InnovationEngineers@DravidianLangTech-EACL 2024: Sentimental Analysis of YouTube Comments in Tamil by using Machine Learning

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[1], **Palanimurugan V**[1], **Pavul chinnappan D**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{palanimuruganv.22aid, pavulchinnappand.22aid}@kongu.edu

Abstract

There is opportunity for machine learning and natural language processing research because of the growing volume of textual data. Although there has been little research done on trend extraction from YouTube comments, sentiment analysis is an intriguing issue because of the poor consistency and quality of the material found there.The purpose of this work is to use machine learning techniques and algorithms to do sentiment analysis on YouTube comments pertaining to popular themes.The findings demonstrate that sentiment analysis is capable of giving a clear picture of how actual events affect public opinion. This study aims to make it easier for academics to find high-quality sentiment analysis research publications.Data normalisation methods are used to clean an annotated corpus of 3200 citation sentences for the study.A system that uses the machine learning algorithms K-Nearest Neighbor (KNN), Naïve Bayes, SVC (Support Vector Machine), and RandomForest is constructed for classification. The accuracy of the system is evaluated using metrics such as the f1-score and correctness score.

1 Introduction

In the dynamic realm of social media, the role of sentiment analysis is increasingly crucial in comprehending the subtleties of user expressions. Traditionally designed for high-resource languages and individual utterances, sentiment analysis tools are encountering new challenges amid bilingual communities and code-mixed writing styles. This project addresses the growing significance of sentiment analysis, specifically focusing on code-mixed Tamil-English expressions prevalent across various social media platforms. Supervised learning approaches, traditionally reliant on annotated data, face limitations when applied to code-mixed languages. Notably, features based on lexical attributes, such as word dictionaries and parts of speech tagging, exhibit suboptimal performance in this multilingual context. To overcome these challenges, our research focuses on sentiment analysis within code-mixed Tamil-English contexts. Central to our approach is the implementation of the Decision Tree algorithm, which offers a robust solution for accurately classifying sentiments within this unique linguistic fusion. This project not only demonstrates exceptional accuracy using detailed metrics like precision, recall, and F1-score but also introduces a substantial corpus for under-resourced code- mixed Tanglish. Marked by high inter-annotator agreement, this dataset serves as a valuable resource for researchers delving into sentiment analysis and linguistic phenomena in code-mixed environments. Positioned at the intersection of sentiment analysis, machine learning, and code-mixed language research, our project extends beyond precise sentiment classification. It serves as a foundational resource for future investigations into the dynamic landscape of multilingual social media expressions.

2 Literature Survey

Certainly, here's a brief list of literature surveys by various authors focusing on sentiment analysis in Tamil Nadu:

The research paper work in Thavareesan and Mahesan (2021) demonstrates a sentiment analysis technique for Tamil texts utilising k-means clustering and k-nearest neighbour classifier. Despite different settings, the technique achieved an accuracy of 89.87% for the UJ_MovieReviews corpus utilising fastText and class-wise clustering. The main focus of the paper present in Kausikaa and Uma (2016) is sentiment analysis, a natural language processing task that involves determining the sentiment or emotion expressed in a given piece of text. In this case, the analysis is applied to tweets in both English and Tamil.It is discovered that the suggested system's F-measure accuracy value, which

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 262–265
March 22, 2024 ©2024 Association for Computational Linguistics

makes use of SVM, is 0.741. The primary focus of the paper work in Se et al. (2016)is on sentiment analysis, particularly suited to data with mixed Tamil codes. The term "code-mixing" describes the typical practice in multilingual cultures of combining two or more languages into a single statement or speech. SVM achieves 75.9% classification accuracy for Tamil movie reviews, a noteworthy achievement in Tamil language study. The primary focus of the paper repersent in Shanmugavadivel et al. (2022) is on sentiment analysis,particularly applicable to data with mixed Tamil codes. Code-mixing, a prevalent practice in multilingual cultures, is the blending of two or more languages inside a single statement or speech.The outcome shows that, with an accuracy of 0.66 using pre-processed Tamil code-mixed data, the hybrid deep learning model in particular, the CNN+BiLSTM model performs better than all the other models used.

The paper in Soumya and Pramod (2020) A review of machine learning methods for sentiment analysis of data with mixed Tamil codes This study examines the impact of pre-processing on Tamil code-mixed data using transfer learning, hybrid deep learning, deep learning, and traditional machine learning models. The study concentrates on removing emojis, punctuation, symbols, numerals, and repeating characters from the data. The hybrid deep learning model CNN+BiLSTM performs better with pre-processed Tamil code-mixed data, with an accuracy of 0.66. The study compares the performance of these models with the state-of-the-art methods, including IndicBERT, logistic regression, random forest, multinomial Naive Bayes, and linear support vector classification. In order to increase the accuracy of sentiment analysis on social media data, future research should focus on multimodal data sets and context-based algorithms.

The primary focus of the paper work in Se et al. (2016) is predicting sentiment in reviews related to Tamil movies using machine learning algorithms.with accurancy For categorising Tamil movie reviews, SVM yields a 75.9% accuracy rate.

The paper inChakravarthi et al. (2020b)The Dravidian-CodeMix-FIRE 2020 track focused on sentiment analysis for code-mixed Tamil and Malayalam in YouTube comments. Researchers aimed to classify sentiments using a weighted-F1 score, addressing linguistic complexities.

The primary focus of the paper work inChakravarthi et al. (2020a) sentiment in social media comments is crucial for decision-making. This study addresses challenges in sentiment analysis, especially in code-mixed text from low-resourced languages like Tamil, presenting a benchmark corpus and sentiment analysis results.

The paper in Hegde et al. (2022)Sentiment Analysis (SA) uses code-mixed data from social media for decision-making. However, low-resource languages like Tulu struggle with annotated data. A gold standard corpus of 7,171 Tulu comments is created, and Machine Learning algorithms are used to evaluate the dataset, showing encouraging performance.

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" or "Unknown State." The project aims to leverage machine learning, specifically the KNN algorithm, to accurately predict the sentiment of textual data.

TEXT	CATEGORY
Thalavaa neenga veera level boss and neega thaan marana mass That bgm..	Positive
Do or Die	Negative
Sema trailer fun movie Comali blockbuster 90s kids like	Unknown_State

Table 1: Dataset Description

4 Dataset Description

The training dataset comprises 33990 samples of code-mixed Tamil-English language, spanning diverse topics, with sentiment labels including Positive, Negative, Mixed Feelings, and Unknown State. The text data underwent TF-IDF vectorization, resulting in the creation of the train_tfdf matrix. The Decision Tree classifier demonstrated exceptional accuracy, achieving approximately 99.97% on the training data. 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.

5 Predictions on Test Data

Text Segments: The test dataset consists of 649 text segments in code-mixed Tamil-English language. Prediction Labels: Predicted sentiment labels were generated using the trained Decision Tree classifier, categorizing each text segment into sentiments such as Positive, Negative, Mixed Feelings, or Unknown State. Model Generalization: The predictions on the test data showcase the model's ability to generalize its learned patterns to previously unseen text, providing insights into its performance on real-world, diverse language expressions. Evaluation: The predicted sentiment labels can be compared with the ground truth labels, if available, to assess the model's accuracy and performance on this new, independent dataset.

6 Workflow

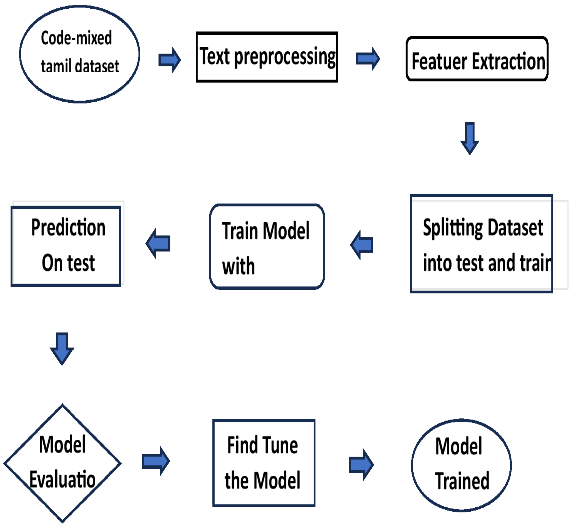

Figure 1: Processed workflow

7 Result

The sentiment is classified into categories such as Positive,Negative, or Unknown State by using different types of machine learning algorithm.

7.1 KNN report

Algorithm used	Accuracy
KNN	0.7345812952815269

Class Label	Precision	Recall	f1 score
Mixed Feelings	0.35	0.01	0.01
Positive	0.23	0.01	0.01
Negative	0.61	0.96	0.74
Unknown State	0.52	0.19	0.28

7.2 NAIVE BAYES report

Algorithm used	Accuracy
Naivebayes	0.598705501618123

Class Label	Precision	Recall	f1-score
Mixed Feelings	0.5	0.002	0.0024
Positive	0.78	0.029	0.056
Negative	0.596	0.99	0.746
Unknown State	0.76	0.031	0.060

7.3 SVC report

Algorithm used	Accuracy
SVC	0.6301853486319505

Class Label	Precision	Recall	f1-score
Mixed Feelings	0.537	0.035	0.066
Positive	0.535	0.148	0.233
Negative	0.634	0.963	0.765
Unknown State	0.632	0.232	0.339

7.4 Randomforest report

Algorithm used	Accuracy
Randomforest	0.99973589910195

Class Label	Precision	Recall	f1-score
Mixed Feelings	1.0	1.0	1.0
Positive	1.0	1.0	1.0
Negative	1.0	1.0	1.0
Unknown State	1.0	1.0	1.0

8 Conclusion

The project analyzes sentiment using the Random-Forest Tree technique, and it achieves remarkable accuracy in a variety of classes. Important parameters including precision, recall, and F1-score are shown in its thorough classification report. The project has the potential to influence linguistic study and is a useful resource for code-mixed research with a carefully annotated dataset. Nonetheless, it encounters obstacles like as overfitting and broad generalization to other settings. Despite these, the project is a commendable addition to sentiment analysis in code-mixed languages because of its solid base and available resources.

References

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, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020b. 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.

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.

N Kausikaa and V Uma. 2016. Sentiment analysis of english and tamil tweets using path length similarity based word sense disambiguation. *International Organization of Scientific Research Journal*, 1:82–89.

Shriya Se, R Vinayakumar, M Anand Kumar, and KP Soman. 2016. Predicting the sentimental reviews in tamil movie using machine learning algorithms. *Indian journal of science and technology*, 9(45):1–5.

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.

S Soumya and KV Pramod. 2020. Sentiment analysis of malayalam tweets using machine learning techniques. *ICT Express*, 6(4):300–305.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53. IEEE.