Code_Maker@DravidianLangTech-EACL 2024: Sentiment Analysis in Code-Mixed Tamil using Machine Learning Techniques

Kogilavani Shanmugavadivel[1]**,Sowbharanika Janani J S**[1]**,**
Navbila K[1]**, Malliga Subramanian**[1]
Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv}@gmail.com
{sowbharanikajananijs.22aid,navbilak.22aid}@kongu.edu
{mallinishanth72}@gmail.com

Abstract

The rising importance of sentiment analysis in online community research is addressed in our project, which focuses on the surge of code-mixed writing in multilingual social media. Targeting sentiments in texts combining Tamil and English, our supervised learning approach, particularly the Decision Tree algorithm, proves essential for effective sentiment classification. Notably, Decision Tree(accuracy: 0.99, macro average F1 score : 0.39), Random Forest exhibit high accuracy (accuracy: 0.99, macro average F1 score : 0.35), SVM (accuracy: 0.78, macro average F1 score : 0.68), Logistic Regression (accuracy: 0.75, macro average F1 score : 0.62), KNN (accuracy: 0.73, macro average F1 score : 0.26) also demonstrate commendable results. These findings showcase the project's efficacy, offering promise for linguistic research and technological advancements. Securing the 8th rank emphasizes its recognition in the field.[1]

1 Introduction

Sentiment analysis is an essential tool in the ever-changing world of social media for understanding the subtleties of user expressions. Sentiment analysis algorithms, which were previously designed for high-resource languages and single utterances, now confront additional difficulties in the age of multilingual societies and code-mixed writing. The expanding importance of sentiment analysis is discussed in this research, especially in light of the prevalence of code-mixed Tamil-English statements on social media platforms. The conventional supervised learning approaches, relying on annotated data, encounter limitations when applied to code-mixed languages. Notably, in this multilingual environment, lexical characteristics like word dictionaries and parts of speech labelling perform less than ideal. To tackle these problems, we concentrate our study on sentiment analysis in code-mixed Tamil-English. Using the Decision Tree technique, our strategy's key component offers a dependable way to classify emotions in this peculiar language fusion. This project not only showcases exceptional accuracy through detailed metrics like precision, recall, and F1-score but also introduces a substantial corpus for under-resourced code-mixed Tanglish. Marked by a high inter-annotator agreement, this dataset stands as a valuable resource for researchers exploring sentiment analysis and linguistic phenomena in code-mixed environments. Our project stands at the intersection of sentiment analysis, machine learning, and code-mixed language research. Its contributions extend beyond accurate sentiment classification, serving as a foundational resource for future investigations in the dynamic landscape of multilingual social media expressions.

Keywords: Sentiment analysis, Code-mixed writing, Decision tree algorithm, Machine learning, Tamil - English dataset

2 Literature Survey

A sentiment analysis of COVID-19 vaccine-related tweets on English-language Twitter is conducted in Liu and Liu (2021). They found that nearly 43 percent of the more than 2.6 million tweets they analysed were positive, 27 percent were neutral, and 30 percent were negative. Based on the research findings that these opinions varied by region and changed over time, health officials may adjust their efforts to educate people about vaccines. The study claims that sentiment analysis on Twitter can be utilised to learn more about the public's perceptions on vaccinations.

In social media especially, sentiment analysis of comments on photos or videos is crucial for decision-making. On social media, comments, however, are often multilingual and lack annotations for languages with low resource availability,

[1] S. K. et al. (2024)

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 129–133

like Tamil 15,744 Tamil and English YouTube comment threads were code-switched, and sentiment analysis was done to establish a gold standard corpus. Results in F-Score, Precision, and Recall were obtained from Chakravarthi et al. (2020) with good inter-annotator agreement.

Raveendirarasa and Amalraj (2020) examines sentiment analysis of texts that move across codes on social networking sites such as Facebook. It suggests a method for applying natural language processing to recognise user behavioural patterns. Clustering-based pre-processing and hyperparameter optimisations are used by the system, which primarily targets Facebook users in Sri Lanka. The model's accuracy was 75 percent, and results for huge and uncommon words were enhanced by sub-word-level LSTM.

Within the field of Natural Language Processing, sentiment analysis (SA) examines user sentiments from internet reviews. It helps consumers comprehend and organise their travels, and search engines depend on it. Devika et al. (2016) focuses on four primary approaches to sentiment analysis: machine learning, semantic analysis, rule-based, and lexicon-based approaches.

In this study, Shanmugavadivel et al. (2022) uses transfer learning, hybrid deep learning, deep learning, and classic machine learning models to investigate the effects of pre-processing Tamil code-mixed data. The goal of the study is to eliminate from the data any emojis, punctuation, symbols, numerals, and repeated letters. With pre-processed Tamil code-mixed data, the hybrid deep learning model CNN+BiLSTM outperforms the others, with an accuracy of 0.66. The study evaluates these models' performance against the most advanced techniques, such as logistic regression, random forest, IndicBERT, multinomial Naive Bayes, and linear support vector classification. Future work should concentrate on multimodal data sets and context-based algorithms to improve the accuracy of sentiment analysis on social media data.

The dynamic field of sentiment analysis (SA) examines user opinions as they are conveyed in written language. It facilitates the gathering of input for manufacturers, governments, and companies. The limitations and future directions of research on implicit aspect extraction for SA have been evaluated in Ganganwar and Rajalakshmi (2019). Grammatical errors, double implicit problems, and semantic concept-centric aspect level sentiment analysis are some of the topics that should be the focus of future research.

Sentiment analysis is vital in the fast-paced world of the internet, especially on social media platforms like Twitter. A method for classifying customer evaluations as positive, negative, or neutral is presented in Rakshitha et al. (2021) utilising text blobs that are retrieved from Twitter APIs. This helps customers choose the best products more effectively, and it also allows businesses to modify as needed in response to customer input

Alshamsi et al. (2020) investigates sentiment analysis utilising many machine learning algorithms and a dataset of tweets. The research examined 16 scholarly works. In summary, the project excels in sentiment analysis, utilizing the Decision Tree algorithm with exceptional accuracy. The comprehensive classification report emphasizes crucial metrics, illuminating the model's robust performance. Beyond sentiment analysis, it stands as a pivotal resource for code-mixed research, marked by a high inter-annotator agreement, showcasing adaptability in exploring diverse linguistic phenomena in code-mixed Tanglish. While addressing concerns like overfitting, the project's strong foundation positions it as a commendable contribution to sentiment analysis in code-mixed languages. Exploration of alternative algorithms holds promise for further enhancement. Concerning text categorization and analysis on Twitter, assessing several classifiers for both balanced and unbalanced datasets. On balanced datasets, the ID3 and Naive Byes classifiers demonstrated greater accuracy levels; on unbalanced datasets, K-NN, Decision Tree, Random Forest, and Random Tree outperformed the others. The study emphasises how crucial it is to comprehend the data produced by social media platforms in order to enhance goods, services, and research.

Validating social media content requires sentiment analysis, especially when managing comments in multiple languages. Sripriya and Divya suggests a model that codes input data based on word frequency and applies a multiclass classification algorithm. The model receives an average weighted F1 score of 0.35 from the Dravidian Codemix dataset. Further learning techniques could enhance the functionality of the model.

Multilingual sentiment analysis is critical for recommendation systems, sentiment summarization, and opinion retrieval. Existing solutions include

machine translation and bilingual dictionary methods. Thilagavathi and Krishnakumari (2016) employs supervised and unsupervised algorithms as well as Tamil language reviews that have been translated into English. The essay provides a product aspect ranking methodology for identifying essential characteristics from online consumer reviews, increasing usability, and influencing consumer opinions.[2]

3 Problem and System Description

The objective of the sentiment analysis project is to automatically analyse and categorize the sentiment expressed in a given text. The sentiment is classified into categories such as "Positive", "Negative", "Mixed feelings" or "Unknown State". The objective of the research is to use machine learning, namely the Decision Tree approach, to consistently predict the emotion of textual data.

4 Dataset Description

The training dataset comprises 33,989 code-mixed Tamil-English language samples encompassing a wide range of themes, with emotion labels such as Positive, Negative, Mixed Feelings, and Unknown State. X-train-tfdf matrix was created by TF-IDF vectorization of the text data. Achieving 99.97 percent accuracy on the training data, the Decision Tree classifier showed remarkable accuracy. The test dataset comprises 649 samples of code-mixed Tamil-English language. Each sample includes a text segment unseen during training, serving to assess the model's generalization to new data. The dataset includes predicted sentiment labels generated by the trained Decision Tree classifier, indicating the model's predictions for the sentiments expressed in the text segments.[3]

Dataset	No. of Comments
Train	33,989
Validation	3,786
Test	649

Table 1: Dataset Description

5 Predictions on Test Data

Text Segments: There are 649 text segments in the test sample that are code-mixed Tamil and English. Prediction Labels: The trained Decision Tree classifier was used to generate predicted sentiment labels, which classified each text segment into sentiments like Positive, Negative, Mixed Feelings, or

Text	Category
Vera level..Waiting for FDFS	Positive
Do or Die	Negative
it not vijay hair,setup paaa	Mixed feelings
598K left for 15M views	unknown state

Table 2: Text and Category examples

Class	Train	Dev
Positive	15,203	2,257
Negative	3,219	480
Mixed feelings	3,031	438
unknown state	4,242	611

Table 3: Class Description

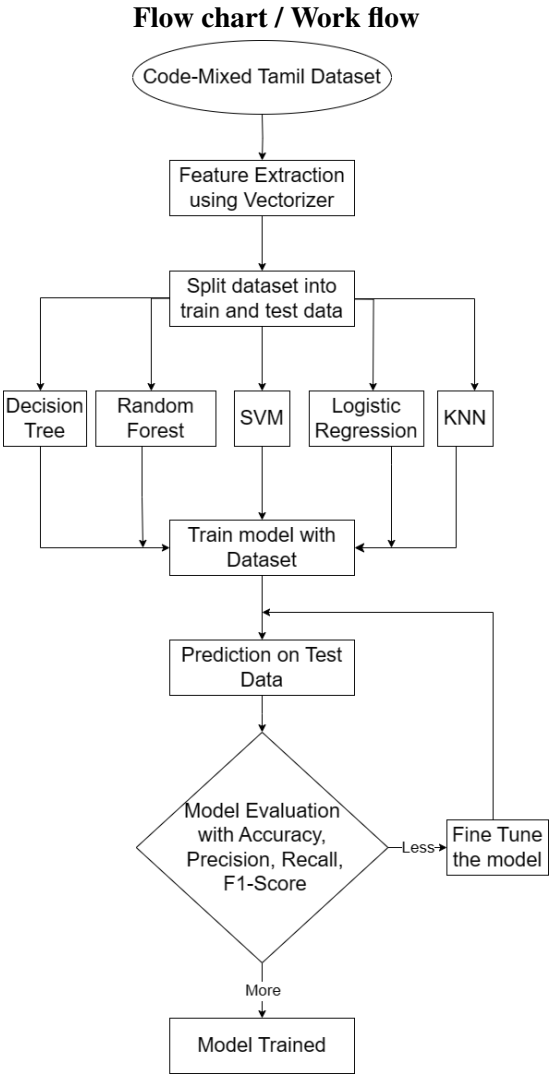

Figure 1: Proposed System Workflow

[2] Chakravarthi et al. (2020)

[3] (heg)

Unknown State. Model Generalisation: The test data predictions show how well the model can use learned patterns to interpret previously read text, providing insight into how well it functions with a range of natural language expressions. Evaluation: To evaluate the model's performance and accuracy on this new, independent dataset, the predicted sentiment labels can be compared with the ground truth labels, if available. Among the criteria considered are the accuracy and macro average F1 score of a model, which are two important markers of its efficiency. First off, the Decision Tree model has a macro average F1 score of 0.39 and an accuracy of 99.79 percent. Furthermore, with a 99.79 percent accuracy rate, the Random Forest model performs admirably. Its macro average F1 score, 0.35, is a little lower.The Support Vector Machine (SVM) is the next model, with a high macro average F1 score of 0.68 and an amazing accuracy of 78.55 percent.In comparison, the accuracy and macro average F1 score of Logistic Regression are 75.11 percent and 0.62, respectively. Ultimately, the macro average F1 score and accuracy of the K-Nearest Neighbours (KNN) model are 0.26 and 73.11 percent, respectively. The Decision Tree model outperforms the others with an impressive accuracy of 99.79 percent. This demonstrates a robust classification performance, setting it out as the top model in terms accuracy.[4]

6 Conclusion

Result:

Model	Accuracy	F1 Score
Decision Tree	0.99	0.39
Random Forest	0.99	0.35
SVM	0.78	0.68
Logistic Regression	0.75	0.62
KNN	0.73	0.26

Table 4: Accuracy and Macro average F1 Score

In summary, the project's standout feature lies in its adept utilization of the Decision Tree algorithm, showcasing exceptional accuracy in analyzing sentiments across a spectrum of classes. The comprehensive classification report deepens our understanding by emphasizing pivotal metrics like precision, recall, and macro average F1-score, illuminating the model's robust performance. Beyond its proficiency in sentiment analysis, the project serves as a pivotal resource for code-mixed research. The meticulously annotated dataset, marked by high inter-annotator agreement, lays a sturdy groundwork for prospective investigations. Its versatility extends into the exploration of diverse linguistic phenomena in code-mixed Tanglish, underscoring its adaptability and potential impact on linguistic research. While recognizing these strengths, it's imperative to address potential concerns like the risk of overfitting and challenges in generalizing to diverse contexts. These considerations pave the way for continual improvement. Nonetheless, the project's strong foundation, anchored by the Decision Tree algorithm and valuable resources, positions it as a commendable contribution to sentiment analysis in code-mixed languages. Exploring alternative algorithms promises to further elevate its already noteworthy capabilities.

[4] Hegde et al. (2022)

References

Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text.

Arwa Alshamsi, Reem Bayari, and Said Salloum. 2020. Sentiment analysis in english texts. *Advances in Science, Technology and Engineering Systems Journal*, 5(6).

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.

M.D. Devika, C. Sunitha, and Amal Ganesh. 2016. Sentiment analysis: A comparative study on different approaches. *Procedia Computer Science*, 87:44–49. Fourth International Conference on Recent Trends in Computer Science Engineering (ICRTCSE 2016).

Vaishali Ganganwar and R Rajalakshmi. 2019. Implicit aspect extraction for sentiment analysis: A survey of recent approaches. *Procedia Computer Science*, 165:485–491.

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.

Siru Liu and Jialin Liu. 2021. Public attitudes toward covid-19 vaccines on english-language twitter: A sentiment analysis. *Vaccine*, 39(39):5499–5505.

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.

Vidyapiratha Raveendirarasa and C.R.J. Amalraj. 2020. Sentiment analysis of tamil-english code-switched text on social media using sub-word level lstm. In *2020 5th International Conference on Information Technology Research (ICITR)*, pages 1–5.

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.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech Language*, 76:101407.

N Sripriya and S Divya. Sentiment analysis model for code-mixed tamil language.

R Thilagavathi and Kalyan Krishnakumari. 2016. Tamil english language sentiment analysis system. *International Journal of Engineering Research Technology (IJERT)*, 4:114–118.