and 62,516 word unigrams are extracted for Tulu and Tamil datasets respectively.

3.3 Classifier Construction

The three learning models, namely: LinearSVC, LR, and Ensemble (LR, DT, and SVM) classifier with hard voting, are trained using TF-IDF of word unigrams to perform SA. LinearSVC learns the decision boundary by separating different classes and it is suitable for linearly separable problems. The LR model provide a simple approach that predicts the probability that an input belongs to a certain class. LinearSVC is used to find a hyperplane that separates data points of different classes in a way that maximizes the margin between them. To improve the accuracy and robustness of predictions, specifically if the individual classifiers have various weaknesses or strengths, the ensemble techniques can be applied. In the proposed methodology, an Ensemble of ML classifiers (LR, DT, and SVM) with hard voting is used.

4 Experiments and Result

The statistics of code-mixed Tamil (Chakravarthi et al., 2020) and Tulu (Hegde et al., 2022) datasets shared by the organisers of the shared task in shown in Table 1. It can be observed that the both datasets are imbalanced.

The performance of the proposed model for Development (Dev) set and Test sets for both the languages are shown in Table 2. The results illustrate that LinearSVC model outperformed the other two models by achieving F1-scores of 0.189 and 0.508

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

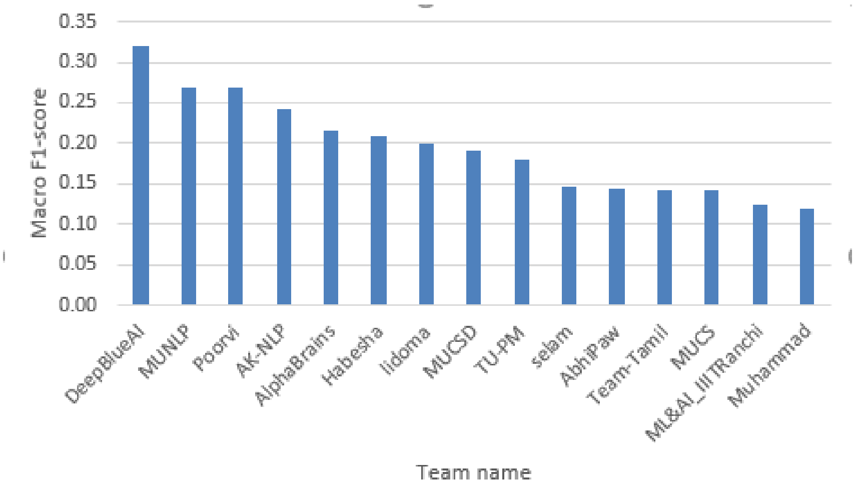

Figure 2: Comparison of macro F1-scores of the LinearSVC model with other teams (participants') for Tamil-English Dataset

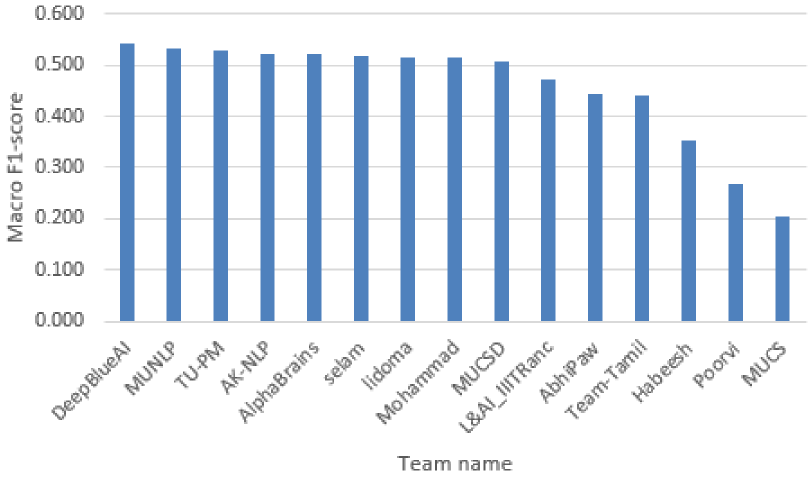

Figure 3: Comparison of macro F1-scores of the LinearSVC model with other teams (participants') for Tulu-English Dataset

Tulu Text	English Translation	Actual Label	Predicted Label	Remarks
ajji baari joruller....jagrthed patherodu	Grandmother is very strict. Should be careful while talking	Mixed Feeling	Positive	The words like "baari", and "ajji" are associated with sentences of 'Positive' class in the Train set. Hence, this sample is classified as 'Positive'.
Ipl,gangnam style,,pili ,,avu matha common adu ippundu,,,	Ipl, Gangam style, tiger, all those things are very common	Neutral	Positive	The model has associated "Ipl" and "Pili" words with 'Positive' class and hence has classified the comment as 'Positive'.
ಕುಸಾಸ್ ಇತ್ತುಂಡು. ಪ್ರಾಕ್ಟಿಸ್ ಕಮ್ಮಿ. ಒಟ್ಟುಗ್ ಓಕೆ.	It was funny, they have less practice, totally okay	Mixed Feeling	Neutral	The words "ಕಮ್ಮಿ", is associated with 'Neutral' sentiment and hence, the model has the comment as "Neutral" class.

Table 3: Samples of misclassified Tulu Test set with their English translation, actual and predicted (using LinearSVC model) labels and remarks

for code-mixed Tamil and Tulu datasets respectively. The comparison of the macro F1-scores of other participants' models with our proposed model for Tamil and Tulu Datasets are shown in Figure 2 and 3 respectively. The misclassified samples of the Tulu Test test along with actual and predicted labels and remarks, for LinearSVC model are shown in Table 3.

5 Conclusion

In this paper, our team MUCSD describes the three proposed models, namely: LR, LinearSVC, and an Ensemble (LR, DT, and SVM) classifier with hard voting, trained with TF-IDF of word unigrams, to perform SA of Tulu and Tamil code-mixed texts. These models are submitted to the "Sentiment Analysis in Tamil and Tulu" shared task at DravidianLangTech@RANLP 2023. Among the proposed models, LinearSVC achieved F1-scores of 0.189 and 0.508 and secured 8th and 9th rank for Tamil and Tulu texts respectively. Different feature sets and different classifications will be investigated to improve the performance of the proposed models in the future.

References

R Anita and CN Subalalitha. 2019. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

Muhammad Zubair Asghar, Shakeel Ahmad, Afsana Marwat, and Fazal Masud Kundi. 2015. Sentiment Analysis on YouTube: A Brief Survey. *arXiv preprint arXiv:1511.09142*.

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

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, 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, 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.

Bijoyan Das and Sarit Chakraborty. 2018. An Improved Text Sentiment Classification Model using TF-IDF and Next Word Negation. *arXiv preprint arXiv:1806.06407*.

Enas Elgeldawi, Awny Sayed, Ahmed R Galal, and Alaa M Zaki. 2021. Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. In *Informatics*, volume 8, page 79. Multidisciplinary Digital Publishing Institute.

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

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, 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 and Hosahalli Lakshmaiah Shashirekha. 2021. Urdu Fake News Detection Using Ensemble of Machine Learning Models.

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.

Doaa Mohey El-Din Mohamed Hussein. 2018. A Survey on Sentiment Analysis Challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338.

A Poornima and K Sathiya Priya. 2020. A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques. In *2020 6th international conference on advanced computing and communication systems (ICACCS)*, pages 493–496. IEEE.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2016. A novel hybrid approach to detect and correct spelling in Tamil text. In *2016 IEEE international conference on information and automation for sustainability (ICIAfS)*, pages 1–6. IEEE.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words. In *2017 IEEE international conference on industrial and information systems (ICIIS)*, pages 1–5. IEEE.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47. IEEE.

CN Subalalitha. 2019. Information extraction framework for Kurunthogai. *Sādhanā*, 44(7):156.

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.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on industrial and information systems (ICIIS)*, pages 320–325. IEEE.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts. In *2020 Moratuwa engineering research conference (MERCon)*, pages 272–276. IEEE.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based part of speech tagging in tamil texts. In *2020 IEEE 15th International conference on industrial and information systems (ICIIS)*, pages 478–482. IEEE.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A Survey of Sentiment Analysis from Social Media Data. *Knowledge and Information Systems*, 60:617–663.